Journal of Machine Learning Research 24 (2023) 1-68 Submitted 6/20; Revised 2/23; Published 11/23

Randomized Spectral Co-Clustering for Large-Scale Directed Networks

Xiao Guo XIAOGUO@NWU.EDU.CN
Center for Modern Statistics
School of Mathematics
Northwest University, Xi'an, China
Yixuan Qiu QIUYIXUAN@SUFE.EDU.CN
School of Statistics and Management
Shanghai University of Finance and Economics,
Shanghai, China
Hai Zhang ZHANGHAI@NWU.EDU.CN
Center for Modern Statistics
School of Mathematics
Northwest University, Xi'an, China
Xiangyu Chang[*] XIANGYUCHANG@XJTU.EDU.CN
Center for Intelligent Decision-Making and Machine Learning
School of Management
Xi'an Jiaotong University, Xi'an, China

Editor: Vahab Mirrokni

Abstract

Directed networks are broadly used to represent asymmetric relationships among units. Co-clustering aims to cluster the senders and receivers of directed networks simultaneously. In particular, the well-known spectral clustering algorithm could be modified as the spectral co-clustering to co-cluster directed networks. However, large-scale networks pose great computational challenges to it. In this paper, we leverage sketching techniques and derive two randomized spectral co-clustering algorithms, one *random-projection-based* and the other *random-sampling-based*, to accelerate the co-clustering of large-scale directed networks. We theoretically analyze the resulting algorithms under two generative models – the stochastic co-block model and the degree-corrected stochastic co-block model, and establish their approximation error rates and misclustering error rates, indicating better bounds than the state-of-the-art results of co-clustering literature. Numerically, we design and conduct simulations to support our theoretical results and test the efficiency of the algorithms on real networks with up to millions of nodes. A publicly available R package RandClust is developed for better usability and reproducibility of the proposed methods.

Keywords: Co-clustering, Directed Network, Random Projection, Random Sampling, Stochastic co-Block Model

[*] Corresponding author.

1. Introduction

Recent advances in computing and measurement technologies have led to an explosion of large-scale network data (Newman, 2018). Networks can describe symmetric (undirected) or asymmetric (directed) relationships among interacting units in various fields, ranging from biology and informatics to social science and finance (Goldenberg et al., 2010). To extract knowledge from complex network structures, many clustering techniques, also known as community detection algorithms, are widely used to group together nodes with similar patterns (Fortunato, 2010). In particular, as asymmetric relationships are essential to the organization of networks, clustering *directed networks* is receiving more and more attention (Dhillon, 2001; Chung, 2005; Boley et al., 2011; Rohe et al., 2016). For large-scale directed network data, an appealing clustering algorithm should have not only the statistical guarantee but also the computational advantage.

To accommodate and explore the asymmetry in directed networks, the notion of *co-clustering* was introduced in Dhillon (2001); Rohe et al. (2016), and such an idea can be traced back to Hartigan (1972). Let $A \in \{0,1\}^{n\times n}$ be the network adjacency matrix such that $A_{ij} = 1$ if there is an edge from node i to node j, and $A_{ij} = 0$ otherwise. By convention, we also assume $A_{ii} = 0$. Then the ith row and column of A represent the outgoing and incoming edges for node i, respectively. Co-clustering refers to simultaneously clustering both the rows and the columns of A so that the nodes in a row cluster share similar sending patterns and the nodes in a column cluster share similar receiving patterns. Hence, we will refer to sending and receiving clusters to row and column clusters, respectively. Compared to standard clustering, where only one set of clusters is obtained, co-clustering a directed network yields two possibly different sets of clusters, which provide more insights and improve the understanding of the organization of directed networks.

Spectral clustering (Von Luxburg, 2007) is a natural and interpretable algorithm to group undirected networks, which first performs the eigendecomposition on a matrix representing the network, for example, the adjacency matrix A, and then runs k-means or other similar algorithms to cluster the resulting leading eigenvectors. Considering the asymmetry in directed networks, the standard spectral clustering algorithm has been modified to the *spectral co-clustering*, in which the eigendecomposition is replaced by the singular value decomposition (SVD), and the k-means is implemented on the left and right leading singular vectors, respectively. As the leading left and right singular vectors approximate the row and column spaces of A, it is expected that the resulting two sets of clusters contain nodes with similar sending and receiving patterns, respectively. A concrete version of the aforementioned algorithm is introduced in Rohe et al. (2016).

The spectral co-clustering is easy to be implemented and has been shown to have many excellent properties (Von Luxburg, 2007; Rohe et al., 2016). However, large-scale directed networks, namely, networks with a huge number of nodes or dense edges – say, millions of nodes and tens of millions of edges, pose significant challenges to the computation of spectral co-clustering. Improving spectral co-clustering's efficiency while maintaining a controllable accuracy becomes an urgent need. In this paper, we consider the problem of co-clustering large-scale directed networks based on randomization techniques, a popular approach to reducing the size of data with limited information loss (Mahoney et al., 2011; Woodruff et al., 2014; Drineas and Mahoney, 2016). Randomization techniques have been widely used in machine learning to speed up fundamental problems such as the least-squares regression and low-rank matrix approximation (see Drineas et al. (2006); Halko et al. (2011); Meng and Mahoney (2013); Nelson and Nguyên (2013); Pilanci and Wainwright (2016); Martinsson (2016); Clarkson and Woodruff (2017); Ye et al. (2017), among many others). The basic idea is to compromise the size of a data matrix by sampling a small subset of the matrix entries or forming linear combinations of the rows or columns. The entries or linear combinations are carefully chosen to preserve the significant information contained in the matrix. Hence, randomization techniques may provide a beneficial way to aid the spectral co-clustering of large-scale directed network data.

For a network with community structures, its adjacency matrix A is low-rank in nature, so the randomization for low-rank matrix approximation can be readily used to accelerate the SVD of A (Halko et al., 2011; Witten and Candès, 2015; Martinsson, 2016). We investigate two specific strategies: the random-projection-based and the random-sampling-based SVD. The random projection strategy compresses the original matrix A into a smaller one, whose rows (columns) are random linear combinations of the rows (columns) of A. In this way, the dimension of A is largely reduced, and the corresponding SVD is thus sped up. As for the random sampling strategy, the starting point is that there exist fast iterative algorithms to compute the partial SVD of a sparse matrix, such as orthogonal iteration and Lanczos iteration (Calvetti et al., 1994; Baglama and Reichel, 2005), whose time complexity is generally proportional to the number of non-zero elements of A. Therefore, a good way of accelerating the SVD of A is first to sample the elements of A to obtain a sparser matrix and then use fast iterative algorithms to compute its SVD. As a whole, the spectral co-clustering with the classical SVD therein replaced by the randomized SVD is called the *randomized spectral co-clustering*.

Given the fast randomization techniques, it is also critical to study the statistical accuracy of the resulting algorithms under certain generative models. To this end, we assume the directed network is generated from the *stochastic co-block model* (ScBM) or the *degree-corrected stochastic co-block model* (DC-ScBM) (Rohe et al., 2016). These two models assume the nodes are partitioned into two sets of non-overlapping blocks, one corresponding to the row cluster and the other to the column cluster. Generally, nodes in the same row (column) cluster are stochastically equivalent senders (receivers). That is, two nodes send out (receive) an edge to (from) a third node with the same probability if these two nodes are in the same row (column) cluster. The difference between these two models is that the degree-corrected model (Karrer and Newman, 2011; Rohe et al., 2016) considers the degree heterogeneity arising in real-life networks. The statistical error of the randomized spectral co-clustering is then studied under these two settings.

The merits of the current work lie in the following aspects:

- We analyze the true singular vector structure of population adjacency matrices generated by ScBMs and DC-ScBMs systematically. The results explain why the spectral co-clustering algorithms work well for directed networks and provide insights on designing the co-clustering algorithms for networks with and without degree heterogeneity. Different from most existing works of literature, we do not assume the row clusters and column clusters have the same number nor assume the ScBMs are full-rank. We also provide insightful conditions which would lead to the success of spectral co-clustering-based algorithms.

- We study the approximation and clustering performance of the two randomization-based algorithms from a statistical point of view. The results provide statistical insights into the randomized algorithms, which are new in randomization works of literature. For the approximation error, we show that under mild conditions, the minimax optimal error rate is attained, although we only use sketched data. More interestingly, in terms of misclustering error rates, we even obtain an improved upper bound for randomization methods relative to the non-randomized ones in Rohe et al. (2012) and Rohe et al. (2016).

- We evaluate the randomization-based spectral co-clustering algorithms on several large-scale networks with up to millions of nodes and tens of millions of edges. They show great efficiency, with most running times less than ten seconds on a regular personal computer, while yielding satisfactory clustering performance. The core algorithm is publicly available via the R package RandClust[1].

1.1 Related works

Randomization techniques have been widely used to speed up SVD and spectral-clustering-based algorithms; see Halko et al. (2011); Witten and Candès (2015); Musco and Musco (2015); Martinsson (2016); Tremblay et al. (2016); Erichson et al. (2019); Tremblay and Loukas (2020); Liao et al. (2020); Woodruff and Yasuda (2022), among others. The novelty of this paper is that we study the effect of randomization from a statistical perspective. We analyze the approximation error of the randomized adjacency matrix under ScBMs and DC-ScBMs carefully and find that the approximation error essentially attains the statistical minimax optimal rate under these two models. In particular, under the random sampling scheme, we deduce an approximation error that can not be obtained by simply combining the results in randomization and SBMs' works of literature. On the other hand, recent years have witnessed a few works studying the randomized algorithms under various statistical models, such as linear regression models, logistic regression models, and constrained regression models; see for example Ma et al. (2015); Raskutti and Mahoney (2016); Pilanci and Wainwright (2016); Wang et al. (2017, 2019); Li and Kang (2019). To the best of our knowledge, this is one of the first couples of works to study the randomization under the statistical network models. Note that Zhang et al. (2022) studied the randomized spectral clustering algorithms for large-scale undirected networks and analyzed the theoretical properties under the the stochastic block models (Holland et al., 1983). Compared with undirected networks, directed networks contain more information and bring the asymmetry that needs to be accommodated. As will be seen in Lemma 2, 3, 9 and 10, the underlying row and column clusters correspond to distinct singular vector structures. Accordingly, our theoretical results (see Theorem 5 for example) show that the estimated row and column clusters perform differently, and in principle, the nodes would be more easily clustered if their target cluster number is the same as the target rank of ScBMs and DC-SsBMs.

Spectral clustering has also been widely studied under various statistical network models, see Rohe et al. (2016, 2012); Qin and Rohe (2013); Lei and Rinaldo (2015); Yun and Proutiere (2016); Su et al. (2019); Abbe (2018); Tang et al. (2022); Arroyo and Levina

[1] https://github.com/XiaoGuo-stat/RandClust

(2022), among many others. In particular, Rohe et al. (2016) and its earlier version Rohe et al. (2012) are seminal works on spectral co-clustering of ScBMs, where they used the Laplacian matrix as the input. Compared with previous works, the merits of this work are as follows. First, our algorithm can handle large-scale networks for which the aforementioned work often fails. Second, we provide delicate analysis of the true singular structure of the population adjacency matrix and give sufficient and interpretable conditions on when the population-wise spectral-clustering-based algorithms could succeed (see Lemma 2, 3, 9 and 10). Note that this is rarely mentioned in Rohe et al. (2016, 2012) and previous literatures. More importantly, we argue in Theorem 14 that the extension of this work from the most considered full-rank ScBMs to rank-deficient ScBMs is possible, which is also not common in previous literatures. Third, as we utilize more advanced techniques, the resulting misclustering bounds are tighter than those in Rohe et al. (2016) and Rohe et al. (2012), although the latter two studied the non-randomized spectral co-clustering; see the following Table 1 and find more thorough discussions in Section 4. Last but not least, we apply the recently developed techniques for the entry-wise perturbation bound of eigenvectors (Abbe et al., 2020) to study the effect of random sampling on the spectral clustering with two underlying clusters, which has been proved to achieve the statistical minimax optimal misclustering error rate without randomization. Our analysis provides insightful results; see Theorem 13.

Table 1: A brief comparison of the misclustering error rates and the corresponding conditions in this work and in Rohe et al. (2016) and Rohe et al. (2012). K, n, α_n, p denote the number of clusters, number of nodes, maximum link probability of edges and the sampling rate in the random sampling scheme, respectively.

	Corollary C.1 in Rohe et al. (2016)	Corollary 4.1 in Rohe et al. (2012)
Bounds	$O(K^2 \text{log} n/n)$	$o(K^3 \text{log} n/\alpha_n^4)$
Conditions	$\alpha_n = O(1)$	$K = O(n^{1/4}/\text{log} n)$
	Theorem 5	Theorem 7
Bounds	$O(K^2/(n\alpha_n))$	$O(K^2/(pn\alpha_n))$
Conditions	(C2)	(C2), $p > 1/2$

The modern computation of SVD can be traced back to the 1960s, when the seminal works Golub and Kahan (1965); Golub and Reinsch (1970) provided the basis for the *EISPACK* and *LAPACK* routines, though the randomized matrix decomposition is a relatively young field. For computing partial SVD of matrices, iterative algorithms flourished; see Calvetti et al. (1994); Baglama and Reichel (2005); Jia and Niu (2003, 2010); Wu and Stathopoulos (2015); Wu et al. (2017). Another branch of algorithms are stochastic and incremental variants of the deterministic iterative algorithms (Oja and Karhunen, 1985; Arora et al., 2013; Shamir, 2015, 2016; Xu et al., 2018). Compared with iterative algorithms, the rationality of random-sampling-based scheme is straightforward: we accelerate the iterative method of Baglama and Reichel (2005) by sampling the original matrix at the price of accuracy. The random-sampling-based scheme can be generalized by using more advanced deterministic or stochastic iterative methods as the starting algorithm. On the other hand, for the random-projection-based scheme, its advantage over iterative methods lies in the following aspects. First, the random-projection-based method enables distributed computing since the matrix multiplications therein can be parallelized. Second, it is communication efficient because only a few passes over the input matrix are required. Overall, as is evidenced in Section 6, randomized methods are more efficient than deterministic iterative algorithms while maintaining good accuracy on large-scale networks. For reference, we summarize the time complexities of mentioned methods in Table 2.

Table 2: A summary of the time complexities of the randomized methods in this work and other methods for computing the SVD. n, K denote the number of nodes and the target rank, respectively. q denotes the power parameter and r, s denote the oversampling parameters. T_0 and T_1 denote the number of iterations. $\|A\|_0$ and $\|A^{\mathrm{rs}}\|_0$ represent the number of non-zero elements in the original adjacency matrix and the sparsified matrix. See Section 2 for more details.

Method	Full SVD	Iterative methods
Time	$O(n^3)$	$O(\|A\|_0 T_0)$
Method	Projection-based SVD	Sampling-based SVD
Time	$O((2q+1)\|A\|_0(K+\max(r,s)))$	$O(\|A^{\mathrm{rs}}\|_0 K T_1)$

The remainder of the paper is organized as follows. Section 2 introduces the randomized spectral co-clustering algorithms for co-clustering large-scale directed networks. Section 3 includes the theoretical analysis of the proposed algorithms under two network models. Section 4 discusses several theoretical aspects and possible extensions on the proposed methods. Section 5 and 6 present the experimental results on simulated and real-world data, respectively. Section 7 concludes the paper. Technical proofs are included in the Appendix.

2. Randomized spectral co-clustering

2.1 A brief review of prior art

For directed networks, co-clustering aims to find two possibly different sets of clusters, namely, row clusters and column clusters, to describe and understand the sending pattern and receiving pattern of nodes, respectively. Suppose there are K^y row clusters and K^z column clusters, and without loss of generality, assume $K^y \leq K^z$. Write the partial SVD of A, the adjacency matrix, as $A \approx U\Sigma V^\intercal$, where the left singular vectors $U \in \mathbb{R}^{n\times K^y}$ and right singular vectors $V \in \mathbb{R}^{n\times K^y}$ approximate the row and column spaces of A, respectively, and $\Sigma \in \mathbb{R}^{K^y\times K^y}$ is a diagonal matrix containing the K^y-largest singular values. On the other hand, U contains the eigenvectors of the symmetric matrix $AA^\intercal$, whose (i, j) entry corresponds to the number of common children of nodes i and j. Similarly, V represents the eigenvectors of $A^\intercal A$, whose (i, j) entry is the number of common parents of i and j. Therefore, U and V contain the sending and receiving information of each node, and clustering U and V respectively would yield clusters with nodes sharing similar sending and receiving patterns.

Based on the explanations above, the well-known spectral clustering is a good paradigm for co-clustering directed networks (Hartigan, 1972; Rohe et al., 2012, 2016). We consider the following two variants of spectral co-clustering algorithms, corresponding to different assumptions on the network. The first one is based on the standard spectral clustering, which first computes the SVD of A, and then uses k-means to cluster the left and right singular vectors of A, respectively (Algorithm 1, SCC). This algorithm is well-suited to networks whose nodes have approximately equal degrees. Whereas for networks whose nodes have heterogeneous degrees, the following algorithm (Algorithm 2, SsCC) is preferred. It first computes the SVD of A and then normalizes the non-zero rows of the left and right singular vectors such that the resulting rows have Euclidean norm 1. The zero rows are remained the same. The k-means clustering is then performed on the normalized rows of the left and right singular vectors, respectively. The normalization step aims to balance the importance of each node to facilitate the subsequent clustering procedures, which was also studied in Rohe et al. (2012, 2016); Lei and Rinaldo (2015), among others. As we will see in Section 3, this step is essential for co-clustering networks with degree heterogeneity.

Algorithm 1 Spectral co-clustering with k-means

```
Input:
    Adjacency matrix $A \in \mathbb{R}^{n\times n}$ of a directed network, number of row clusters $K^y$, and number of
    column clusters $K^z$ ($K^y \leq K^z$).

Compute the partial SVD of $A$, with left and right singular vectors $U \in \mathbb{R}^{n\times K^y}$ and $V \in \mathbb{R}^{n\times K^y}$.
Run $k$-means on $U$ with $K^y$ target clusters and on $V$ with $K^z$ clusters.
Output the co-clustering results.
```

Algorithm 2 Spectral co-clustering with spherical k-means

```
Input:
    Adjacency matrix $A \in \mathbb{R}^{n\times n}$ of a directed network, number of row clusters $K^y$, and number of
    column clusters $K^z$ ($K^y \leq K^z$).

Compute the partial SVD of $A$, with left and right singular vectors $U \in \mathbb{R}^{n\times K^y}$ and $V \in \mathbb{R}^{n\times K^y}$.
Construct $U'$ and $V'$, whose rows are normalized rows of $U$ and $V$, respectively. The zero rows
    are remained the same.
Run $k$-means on $U'$ with $K^y$ target clusters and on $V'$ with $K^z$ clusters.
Output the co-clustering results.
```

Now we discuss the time complexity of Algorithm 1 and 2. It is well-known that the classical full SVD generally takes $O(n^3)$ time, which is time-consuming when n is large. But in fact, only the partial SVD of A is needed, which can be done by fast iterative methods (Calvetti et al., 1994; Baglama and Reichel, 2005). They generally take $O(n^2K^yT_0)$ time, where T_0 is the iteration number corresponding to a certain error, and it can be large when n is large. For k-means, finding its optimal solution is NP-hard, and hence efficient heuristic algorithms are commonly employed. In this paper, we use the Lloyd's algorithm to solve k-means, whose time complexity is proportional to n. Alternatively, one can use a more delicate $(1+\epsilon)$-approximate k-means (Kumar et al., 2004) for a good approximate solution within a constant fraction of the optimal value. Based on the discussions above, the time complexities of Algorithm 1 and 2 are dominated by the SVD, which encourages the use of randomization techniques to speed up the computation of SVD for further improving the spectral co-clustering.

2.2 Random-projection-based spectral co-clustering (RP-SCC)

The basic idea of the *Random-Projection-based Spectral Co-Clustering* (RP-SCC) is to compress the adjacency matrix A into a smaller matrix, and then apply a standard SVD to the compressed one, thus saving the computational cost. The approximate SVD of the original A can be recovered by postprocessing the SVD of the smaller matrix (Halko et al., 2011; Witten and Candès, 2015; Martinsson, 2016).

For an asymmetric matrix A with a target rank K^y, the objective is to find orthonormal bases $Q, T \in \mathbb{R}^{n\times K^y}$ such that

$$A \approx QQ^\intercal ATT^\intercal := A^{\mathrm{rp}}.$$

It is not hard to see that $QQ^\intercal$ projects the column vectors of A to the column space of Q, and $TT^\intercal$ projects the row vectors of A to the column space of T. Therefore, Q and T approximate the column and row spaces of A, respectively. In randomization methods, Q and T can be built via *random projection* (Halko et al., 2011). Take Q as an example, one first constructs an $n \times K^y$ random matrix whose columns are random linear combinations of the columns of A, and then orthonormalizes the K^y columns using the QR decomposition to obtain the orthonormal matrix Q. Once Q and T are constructed, the standard SVD is performed on $Q^\intercal AT$, and the approximate SVD of A can be achieved by left-multiplying Q and right-multiplying T. The whole procedure of the random-projection-based SVD can be summarized as the following steps:

- *Step 1:* Construct two test matrices $\Omega, \Gamma \in \mathbb{R}^{n\times K^y}$ with independent standard Gaussian entries.
- *Step 2:* Obtain Q and T via the QR decomposition $A\Omega \to QR_1$ and $A^\intercal\Gamma \to TR_2$.
- *Step 3:* Compute SVD of $Q^\intercal AT \to U_s\Sigma V_s^\intercal$.
- *Step 4:* Output the approximate SVD of A as $A \approx U^{\mathrm{rp}}\Sigma(V^{\mathrm{rp}})^\intercal$, where $U^{\mathrm{rp}} := QU_s$ and $V^{\mathrm{rp}} := TV_s$.

To fix ideas, RP-SCC generally refers to spectral co-clustering with the SVD therein replaced by the random-projection-based SVD. While in some places to follow, RP-SCC and RP-SsCC refer particularly to random-projection-based SVD coupled with Algorithm 1 and 2, respectively.

In actual implementation, the *oversampling* and *power iteration* schemes can be used to improve the performance of the randomized SVD (Halko et al., 2011; Martinsson, 2016). Oversampling uses extra r and s ($K^y + r$ and $K^y + s$ in total) random projections to form the sketch matrices $A\Omega$ and $A^\intercal\Gamma$ in *Step 2*, which reduce the information loss when the rank of A is not exactly K^y. The power iteration scheme employs $(AA^\intercal)^q A\Omega$ and $(A^\intercal A)^q A^\intercal\Gamma$

instead of $A\Omega$ and $A^\intercal\Gamma$ in *Step 2*. This treatment improves the quality of the sketch matrix when the singular values of A are not rapidly decreasing.

The time complexity of RP-SCC is dominated by the matrix multiplication operations $(AA^\intercal)^q A\Omega$ and $(A^\intercal A)^q A^\intercal\Gamma$ in *Step 2*, which takes $O((2q+1)\|A\|_0(K^y+\max(r,s)))$ time. Note that the classical SVD in *Step 3* is cheap as the matrix dimension is as low as $K^y + \max(r,s)$. In addition, the time of *Step 2* can be further improved if one uses structured random test matrices or performs the matrix multiplications in parallel. The random-projection-based SVD is numerically stable, and comes with its good theoretical guarantee (Halko et al., 2011; Witten and Candès, 2015; Martinsson, 2016).

2.3 Random-sampling-based spectral co-clustering (RS-SCC)

The *Random-Sampling-based Spectral Co-Clustering* (RS-SCC) is based on the fact that real-world networks are often sparse (Watts and Strogatz, 1998; Chang et al., 2019) that the number of non-zero elements in the adjacency matrix A is $O(n^\alpha)$ with $0<\alpha<2$. It is known that the time complexity of fast iterative algorithms for partial SVD is proportional to the number of non-zero elements of the matrix (Calvetti et al., 1994; Baglama and Reichel, 2005). Consequently, RS-SCC makes the partial SVD of A more efficient by randomly sampling the elements of A, followed by a fast iterative algorithm to compute the leading singular vectors of the sparsified matrix. The partial SVD of A can then be approximated by that of the sparsified matrix.

We use the following simple strategy to construct the sparsified matrix A^{rs}: each element of A is sampled with equal probability p, and the elements that are not sampled are forced to be zero. Formally, for each pair of (i,j),

$$A^{\text{rs}}_{ij} = \begin{cases} \frac{A_{ij}}{p}, & \text{if } (i,j) \text{ is selected,} \\ 0, & \text{if } (i,j) \text{ is not selected,} \end{cases} \tag{2.1}$$

where A_{ij} is divided by p to remove bias as we will see in Section 3. If the sampling probability p is not too small, then A^{rs} is close to A with little information loss. With A^{rs} at hand, the random-sampling-based SVD follows:

- *Step 1:* Form the sparsified matrix A^{rs} via (2.1).
- *Step 2:* Compute the partial SVD of A^{rs} using the fast iterative algorithm in Calvetti et al. (1994) or Baglama and Reichel (2005) such that $A^{\text{rs}} \approx U^{\text{rs}}_{n\times K^y}\Sigma_{K^y\times K^y}(V^{\text{rs}})^\intercal_{K^y\times n}$.

Generally, RS-SCC refers to spectral co-clustering with the SVD therein replaced by the random-sampling-based SVD, while in some places to follow, we may use RS-SCC and RS-SsCC to distinguish Algorithm 1 and 2.

The time complexities of *Step 1* and *Step 2* are approximately $O(\|A\|_0)$ and $O(\|A^{\text{rs}}\|_0 K^y T_1)$, where $\|A\|_0$ denotes the number of non-zero elements in A, and T_1 is the number of iterations. The number of edges in a real-world network is typically far below $O(n^2)$, thus making RS-SCC rather efficient.

3. Theoretical analysis

3.1 Preliminaries

We analyze the theoretical properties of the randomized spectral co-clustering algorithms under two generative models, ScBM and DC-ScBM (Rohe et al., 2016). In ScBM, nodes in a common row cluster are stochastically equivalent senders in the sense that they send out an edge to a third node with equal probabilities. Similarly, nodes in a common column cluster are stochastically equivalent receivers, as they receive an edge from a third node with equal probabilities. In DC-ScBM, however, the probabilities depend not only on the row or column clusters but also on propensity parameters for each node.

To provide the formal definitions of these two models, we first introduce the following notation. Recall that for a directed network $A \in \mathbb{R}^{n\times n}$, we assume there exist K^y row clusters and K^z column clusters with $K^y \leq K^z$. For $i = 1, ..., n$, let $g_i^y \in \{1, ..., K^y\}$ and $g_i^z \in \{1, ..., K^z\}$ denote the assignments of the row cluster and column cluster of node i, respectively. Alternatively, the cluster assignments can be represented by membership matrices defined as follows. Let $\mathbb{M}_{n,K}$ be the set of all $n \times K$ matrices that have exactly one 1 and $K - 1$ 0's in each row, and $Y \in \mathbb{M}_{n,K^y}$ and $Z \in \mathbb{M}_{n,K^z}$ are two matrices such that $Y_{ig_i^y} = 1$ and $Z_{ig_i^z} = 1$ for each i. Y and Z are then called row and column membership matrices, respectively. For $1 \leq k \leq K^y$, let $G_k^y = \{1 \leq i \leq n : g_i^y = k\}$ be the set of nodes belonging to row cluster k, and denote by $n_k^y = |G_k^y|$ its size. Similarly, for $1 \leq k \leq K^z$, define $G_k^z = \{1 \leq i \leq n : g_i^z = k\}$ and $n_k^z = |G_k^z|$ for column cluster k. For any matrix B and proper index sets I and J, B_{I*} and B_{*J} denote the sub-matrices of B that consist of the rows in I and columns in J, respectively. $\|B\|_{\mathrm{F}}$, $\|B\|_2$, and $\|B\|_\infty$ are the Frobenius norm, spectral norm, and the element-wise maximum absolute value of B, respectively. Finally, $\mathrm{diag}(B)$ denotes a diagonal matrix whose diagonal entries are the same as those of B.

3.2 Stochastic co-block model

We use the following definition for ScBM.

Definition 1 (ScBM, Rohe et al., 2016) *Let $Y \in \mathbb{M}_{n,K^y}$ and $Z \in \mathbb{M}_{n,K^z}$ be the row and column membership matrices, respectively. Let $B \in [0, 1]^{K^y\times K^z}$ be the connectivity matrix whose (k, l)th element is the probability of a directed edge from any node in the row cluster k to any node in the column cluster l. Given (Y, Z, B), each element of the network adjacency matrix $A = (a_{ij})_{1\leq i,j\leq n}$ is generated independently as $a_{ij} \sim \mathrm{Bernoulli}(B_{g_i^y g_j^z})$ if $i \neq j$, and $a_{ij} = 0$ if $i = j$.*

Throughout this subsection, we would consider the ScBM parameterized by (Y, Z, B). Note that when $Y = Z$, and B and A are symmetric, then ScBMs reduce to the stochastic block models (SBMs) (Holland et al., 1983). Define $P = YBZ^\intercal$, and let σ_n and γ_n denote its maximum and minimum non-zero singular values, respectively. The formulation of P makes sense throughout this subsection. It is easy to see that P is the population version of A in the sense that $\mathbb{E}(A) = P - \mathrm{diag}(P)$. We assume throughout this subsection that $\mathrm{rank}(P) = \mathrm{rank}(B) = K^y$, though relaxing this assumption to $\mathrm{rank}(P) = \mathrm{rank}(B) \leq K^y$ is also feasible as we will discuss in Section 4. The following Lemma 2 reveals the structure of the singular vectors of P.

Lemma 2 *Denote the SVD of the population matrix $P = YBZ^\intercal$ by $\bar{U}_{n\times K^y}\bar{\Sigma}_{K^y\times K^y}\bar{V}^\intercal_{K^y\times n}$. Define $\Delta_y = \mathrm{diag}(\sqrt{n_1^y},...,\sqrt{n_{K^y}^y})$, $\Delta_z = \mathrm{diag}(\sqrt{n_1^z},...,\sqrt{n_{K^z}^z})$, and denote the SVD of $\Delta_y B \Delta_z$ by $L_{K^y\times K^y}D_{K^y\times K^y}R^\intercal_{K^y\times K^z}$. Then the following arguments hold for any $1\leq i\neq j\leq n$.*

(1) If $Y_{i*}=Y_{j*}$, then $\bar{U}_{i*}=\bar{U}_{j*}$; otherwise

$$\|\bar{U}_{i*}-\bar{U}_{j*}\|_2=\sqrt{(n^y_{g_i^y})^{-1}+(n^y_{g_j^y})^{-1}}.$$

(2) If $Z_{i*}=Z_{j*}$, then $\bar{V}_{i*}=\bar{V}_{j*}$; otherwise

$$\|\bar{V}_{i*}-\bar{V}_{j*}\|_2=\left\|\frac{R_{g_i^z*}}{\sqrt{n^z_{g_i^z}}}-\frac{R_{g_j^z*}}{\sqrt{n^z_{g_j^z}}}\right\|_2.$$

Moreover, if $\Delta_z^{-1}R$'s rows are mutually distinct such that there exists a deterministic sequence $\{\xi_n\}_{n\geq 1}$ satisfying

$$\min_{1\leq k\neq l\leq K^z}\left\|\frac{R_{k*}}{\sqrt{n_k^z}}-\frac{R_{l*}}{\sqrt{n_l^z}}\right\|_2\geq \xi_n>0, \tag{C1}$$

then $\|\bar{V}_{i*}-\bar{V}_{j*}\|_2\geq\xi_n>0$.

The following lemma provides an explicit condition on B which suffices for (C1).

Lemma 3 *Under the same parameter setting as in Lemma 2, if the columns of B are mutually distinct such that*

$$\min_{1\leq k\neq l\leq K^y}\|B_{*k}-B_{*l}\|_2\geq\mu_n$$

for some $\mu_n>0$, then (C1) holds with $\xi_n=\mu_n\cdot\min_{1\leq k\leq K^y}(n_k^y)^{1/2}/\sigma_n$, where σ_n is the maximum singular value of P.

Lemma 2 and 3 provide the following important insights. The left singular vectors $\bar{U}$ of P reveal the true row clusters in the sense that two rows of $\bar{U}$ are identical if and only if the corresponding nodes are in the same row cluster. In addition, for two nodes in distinct row clusters, the distance between their corresponding rows of $\bar{U}$ is determined by their row cluster sizes. However, the story for the column clusters is slightly different. Nodes in common column clusters have equal rows in $\bar{V}$, but the converse is generally not true which is caused by the fact that $K^z\geq K^y=\mathrm{rank}(B)$. Nonetheless, in Lemma 3, we see that if the columns of B are mutually distinct, then the converse is also true. In particular, a larger minimum distance of the column pairs in B would possibly lead to a larger minimum distance of the rows pairs in $\bar{V}$. Based on these facts, one would expect that the spectral co-clustering algorithms (Algorithm 1 and 2) would estimate the true underlying clusters well if the singular vectors of A are close enough to those of P, which by the Davis-Kahan-Wedin theorem (O'Rourke et al., 2018) would hold if A and P are close in some sense. Moreover, A is approximated by A^{rp} in RP-SCC and by A^{rs} in RS-SCC, respectively, so in subsequent sections, we first study the deviation of A^{rp} and A^{rs} from P, and then examine the clustering performance of the proposed methods.

3.2.1 Performance of RP-SCC in ScBMs

First, Theorem 4 quantifies the spectral deviation of A^{rp} from P.

Theorem 4 *Let $A^{\mathrm{rp}} = QQ^\intercal ATT^\intercal$ be the random projection approximation to A with target rank K^y. Assume that the oversampling parameters (r, s) satisfy $K^y+r \leq n$ and $K^y+s \leq n$, and the test matrices have i.i.d. standard Gaussian entries. If*

$$\mathrm{max}_{kl} B_{kl} \leq \alpha_n \text{ for some } \alpha_n \geq c_0 \,\mathrm{log}n/n \text{ and } c_0 > 0, \tag{C2}$$

and

$$\min(r, s) \geq 4,\ \max(r\mathrm{log}r, s\mathrm{log}s) \leq n,\ q = c_1 \cdot n^{1/\tau} \tag{C3}$$

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

$$\|A^{\mathrm{rp}} - P\|_2 \leq c_2\sqrt{n\alpha_n}, \tag{3.1}$$

with probability at least $1 - 6r^{-r} - 6s^{-s} - 3n^{-\epsilon}$.

Theorem 4 implies that the randomized adjacency matrix A^{rp} concentrates around P at the rate of $\sqrt{n\alpha_n}$, where $n\alpha_n$ can be regarded as the upper bound of the expected degree in the network A. (C2) prevents the network from being too sparse, and it is a common requirement in SBMs literature; see Lei and Rinaldo (2015), among others. (C3) ensures that the error caused by the random projection, namely, $\|A^{\mathrm{rp}} - A\|_2$, is dominated by the error caused by the ScBMs, namely, $\|A - P\|_2$. The bound (3.1) achieves the statistical minimax optimal error rate (Gao et al., 2017). In this sense, the random projection pays no price under the framework of ScBMs.

Denote

$$\tau = \mathrm{min}_{l\neq k} \sqrt{(n_k^y)^{-1} + (n_l^y)^{-1}} \quad \text{and} \quad \delta = \min_{1\leq k\neq l\leq K^z} \|\frac{R_{k*}}{\sqrt{n_k^z}} - \frac{R_{l*}}{\sqrt{n_l^z}}\|_2, \tag{3.2}$$

where recall that R denotes the right singular matrix of $\Delta_y B \Delta_z$. The next theorem provides an upper bound for the proportion of misclustered nodes.

Theorem 5 *Suppose (C2) and (C3) hold and other parameter settings are the same with those in Theorem 4. Denote the sets of misclustered nodes with respect to the row and column clusters respectively by $M^y \subseteq \{1, ..., n\}$ and $M^z \subseteq \{1, ..., n\}$, and recall τ and δ defined in (3.2). Then under ScBMs, the following results hold for RP-SCC.*

(1) With probability larger than $1 - 6r^{-r} - 6s^{-s} - 3n^{-\epsilon}$ for any $\epsilon > 0$, the misclustering rate with respect to the row clusters satisfies

$$\frac{|M^y|}{n} \leq c_3^{-1}\frac{K^y\alpha_n}{\tau^2\gamma_n^2}, \tag{3.3}$$

provided that $\frac{K^y\alpha_n n}{n_k^y\tau^2\gamma_n^2} \leq c_3$ for any $k = 1, ..., K^y$ and some constant $c_3 > 0$.

(2) With probability larger than $1-6r^{-r}-6s^{-s}-3n^{-\epsilon}$ for any $\epsilon>0$, the misclustering rate with respect to the column clusters satisfies

$$\frac{|M^z|}{n}\leq c_4^{-1}\frac{K^y\alpha_n}{\delta^2\gamma_n^2}. \tag{3.4}$$

provided that $\frac{K^y\alpha_n n}{n_k^z\delta^2\gamma_n^2}\leq c_4$ for any $k=1,...,K^z$ and some constant $c_4>0$.

Theorem 5 provides upper bounds for the misclustering rates with respect to row clusters and column clusters, as indicated in (3.3) and (3.4). Recalling Lemma 2, we can see that the clustering performance depends on the minimum row distances τ and δ of the population singular vectors $\bar{U}$ and $\bar{V}$. As expected, larger distances imply more accurate clusters. Note that $\frac{K^y\alpha_n n}{n_k^y\tau^2\gamma_n^2}\leq c_3$ and $\frac{K^y\alpha_n n}{n_k^z\delta^2\gamma_n^2}\leq c_4$ are required to ensure the validity of the results. They actually ensure that each true cluster has nodes that are correctly clustered. These conditions can be easily met. Moreover, when $K^y=K^z$, Lemma 2 implies that the column clusters and the row clusters behave similarly with similar misclustering error bounds. In Section 4, we will discuss the misclustering error bounds and compare them with the state-of-art in more detail.

3.2.2 Performance of RS-SCC in ScBMs

We first provide the deviation of $A^{\rm rs}$ from P in the sense of the spectral norm.

Theorem 6 *Let $A^{\rm rs}$ be the random sampling approximation to A with sampling probability p. Suppose (C2) holds, then for any $\nu>0$ and $0<p\leq 1$, there exist constants $c_5>0$ and $c_6>0$ such that*

$$\|A^{\rm rs}-P\|_2\leq c_5\max\Big\{\sqrt{\frac{n\alpha_n}{p}},\ \frac{\sqrt{\log n}}{p},\ \Delta(n,\alpha_n,p)\Big\}, \tag{3.5}$$

where

$$\Delta(n,\alpha_n,p):=\sqrt{\frac{n\alpha_n^2}{p}}\Big(1+p^{1/4}\cdot\max\big(1,\ \sqrt{\frac{1}{p}-1}\big)\Big), \tag{3.6}$$

with probability larger than $1-2n^{-\nu}-\exp\Big(-c_6np\big(1+p^{1/4}\cdot\max(1,\ \sqrt{\frac{1}{p}-1})^2\big)\Big)$.

Theorem 6 says that $A^{\rm rs}$ concentrates around P at the rate shown in (3.5). As expected, the rate decreases as p increases. Note that (3.5) simplifies to $O((\sqrt{n\alpha_n/p})$, provided that $p>1/2$.

In what follows, for notational simplicity, we denote

$$\Phi(n,p,\alpha_n):=\max\Big\{\sqrt{\frac{n\alpha_n}{p}},\ \frac{\sqrt{\log n}}{p},\ \Delta(n,\alpha_n,p)\Big\}, \tag{3.7}$$

where Δ is defined in (3.6). The next theorem provides an upper bound for the misclustering error rates of RS-SCC under ScBMs.

Theorem 7 *Suppose (C2) holds and other parameter settings are identical with those in Theorem 6. Denote the sets of misclustered nodes with respect to the row and column clusters respectively by $M^y \subseteq \{1, ..., n\}$ and $M^z \subseteq \{1, ..., n\}$, and recall the terms defined in (3.2) and (3.7). Then under ScBMs, the following results hold for RS-SCC.*

(1) With probability larger than $1 - 2n^{-\nu} - \exp\Big(- c_6 np\big(1 + p^{1/4} \cdot \max(1, \sqrt{\tfrac{1}{p} - 1})^2\big)\Big)$ for any $\nu > 0$, the misclustering rate with respect to the row clusters satisfies

$$\frac{|M^y|}{n} \leq c_7^{-1} \frac{K^y \Phi^2(n, p, \alpha_n)}{n\tau^2 \gamma_n^2}. \tag{3.8}$$

provided that $\frac{K^y \Phi^2(n,p,\alpha_n)}{n_k^y \tau^2 \gamma_n^2} \leq c_7$ for any $k = 1, ..., K^y$ and some constant $c_7 > 0$.

(2) With probability larger than $1 - 2n^{-\nu} - \exp\Big(- c_6 np\big(1 + p^{1/4} \cdot \max(1, \sqrt{\tfrac{1}{p} - 1})^2\big)\Big)$ for any $\nu > 0$, the misclustering rate with respect to the column clusters satisfies

$$\frac{|M^z|}{n} \leq c_8^{-1} \frac{K^y \Phi^2(n, p, \alpha_n)}{n\delta^2 \gamma_n^2}. \tag{3.9}$$

provided that $\frac{K^y \Phi^2(n,p,\alpha_n)}{n_k^z \delta^2 \gamma_n^2} \leq c_8$ for any $k = 1, ..., K^z$ and some constant $c_8 > 0$.

The proof of Theorem 7 is similar to that of Theorem 5, hence we omit it. (3.8) and (3.9) provide upper bounds for the proportion of the misclustered nodes in the estimated row clusters and column clusters, respectively. As in the random projection scheme, the minimum non-zero row distance in the true singular vectors $\bar{U}$ and $\bar{V}$, i.e., τ and δ, play an important role in the clustering performance. Note that we require similar technical conditions as those in Theorem 5, and they could be achieved with ease as we will discuss in Section 4.

3.3 Degree-corrected stochastic co-block model

In ScBMs, the nodes within each row cluster and column cluster are stochastic equivalent. While in real networks, there exists hubs whose edges are far more than those of the non-hub nodes (Karrer and Newman, 2011). To model such degree heterogeneity, DC-ScBMs introduce extra parameters $\theta^y = (\theta^y_1, \theta^y_2, \ldots, \theta^y_n)^\top \in \mathbb{R}^n_+$ and $\theta^z = (\theta^z_1, \theta^z_2, \ldots, \theta^z_n)^\top \in \mathbb{R}^n_+$, which represent the propensity of each node to send and receive edges. We used the following definition for DC-ScBMs.

Definition 8 (DC-ScBM, Rohe et al., 2016) *Let $Y \in \mathbb{M}_{n,K^y}$ and $Z \in \mathbb{M}_{n,K^z}$ be the row and column membership matrices, respectively. Let $B \in [0,1]^{K^y \times K^z}$ be the connectivity matrix whose (k,l)th element is the probability of a directed edge from any node in the row cluster k to any node in the column cluster l. Let $\theta^y \in \mathbb{R}^n$ and $\theta^z \in \mathbb{R}^n$ be the node propensity parameters. Given $(Y, Z, B, \theta^y, \theta^z)$, each element of the network adjacency matrix $A = (a_{ij})_{1 \leq i,j \leq n}$ is generated independently as $a_{ij} \sim \mathrm{Bernoulli}(\theta^y_i \theta^z_j B_{g^y_i g^z_j})$ if $i \neq j$, and $a_{ij} = 0$ if $i = j$.*

From the above definition, we see that the probability of an edge from node i to j depends on not only the row cluster and column cluster they respectively lie in, but also the propensity of them to send and receive edges, respectively. Note that θ^y and θ^z would make the model non-identifiable except that additional assumptions are enforced. In this paper, we assume $\max_{i\in G_k^y}\theta_i^y = 1$ and $\max_{i\in G_k^z}\theta_i^z = 1$ for each $k = 1, ..., K^y$ and $k = 1, ..., K^z$, respectively.

Throughout this subsection, we would consider the DC-ScBM parameterized by $(\theta^y, \theta^z, Y, Z, B)$. Note that when θ^y and θ^z take 1 at all their entries, then the DC-ScBM reduces to ScBM. With a slight abuse of notation, define $P = \mathrm{diag}(\theta^y)YBZ^\intercal\mathrm{diag}(\theta^z)$, which is actually the population version of A in the sense that $\mathbb{E}(A) = P - \mathrm{diag}(P)$. The formulation of P would be used throughout this subsection. Assume $\mathrm{rank}(P) = \mathrm{rank}(B) = K^y$, and denote the maximum and minimum singular values of P by σ_n and γ_n, respectively.

Before analyzing the singular structure of P, we now introduce some notations. Let ϕ_k^y and ϕ_k^z be $n\times 1$ vectors that consistent with θ^y and θ^z respectively on G_k^y and G_k^z and zero otherwise. Thus, $\sum_{k=1}^{K^y}\phi_k^y = \theta^y$ and $\sum_{k=1}^{K^z}\phi_k^z = \theta^z$. Let $\Psi^y = \mathrm{diag}(\|\phi_1^y\|_2, ..., \|\phi_{K^y}^y\|_2)$, $\Psi^z = \mathrm{diag}(\|\phi_1^z\|_2, ..., \|\phi_{K^z}^z\|_2)$ and $\Psi^y B\Psi^z = \tilde{B}$. Define $\tilde{\theta}^y$ and $\tilde{\theta}^z$ be $n\times 1$ vectors such that their ith elements are $\theta_i^y/\|\phi_{g_i}^y\|_2$ and $\theta_i^z/\|\phi_{g_i}^z\|_2$, respectively. The next lemma reveals the singular structure in P.

Lemma 9 *Denote the SVD of the population matrix* $P = \mathrm{diag}(\theta^y)YBZ^\intercal\mathrm{diag}(\theta^z)$ *by* $\bar{U}_{n\times K^y}$ $\bar{\Sigma}_{K^y\times K^y}\bar{V}_{K^y\times n}^\intercal$. *Denote the SVD of* $\tilde{B}$ *by* $H_{K^y\times K^y}D_{K^y\times K^y}J_{K^y\times K^z}^\intercal$. *For any two vectors* a *and* b, $\cos(a, b)$ *is defined to be* $a^\intercal b/\|a\|_2\|b\|_2$. *Then the following arguments hold for any* $1\le i\ne j\le n$.

(1) $\bar{U}_{i*} = \tilde{\theta}_i^y H_{k*}$ *for* $i\in G_k^y$, *where* H *is an* $K^y\times K^y$ *orthonormal matrix. As a result, if* $Y_{i*} = Y_{j*}$, *then* $\cos(\bar{U}_{i*}, \bar{U}_{j*}) = 1$; *otherwise* $\cos(\bar{U}_{i*}, \bar{U}_{j*}) = 0$.

(2) $\bar{V}_{i*} = \tilde{\theta}_i^z J_{k*}$ *for* $i\in G_k^z$, *where* J *is an* $K^z\times K^y$ *matrix with orthonormal columns. As a result, if* $Z_{i*} = Z_{j*}$, *then* $\cos(\bar{V}_{i*}, \bar{V}_{j*}) = 1$; *otherwise*

$$\cos(\bar{V}_{i*}, \bar{V}_{j*}) = \cos((\tilde{B}_{*g_i^z})^\intercal H\bar{\Sigma}^{-1}, (\tilde{B}_{*g_j^z})^\intercal H^{-1}\bar{\Sigma}^{-1}).$$

Moreover, if there exists a deterministic sequence $\{\xi_n'\}_{n\ge 1} < 1$ *such that*

$$\cos((\tilde{B}_{*g_i^z})^\intercal H\bar{\Sigma}^{-1}, (\tilde{B}_{*g_j^z})^\intercal H\bar{\Sigma}^{-1}) \le \xi_n', \tag{C4}$$

then $\cos(\bar{V}_{i*}, \bar{V}_{j*}) \le \xi_n'$.

The following lemma provides sufficient conditions for (C4) to hold.

Lemma 10 *Under the same parameter setting as in Lemma 9, if the columns of* $\tilde{B}$ *are* not *mutually proportional such that there exists a deterministic sequence* $\{\zeta_n\}_{n\ge 1} < 1$ *satisfying*

$$\max_{1\le k\ne l\le K^z}\cos(\tilde{B}_{*k}, \tilde{B}_{*l}) \le \zeta_n,$$

and $\underline{\iota}_n \le \min_{1\le k\le K^z}\|\tilde{B}_{*k}\|_2 \le \max_{1\le k\le K^z}\|\tilde{B}_{*k}\|_2 \le \bar{\iota}_n$, *then (C4) holds with*

$$\xi_n' = \sqrt{1 - \left(\frac{\underline{\iota}_n}{\bar{\iota}_n}\cdot\frac{\sigma_{\min}(H)}{\sigma_{\max}(H)}\cdot\frac{\sigma_{\min}(P)}{\sigma_{\max}(P)}\right)^2(1-\zeta_n)^2},$$

where $\sigma_{\min}(\cdot)$ and $\sigma_{\max}(\cdot)$ denote the minimum and maximum non-zero singular value of matrices.

Lemma 9 says that the directions of two rows in $\bar{U}$ ($\bar{V}$) are the same if and only if the corresponding nodes lie in the same row (column) cluster. For example, if node i and node j are in the same row cluster k, then $\bar{U}_{i*}$ and $\bar{U}_{i*}$ both have direction H_{k*}. On the other hand, if two nodes are in different row (column) clusters, there exists an angle between the corresponding rows of $\bar{U}$ ($\bar{V}$). In particular, two rows of $\bar{U}$ are perpendicular if the corresponding nodes lie in different row clusters. While on the column's side, if two nodes are in different column clusters, then the angle between the corresponding rows of $\bar{V}$ depends generally on the angles between the indicated rows in a "normalized" connectivity matrix $\tilde{B}^{\intercal}H\bar{\Sigma}^{-1}$. In Lemma 10, we provide understandable and reasonable conditions which lead to an explicit bound of the angle. Except for these facts, Lemma 9 essentially explains why a normalization step is needed in Algorithm 2 before the k-means. It is well-known that k-means clusters nodes together if they are close in the sense of Euclidean distance. The normalization step forces any two rows of $\bar{U}$ or $\bar{V}$ to lie in the same position if the corresponding nodes are in the same row cluster or column cluster. In such a way, the k-means could succeed when applied to the sample version singular vectors.

In Theorem 4 and 6, we have proved that the randomized adjacency matrices A^{rp} and A^{rs} concentrate around the population P under the ScBMs, where we actually did not make use of the explicit structure of P but only the facts that P is the population of A, and P is of rank K^y. Hence the same results hold here for the DC-ScBMs. Next, we use these results combining with Lemma 9 to analyze the clustering performance of RP-SsCC and RS-SsCC.

3.3.1 Performance of RP-SsCC in DC-ScBMs

Define

$$\kappa^y := \max_i \, (\tilde{\theta}_i^y)^{-2} \quad \text{and} \quad \kappa^z := \max_i \, (\tilde{\theta}_i^z)^{-2} \|(\tilde{B}_{*g_i})^{\intercal} H\bar{\Sigma}^{-1}\|_2^{-2}, \tag{3.10}$$

and

$$\eta(P) = \max_{k\neq l} \cos(\tilde{B}_{*k})^{\intercal} H\bar{\Sigma}^{-1}, (\tilde{B}_{*l})^{\intercal} H\bar{\Sigma}^{-1}), \tag{3.11}$$

where recall that H is the left singular matrix of $\tilde{B}$. The next theorem quantifies the clustering performance of RP-SsCC.

Theorem 11 *Suppose (C2) and (C3) hold and other parameter settings are identical with those in Theorem 4. Recall the terms defined in (3.10) and (3.11) and that the minimum and maximum non-zero singular value of P are γ_n and σ_n, respectively. Denote the sets of misclustered nodes with respect to the row and column clusters by $M^y \subseteq \{1, ..., n\}$ and $M^z \subseteq \{1, ..., n\}$, respectively. Then under DC-ScBMs, the following results hold for RP-SsCC.*

(1) With probability larger than $1 - 6r^{-r} - 6s^{-s} - 3n^{-\epsilon}$ for any $\epsilon > 0$, the misclustering rate with respect to the row clusters satisfies

$$\frac{|M^y|}{n} \leq c_9^{-1} \frac{\kappa^y K^y \alpha_n}{\gamma_n^2}. \tag{3.12}$$

provided that $\frac{n\kappa^y K^y \alpha_n}{\gamma_n^2 n_k^y} \leq c_9$ *for any* $k = 1, ..., K^y$ *for some constant* $c_9 > 0$.

(2) With probability larger than $1 - 6r^{-r} - 6s^{-s} - 3n^{-\epsilon}$ *for any* $\epsilon > 0$*, the misclustering rate with respect to the column clusters satisfies*

$$\frac{|M^z|}{n} \leq c_{10}^{-1} \frac{\kappa^z K^y \alpha_n}{(1-\eta(P))\gamma_n^2}. \tag{3.13}$$

provided that $\frac{n\kappa^z K^y \alpha_n}{(1-\eta(P))\gamma_n^2 n_k^z} \leq c_{10}$ *for any* $k = 1, ..., K^z$ *and some constant* $c_{10} > 0$.

The quantity κ_y (κ_z) can be thought of as the maximum node heterogeneity in sending (receiving) edges across all row (column) clusters, respectively. The quantity $\eta(P)$ indicates the minimum non-zero angles among the rows of the population singular vectors $\bar{V}$ (see (2) of Lemma 9), valued by cosine. (3.12) and (3.13) provide upper bounds for the proportion of the misclustered nodes with respect to row and column clusters, respectively. It can be seen that larger node degree heterogeneity may lead to poorer clustering performance. And different from the row clusters, the performance of the estimated column clusters additionally depend on $\eta(P)$. As expected, smaller $\eta(P)$ indicates better clustering performance. The conditions $\frac{n\kappa^y K^y \alpha_n}{\gamma_n^2 n_k^y} \leq c_9$ and $\frac{n\kappa^z K^y \alpha_n}{(1-\eta(P))\gamma_n^2 n_k^z} \leq c_{10}$ have the same effect and meaning with those in Theorem 5.

3.3.2 Performance of RS-SsCC in DC-ScBMs

The next theorem reflects the clustering performance of RS-SsCC.

Theorem 12 *Suppose (C2) holds and the other parameters are the same with those in Theorem 6. Recall that the minimum non-zero singular value of* P *is* γ_n *and recall* $\Phi(n, p, \alpha_n)$ *defined in (3.7). Denote the sets of misclustered nodes with respect to the row and column clusters by* $M^y \subseteq \{1, ..., n\}$ *and* $M^z \subseteq \{1, ..., n\}$*, respectively. Then under DC-ScBMs, the following results hold for RS-SsCC.*

(1) With probability larger than $1 - 2n^{-\nu} - \exp\Big(-c_6 np\big(1 + p^{1/4} \cdot \max(1, \sqrt{\frac{1}{p} - 1})^2\big)\Big)$ *for any* $\nu > 0$*, the misclustering rate with respect to the row clusters satisfies*

$$\frac{|M^y|}{n} \leq c_{11}^{-1} \frac{\kappa^y K^y \Phi^2(n, \alpha_n, p)}{\gamma_n^2 n}. \tag{3.14}$$

provided that $\frac{\kappa^y K^y \Phi^2(n,\alpha_n,p)}{\gamma_n^2 n_k^y} \leq c_{11}$ *for any* $k = 1, ..., K^y$ *and some constant* $c_{11} > 0$.

(2) With probability larger than $1 - 2n^{-\nu} - \exp\Big(-c_6 np\big(1 + p^{1/4} \cdot \max(1, \sqrt{\frac{1}{p} - 1})^2\big)\Big)$ *for any* $\nu > 0$*, the misclustering rate with respect to the column clusters satisfies*

$$\frac{|M^z|}{n} \leq c_{12}^{-1} \frac{\kappa^z K^y \Phi^2(n, \alpha_n, p)}{(1-\eta(P))\gamma_n^2 n}. \tag{3.15}$$

provided that $\frac{\kappa^z K^y \Phi^2(n,\alpha_n,p)}{(1-\eta(P))\gamma_n^2 n_k^z} \leq c_{12}$ *for any* $k = 1, ..., K^z$ *and some constant* $c_{12} > 0$.

We omit the proof of Theorem 12 since it is similar to that of Theorem 11. (3.14) and (3.15) provide upper bounds for the proportion of misclustered nodes with respect to the row clusters and column clusters, respectively. Similar to the results of the random projection paradigm, the clustering performance depends on the degree heterogeneity. And for the column cluster, it additionally depends on the minimum non-zero angles among the rows in the population singular vector $\bar{V}$.

4. Discussions

We discuss the theoretical aspects of the proposed randomized methods.

On the misclustering error rates and comparison with the state-of-the-art. Note that the misclustering error rates depend on one unknown parameter, namely, the minimum non-zero singular value of the population matrix. Here we specify the bounds and compare them with the state-of-the-art by consider the following four-parameter ScBM. The underlying number of row clusters and column clusters are the same, i.e., $K^y = K^z = K$. Each cluster has a balanced size n/K. For any pair of nodes (i, j), a directed edge from i to j is generated with probability α_n if the row cluster of i is identical to the column cluster of j, and with probability $\alpha_n(1-\lambda)$ otherwise. Formally,

$$P = YBZ^\intercal = Y(\alpha_n \lambda I_K + \alpha_n(1-\lambda)1_K 1_K^\intercal)Z^\intercal.$$

In this case, the minimum non-zero singular value of P is $n\alpha_n\lambda/K$ (Rohe et al., 2011), the row and column clusters has the same misclustering error rates, and there is no degree heterogeneity. In what follows, we examine the misclustering rate in Theorem 5 (Theorem 11) and Theorem 7 (Theorem 12), respectively. The misclustering error rates mentioned below have been summarized in Table 1.

- The bound in (3.3) reduces to $O(K^2/(n\alpha_n))$ under the four-parameter ScBM. If $\alpha_n = O(\text{log}n/n)$, then $O(K^2/(n\alpha_n))$ vanishes as n increases provided that $K = o(\sqrt{\text{log}n})$ and (C4) is automatically satisfied. While in Rohe et al. (2012) (see Corollary 4.1 therein), the misclustering rate is $o(K^3\text{log}n/\alpha_n^4)$, and $K = O(n^{1/4}/\text{log}n)$ is required to make the results hold. In addition, in Rohe et al. (2016) (see Corollary C.1 therein), the misclustering rate is $O(K^2\text{log}n/n)$ provided that α_n is fixed, which is not better than the $O(K^2/n)$ in our case.

- The bound in (3.8) simplifies to $O(K^2/(pn\alpha_n))$ provided that $p > 1/2$, which is tighter than those in Rohe et al. (2012) and Rohe et al. (2016).

The major reason why the randomized algorithms lead to even better misclustering error rates than the non-randomized algorithms do in Rohe et al. (2016) and Rohe et al. (2012) is that we derive the approximation bounds of $\|A^{\text{rp}} - P\|_2$ and $\|A^{\text{rs}} - P\|_2$ on the basis of the tightest concentration bound of $\|A - P\|_2$ (Lei and Rinaldo, 2015; Chin et al., 2015), which was originally developed using combinatorial arguments (Feige and Ofek, 2005). While Rohe et al. (2016) studied the misclustering error rates of the Laplacian-matrix-based spectral co-clustering, where they made use of the Bernstein-type bound and needed to additionally handle the degree matrix.

On the information-theoretic threshold and the optimal misclustering rate. As we have mentioned, the approximation error rates of the randomized adjacency matrices achieve the minimax optimal rates established in Zhang et al. (2016). We discuss the misclustering error rate in what follows. It should be noted that the original spectral (co-)clustering generally could not attain the minimax optimal misclustering error rate under SBMs (Zhang et al., 2016) except that the number of clusters $K = 2$ (Abbe et al., 2020), though Gao et al. (2017) developed a refined spectral clustering algorithm which could attain the optimal rates but with time complexity $O(n^3)$. Hence our randomized version inherits such limitations when $K > 2$. Next we examine the case $K = 2$ in more detail.

Consider the following two blocked symmetric ScBM (i.e., SBM) $\text{SBM}(n, a\frac{\log n}{n}, b\frac{\log n}{n}, J)$. Let n be even, $J \subseteq [n]$ with $|J| = n/2$. Each entry $A_{ij} (i < j)$ of the symmetric adjacency matrix $A = (A_{ij})$ is independently generated as $\mathbb{P}(A_{ij} = 1) = a\frac{\log n}{n}$ if $i \sim j$, and $\mathbb{P}(A_{ij} = 1) = b\frac{\log n}{n}$ otherwise, where $i \sim j$ means that $i \in J, j \in J$ or $i \in J^c, j \in J^c$. Under this regime, the population adjacency matrix P is a rank-2 matrix with two nonzero eigenvalues $\lambda_1^* = (a+b)\log n/2$ and $\lambda_2^* = (a-b)\log n/2$ whose corresponding eigenvectors are $u_1^* = \frac{1}{\sqrt{n}}\mathbb{I}_n$ and $u_2^* = \frac{1}{\sqrt{n}}\mathbb{I}_J - \frac{1}{\sqrt{n}}\mathbb{I}_J^c$.

Let $z \in \mathbb{R}^n$ be the true labels, specifically, $z_i = 1$ if $i \in J$ and $z_i = -1$ if $i \in J^c$. Then, the clustering aims to estimate the unknown z by $\hat{z} \in \mathbb{R}^n$. Under the $\text{SBM}(n, a\frac{\log n}{n}, b\frac{\log n}{n}, J)$, Abbe et al. (2015) and Mossel et al. (2015) proved that exact recovery ($\hat{z}$ equal to z or $-z$ with probability tending to 1) is information-theoretically possible if and only if $\sqrt{a} - \sqrt{b} > \sqrt{2}$. On the other hand, when the exact recovery is impossible, that is $\sqrt{a} - \sqrt{b} \in (0, \sqrt{2}]$, Zhang et al. (2016) provided the following minimax misclustering error rate (Abbe et al., 2020)

$$\inf_{\hat{z}} \sup \mathbb{E}r(\hat{z}, z) = \exp\left(-(1+o(1)) \cdot (\sqrt{a} - \sqrt{b})^2 \frac{\log n}{2}\right),$$

where the misclustering rate $r(\hat{z}, z)$ is defined as

$$r(\hat{z}, z) = \min_{s \in \{\pm 1\}} n^{-1} \sum_{i=1}^{n} \mathbb{I}_{\hat{z}_i \neq s z_i}. \tag{4.1}$$

By developing technical tools for entry-wise perturbation bound of eigenvectors, Abbe et al. (2020) proved that the vanilla spectral method based on the adjacency matrix (first compute u_2, the eigenvector of A corresponding to its second largest eigenvalue λ_2; then set $\hat{z} = \text{sgn}(u_2)$) can achieve exact recovery when it is information-theoretic possible, and can attain the optimal minimax misclutering rate otherwise.

To see whether such optimal results could be obtained or how much distortion would be produced by the randomized spectral clustering algorithms. We consider the vanilla spectral method based on the sparisified adjacency matrix A^{rs} (see (2.1)). With a slight abuse of notation, let u_2 be the eigenvector of A^{rs} associated with its second largest λ_2, and the estimated labels are obtained by $\hat{z} = \text{sgn}(u_2)$. We show in the following theorem that the vanilla spectral method based on A^{rs} could yield near-optimal results while bringing some distortion, which is caused by the random sampling of A. In particular, the exact recovery succeeds when $\sqrt{a} - \sqrt{b} > \sqrt{2/p}$, which is a little more stringent than the information-theoretic threshold $\sqrt{2}$. Moreover, when $\sqrt{a} - \sqrt{b} \in (0, \sqrt{2/p}]$, the misclustering error rate turns out to be $n^{-(1+o(1))(p(\sqrt{a}-\sqrt{b})^2/2)}$, which is also slightly inferior to the minimax optimal rate $n^{-(1+o(1))((\sqrt{a}-\sqrt{b})^2/2)}$.

Theorem 13 *(1) If $\sqrt{a}-\sqrt{b} > \sqrt{2/p}$, then there exists $\eta = \eta(a, b, p) > 0$ and $s \in \{\pm 1\}$ such that with probability $1 - o(1)$,*

$$\sqrt{n} \min_{i\in[n]} s z_i (u_2)_i \geq \eta.$$

As a consequence, the vanilla spectral method based on the sparsified matrix A^{rs} achieves exact recovery.

(2) Let the misclustering rate $r(\hat{z}, z)$ be defined in (4.1). If $\sqrt{a}-\sqrt{b} \in (0, \sqrt{2/p}]$, then

$$\mathbb{E}r(\hat{z}, z) \leq n^{-(1+o(1))(p(\sqrt{a}-\sqrt{b})^2/2)}.$$

On the extensions to rank-deficient ScBMs. In this work, we mainly consider the ScBMs (DC-SsBMs) with $B_{K^y\times K^z}$ being of full row rank (recall that $K^y \leq K^z$), where the coupling of the target rank and the smaller target cluster size would make the result nice and interpretable. In real applications, however, the number of clusters may be larger than the dimension (Tang et al., 2022). We argue that all the results could be generated to rank-deficient case. We provide in the following theorem some inspirations.

Theorem 14 *Consider an ScBM parameterized by (Y, Z, B), where B is rank deficient with $\operatorname{rank}(B) = K' < K^y$. $P = YBZ^{\intercal}$ is the population adjacency matrix with its SVD being $\bar{U}_{n\times K'}\bar{\Sigma}_{K'\times K'}\bar{V}^{\intercal}_{K'\times n}$. Define $\Delta_y = \operatorname{diag}(\sqrt{n^y_1}, ..., \sqrt{n^y_{K^y}})$, $\Delta_z = \operatorname{diag}(\sqrt{n^z_1}, ..., \sqrt{n^z_{K^z}})$, $\bar{B} = B\Delta_z$, and denote the SVD of $\Delta_y B\Delta_z$ by $L_{K^y\times K'}D_{K'\times K'}R^{\intercal}_{K'\times K^z}$. If $Y_{i*} = Y_{j*}$, then $\bar{U}_{i*} = \bar{U}_{j*}$; otherwise if the rows of $\bar{B}$ are mutually distinct such that*

$$\min_{1\leq k\neq l\leq K^y} \|\bar{B}_{k*} - \bar{B}_{l*}\|_2 \geq \nu_n,$$

and $0 < \bar{\Sigma}_{ii} \leq \mu_n$, then

$$\|\bar{U}_{i*} - \bar{U}_{j*}\|_2 = \left\| \frac{L_{g^y_i *}}{\sqrt{n^y_{g^y_i}}} - \frac{L_{g^y_j *}}{\sqrt{n^y_{g^y_j}}} \right\|_2 \geq \frac{\nu_n}{\mu_n}.$$

Theorem 14 shows that the population left singular vectors are well-separated provided that the rows of $\bar{B}$ are mutually distinct. Therefore, the analysis on the clustering performance naturally follows with extra conditions and notations.

5. Numerical studies

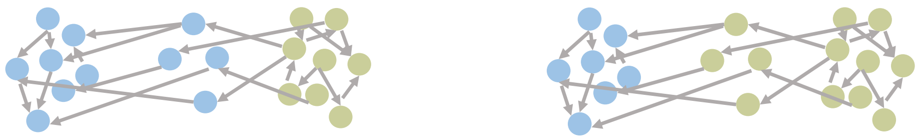
We evaluate the finite sample performance of RP-SCC (RP-SsCC) and RS-SCC (RS-SsCC), and compare them with SCC (SsCC).

In accordance with the theoretical results, we use the following two measures to examine the empirical performance of the three methods. The first is the approximation error, defined by $\|\tilde{A} - P\|_2$, where $\tilde{A}$ can be A, A^{rp}, or A^{rs}. The second is the misclustering error rate with respect to the row clusters and column clusters, defined by $\min_{J\in E_{K^y}} \frac{1}{2n}\|\tilde{Y}J-Y\|_0$ and $\min_{J\in E_{K^z}} \frac{1}{2n}\|\tilde{Z}J-Z\|_0$, respectively, where E_{K^y} and E_{K^z} stand respectively for the sets of K^y and K^z dimensional permutation matrices, $\tilde{Y}$ can be the estimated row membership matrix of RP-SCC, RS-SCC and SCC, and $\tilde{Z}$ is similarly defined with respect to the column clusters. We consider 8 model set-ups, each of which is designed to imitate typical directed network structures. For each model set-up, we consider two parameter settings for the link probability matrix B. One corresponds to the fixed B setting, that is to say, each element of B is kept a constant as the network size n increases. The other corresponds to the more challenging high-dimensional setting, namely, each element of B is vanishing as n increases, which is a more realistic setting. For the readability, model set-ups 3-8 and their topological representations, and the matrix representations of 8 models are relegated to Section E of the Appendix.

Model set-up 1 (Networks with 'transmission' nodes) In such networks, there exists a set of nodes which only receive edges from one set of nodes (set 1) and send edges to another set of nodes (set 2), and hence termed as 'transmission nodes'. While for nodes from set 1 or set 2, except those edges linked to 'transmission nodes', they send to and receive from nodes within their own set. As a result, the sending clusters and receiving clusters are different. In this model-set up, $K^y = K^z = 2$, and we consider the following two cases for the link probability matrix B:

$$B_1 := \begin{bmatrix} 0.05 & 0 \\ 0 & 0.05 \end{bmatrix}, \quad B_2 := \begin{bmatrix} \frac{\log n}{n} & 0 \\ 0 & \frac{\log n}{n} \end{bmatrix}.$$

Figure 1 shows the topological example of model set-up 1.



(a) Sending clusters

(b) Receiving clusters

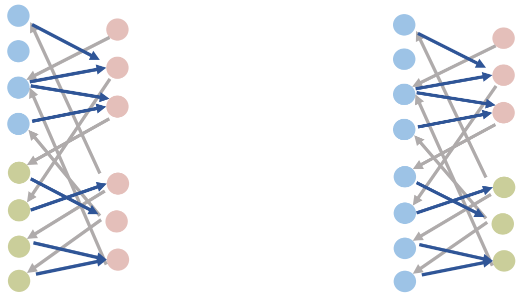
Figure 1: Illustration for networks under model set-up 1. Colors indicate clusters.

Model set-up 2 (Bipartite networks) In such networks, edges only exists between nodes from different clusters. The sending and receiving clusters could be different by incorporating different sending and receiving patterns. In this model-set up, $K^y = K^z = 3$, and we consider the following two cases for the link probability matrix B:

$$B_1 := \begin{bmatrix} 0 & 0.2 & 0 \\ 0 & 0 & 0.2 \\ 0.1 & 0 & 0 \end{bmatrix}, \quad B_2 := \begin{bmatrix} 0 & \frac{1.5\log n}{n} & 0 \\ 0 & 0 & \frac{1.5\log n}{n} \\ \frac{\log n}{n} & 0 & 0 \end{bmatrix}.$$

Figure 2 shows the topological example of model set-up 2.

In the random projection scheme, the oversampling parameter is 10, the power parameter is 2, and the test matrices are generated with i.i.d. standard Gaussian entries. In the



(a) Sending clusters

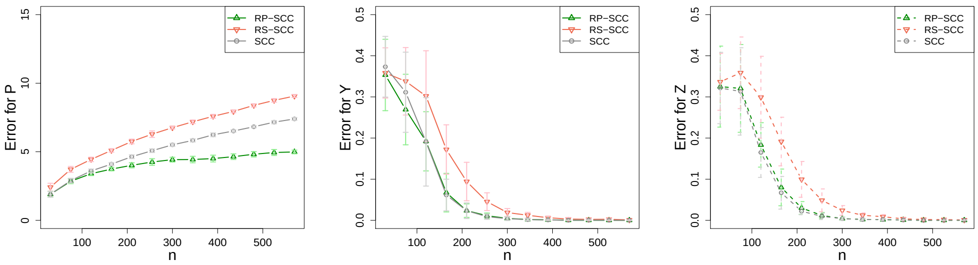
(b) Receiving clusters

Figure 2: Illustration for networks under model set-up 2. Colors indicate clusters.

random sampling scheme, the sampling rate is 0.7. We use the R package irlba to compute the singular vector iteratively after the sampling step. We evaluate how the approximation error for P, the estimation error for the row clusters Y, and the estimation error for the column clusters Z alter as the network size n increases, respectively. For the sake of readability, we only display the averaged results together with the standard deviations over 20 replications of model set-up 1 (case1) and 2 (case1) in Figure 3-4. The remaining results can be found in Section E of the Appendix.

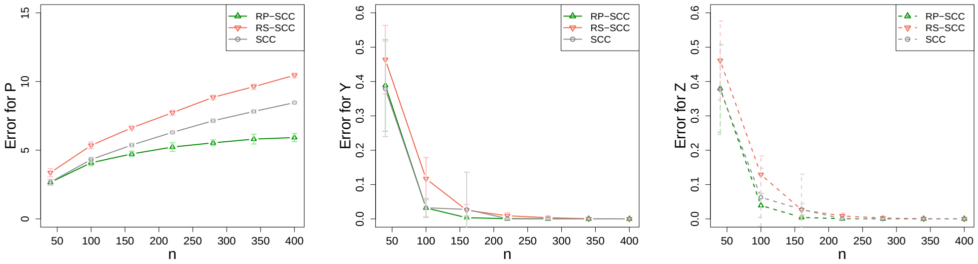
We can make the following observations from the results. First, for the approximation error, the three methods show similar tendencies as the sample size increases, all grow at rate $o(n)$, indicating that $\tilde{A}$'s concentrate around the population P. The slight differences come from pre-constants because we have shown in Section 3 that $\|\tilde{A} - P\|_2$ attains the order-wise minimax optimal rate with large probability. Second, for the misclustering error, all three methods yield decreasing misclustering rates as n increases. The standard deviations generally decrease as n increases, though bad clustering performances would lead to small deviations of all methods when n is relatively small. The RP-SCC (RP-SsCC) and RS-SCC (RS-SsCC) perform just slightly worse than SCC (SsCC), especially when n is large, which is the focus of this work. In particular, RP-SCC (RP-SsCC) generally leads to better clustering performances than RS-SCC (RS-SsCC) does, and more interestingly, the former are even more stable than SCC (SsCC); see Figure 21 and 29 in the Appendix for example. In addition, one may find that the performance of the estimated row clusters are better than that of the column clusters in mode set-up 3 and 4, which is because $K^y < K^z$ therein and is consistent with our theoretical results. Finally, we note that the misclustering error might decrease slowly. This because when the link probability is $O(\log n/n)$, the misclustering rate is as slow as $O(\log n)$, indicated by our theory. In addition, the theoretical bounds hold in the sense of probability.

We additionally compare the RP-SCC with the randomized block Krylov method (Musco and Musco, 2015) based spectral co-clustering, which utilizes Krylov subspaces to obtain the approximated row spaces of A. As mentioned in Woodruff and Yasuda (2022), this method is the *state-of-the-art* general case algorithm for finding low rank approximation to a matrix. For simplicity, we use RB-Krylov to denote both the approximate SVD method in Musco and Musco (2015) and its corresponding spectral co-clustering. The power parameter in RB-Krylov is also set to be 2. Note that under this parameter set-up, the time complexities



(a) (b) (c)

Figure 3: Simulation results of case 1 under model set-up 1.



(a) (b) (c)

Figure 4: Simulation results of case 1 under model set-up 2.

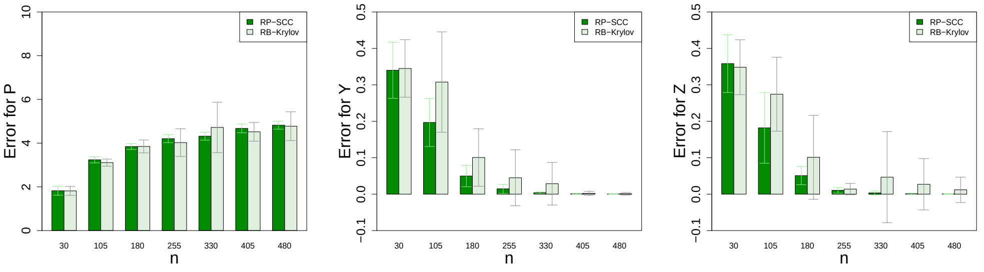
of RB-Krylov and RP-SCC is of the same order. We compare the approximation error and misclustering error of two methods under the case 1 of model set-ups 1 and 2. The bar plot together with the standard deviations over 20 replications are displayed in Figure 5-6. It shows that for the approximation error, RB-Krylov is better than or similar to RP-SCC. While for the misclustering error, our RP-SCC is better than RB-Krylov for $q = 2$. We note that when q increases, the clustering performance of RB-Krylov tends to be similar with that of RP-SCC. Other models also lead to similar results, we omit them for the sake of space. In the next section, we will show that under our experimental set-up, RP-SCC is experimentally faster than RB-Krylov with better or comparable clustering accuracy.

6. Real data analysis

We empirically evaluate the randomized spectral co-clustering algorithms on real network datasets, considering both the clustering accuracy and the computational efficiency.

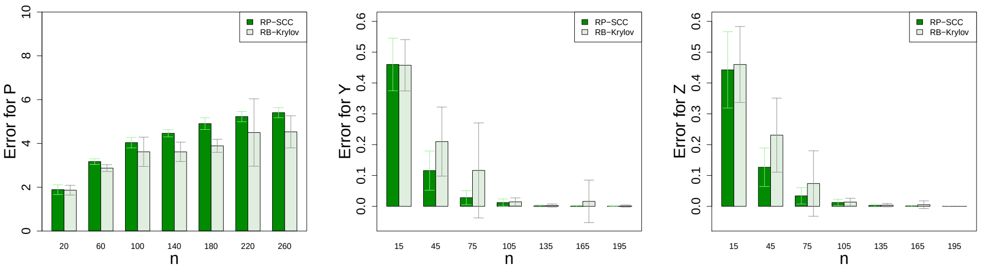
6.1 Accuracy comparison on small-scale networks

We compare the clustering performance of the proposed methods RP-SCC (RP-SsCC) and RS-SCC (RS-SsCC) with SCC (SsCC), two deterministic iterative-algorithm-based spec-



(a) (b) (c)

Figure 5: Comparison of RP-SCC and RB-Krylov under case 1 of model set-up 1.



(a) (b) (c)

Figure 6: Comparison of RP-SCC and RB-Krylov under case 1 of model set-up 2.

tral co-clustering algorithms, denoted by svds and irlba, and the randomized block Krylov method mentioned in Section 5, denoted by RB-Krylov. The latter three methods use the corresponding approximate SVD algorithms coupled with the k-means to obtain the estimated co-clusters. Specifically, svds and irlba use respectively the implicitly restarted Lanczos algorithm (Calvetti et al., 1994) (svds in R package RSpectra (Qiu and Mei, 2019)) and the augmented implicitly restarted Lanczos bidiagonalization algorithm (Baglama and Reichel, 2005) (irlba in R package irlba (Baglama et al., 2019)) to compute the SVD in SCC (SsCC). RB-Krylov uses randomized block Krylov iteration method to find approximate SVD (Musco and Musco, 2015). To fix ideas, we call the five methods including RP-SCC, RS-SCC, svds, irlba and RB-Krylov the approximate methods for SCC (SsCC) in what follows. For RP-SCC (RP-SsCC), the oversampling parameter is 10, the power parameter is 2, and the test matrices are generated with i.i.d. standard Gaussian entries. For RS-SCC (RS-SsCC), the sampling rate is 0.8 and we use the R package irlba to compute the singular vector iteratively after the sampling step. For svds and irlba, the tolerance parameter is set to be 10^{-5}. For RB-Krylov, the power parameter is also 2, and the test matrices are generated with i.i.d. standard Gaussian entries.

We consider two directed networks, one is the statisticians citation network (Ji and Jin, 2016) and the other is the European email network (Yin et al., 2017).

Statisticians citation network This network describes the citation relationships between statisticians who published at least one paper in four top journals of statistics from 2003 to the first half of 2012. If author i cited at least one paper written by author j, then there is a directed edge from node i to node j. The largest component of this network has 2,654 nodes and 21,568 edges.

To decide the target rank, we evaluate the top 50 singular values of the associated adjacency matrix A. As indicated in Figure 7, there is an eigen-gap between the third and fourth singular values, suggesting that the target rank is 3 (Rohe et al., 2016). Before doing clustering, we first evaluate the similarities and patterns of singular vectors found by different methods. Define the *movement score* for node i to be the Euclidean norm of difference between the ith row of the right and left (approximated) singular vectors. Figure 8 shows the histograms of the movement scores by six methods. The first five histograms turn out to be very similar, indicating that all methods except RB-Krylov lead to similar singular vectors. While the singular vectors found by RB-Krylov is not accurate under our parameter set-up. In addition, the asymmetric nature of this citation network is again evidenced since there exists nodes with movement score away from zero.

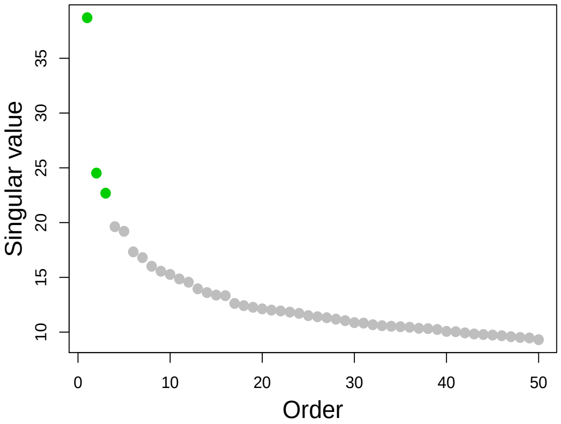
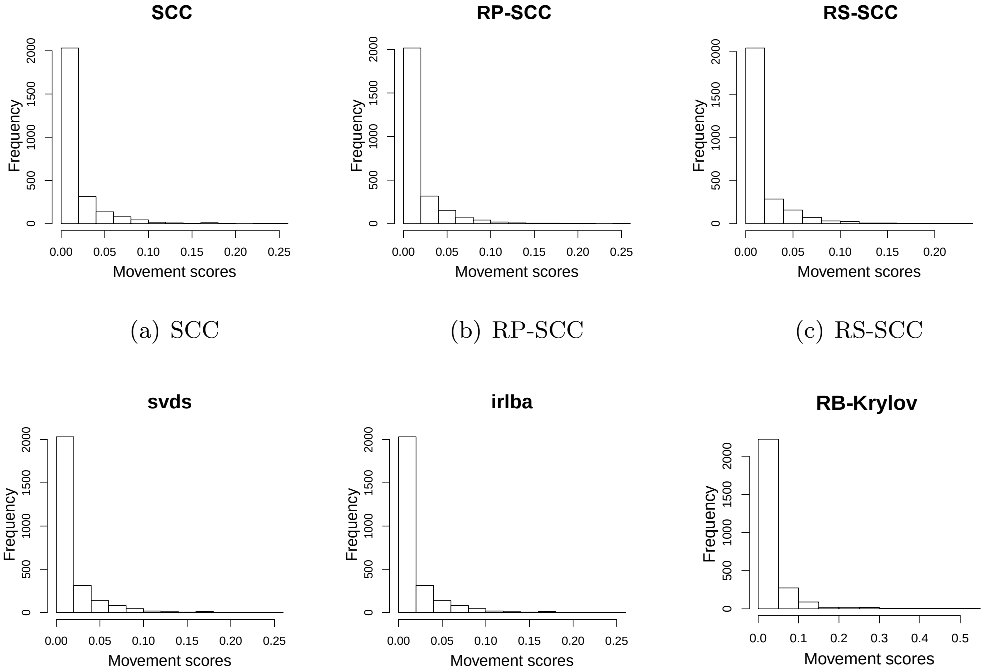


Figure 7: The top 50 singular values of the adjacency matrix of the statisticians citation network.

Now, we evaluate the clustering performance of these methods. We actually test two set of algorithms, one based on SCC (Algorithm 1) and the other based on SsCC (Algorithm 2). We use the *silhouette method* to select respectively the target number of row clusters (K^y) and column clusters (K^z). In our setting, the silhouette method typically evaluates the clustering performance of the k-means output with the original left and right singular vectors as the input with respect to different k via the average silhouette width, for which larger value indicates better performance. With the selected number of row clusters and column clusters, we compare the six methods by computing the ARI (Hubert and Arabie, 1985; Manning et al., 2010) between the SCC (SsCC) and the five SCC (SsCC)-based



(a) SCC (b) RP-SCC (c) RS-SCC

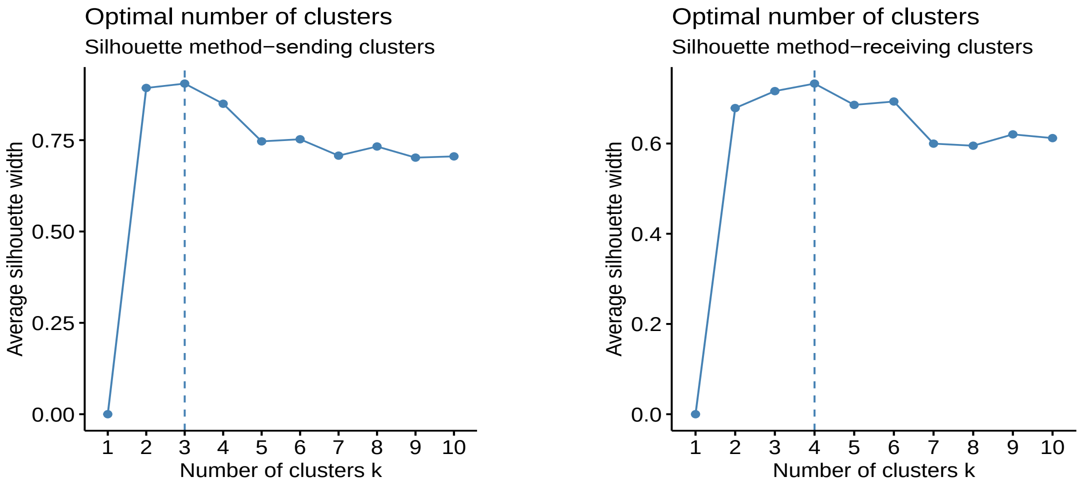
(d) svds (e) irlba (f) RB-Krylov

Figure 8: Histogram of movement scores of different methods for the statisticians citation network.

approximate methods, respectively. Larger ARI indicates that the approximate method lead to more consistency results with that of SCC (SsCC).

For the SCC-based algorithms, the number of sending clusters and receiving clusters turn out to be 3 and 4, respectively; see Figure 9 for details. The box plots of relative ARI over 20 replications are shown in Figure 10. For both side of clusters, the average ARI's of all methods except RB-Krylov are larger than 0.87, showing that these methods could well approximate the SCC, though RP-SCC and RS-SCC are slightly inferior to the deterministic iterative-algorithm-based methods. In addition, the RB-Krylov is less stable than other methods. We note that the performance of RB-Krylov can be improved by enlarging the power parameter but with efficiency sacrifice. We also display the embedding of nodes provided by their corresponding components of the first three left singular vectors in Figure 11, where colors indicate clusters. The results corresponding to the right singular vectors are relegated to Section E of the Appendix. We see that all methods except RB-Krylov yield similar clusters up to certain rotations of singular vectors and clusters. In addition, the sending clusters (authors cited by others) are more concentrated than the receiving clusters (authors citing others), which agrees with the common logic.

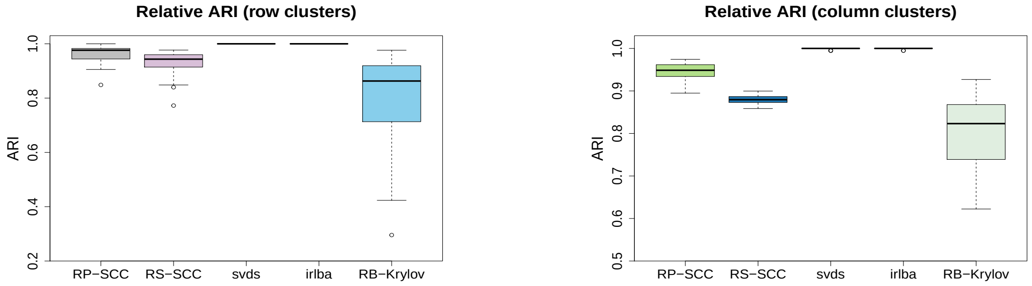
For the SsCC-based algorithms, Figure 12 shows that the optimal number of sending clusters and receiving clusters are both 5. Recall that the target rank is 3, and thus this set-up corresponds to the rank-deficient model, which has been discussed in Section 4.



(a) Sending clusters

(b) Receiving clusters

Figure 9: Optimal number of clusters of the statisticians citation network selected by the average silhouette method based on the SCC.



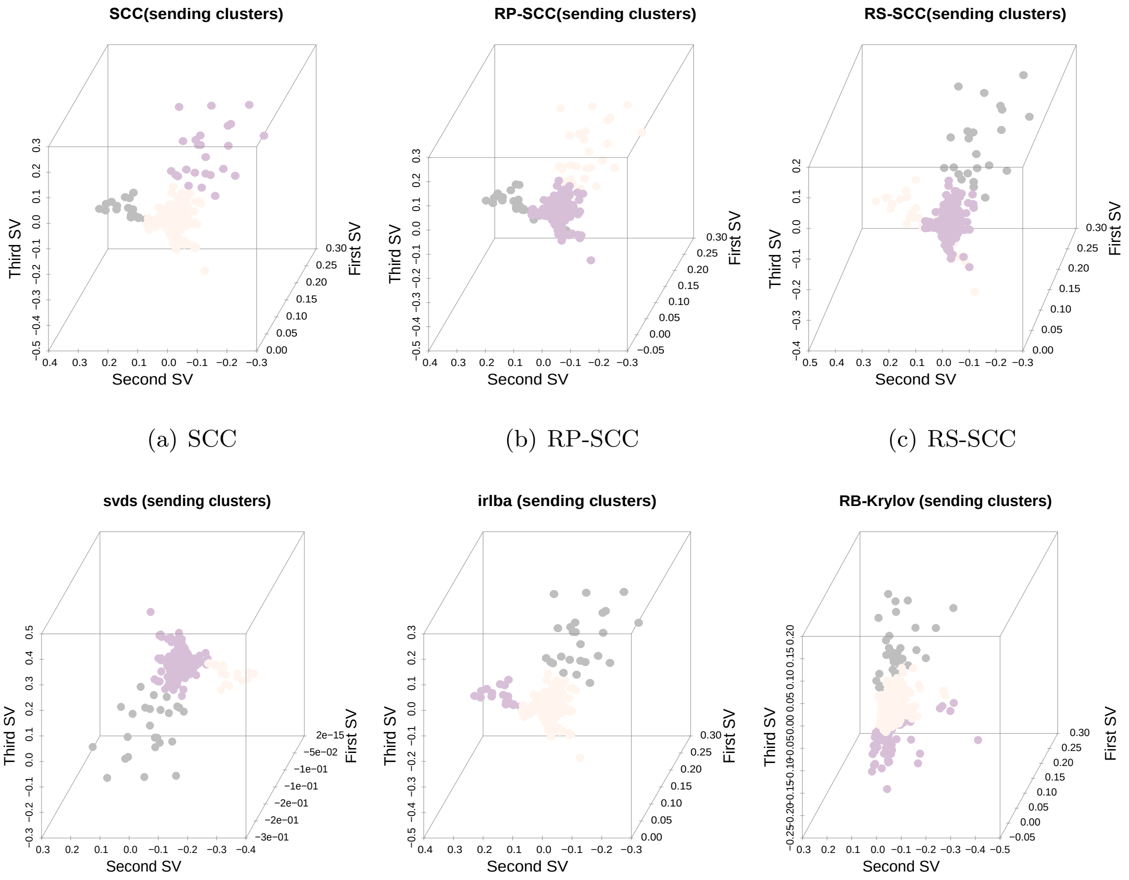
(a) Sending clusters

(b) Receiving clusters

Figure 10: Relative ARI between the SCC and SCC-based five approximate methods on the statisticians citation network.

Figure 13 displays the box plots of relative ARI. For both side of clusters, it turns out that RP-SsCC performs best, followed by RB-Krylov, and other three methods seem to perform poorly. This does not contradict with the results of SCC-based algorithms. Because the normalization operator is not stable to noise, two close singular vectors could be far from each other after normalization. Our theory also indicate the hardness of clustering under DC-ScBMs. Nevertheless, RP-SsCC shows great clustering performance and certain degree of robustness.

European email network This network was generated using the email data from a large European research institution. If person i sent at least one email to person j, then there is a



(a) SCC (b) RP-SCC (c) RS-SCC

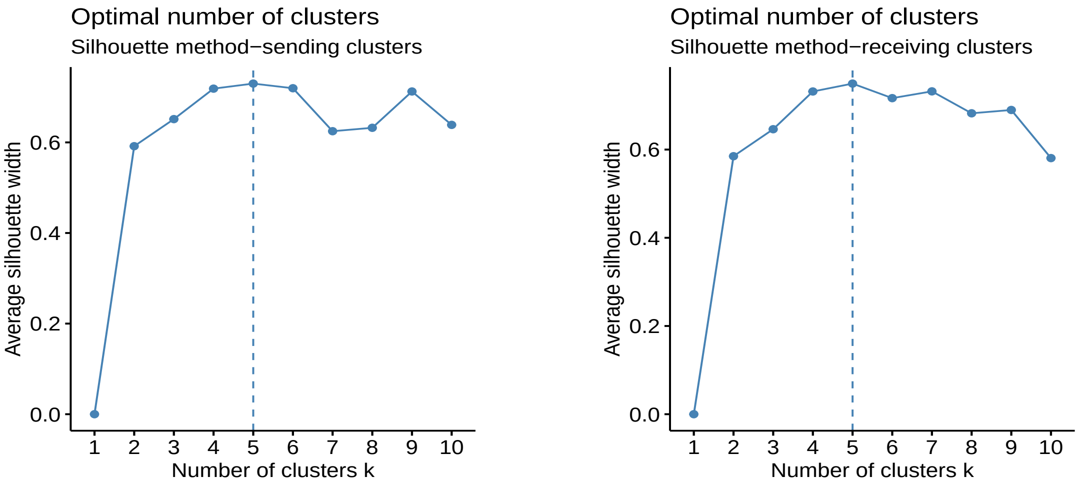
(d) svds (e) irlba (f) RB-Krylov

Figure 11: Sending clusters of the statisticians citation network detected by SCC and five SCC-based approximate algorithms.

directed edge from person i to j. The largest component of this network has 986 nodes and 24,929 edges. All treatments are similar to that for the statisticians citation network, hence we mainly discuss the results. All the mentioned figures are in Section E of the Appendix.

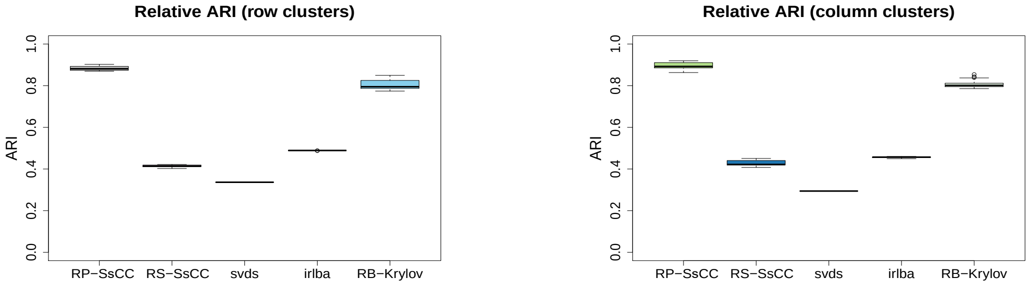
We choose the target rank to be 2 by evaluating the eigen-gap of the original singular values (Figure 36). The histogram of movement score (Figure 37) shows that all methods except RB-Krylov lead to similar singular vectors.

For the SCC-based methods, the optimal number of both sending clusters and receiving clusters turn out to be 2 (Figure 38). The average relative ARI of SCC-based methods are all above 0.9 (Figure 39), suggesting that approximate methods are comparable to SCC, though RB-Krylov is inferior to the other four methods. The two-dimensional embedding of estimated singular vectors (Figure 40 and 42) show similar pattern across different methods except RB-Krylov.



(a) Sending clusters (b) Receiving clusters

Figure 12: Optimal number of clusters of the statisticians citation network selected by the average silhouette method based on the SsCC.



(a) Sending clusters (b) Receiving clusters

Figure 13: Relative ARI between the SsCC and SsCC-based five approximate methods on the statisticians citation network.

For the SsCC-based methods, the optimal number of both sending clusters and receiving clusters turn out to be 4 (Figure 41), thus corresponding to rank-deficient models. Regarding the clustering performance (Figure 43), RP-SsCC, RS-SsCC and irlba turn out to be superior than the other two methods. The reason that lead to the deterioration of approximate methods might due to the noise accumulation of the normalization step therein.

6.2 Time comparison and accuracy evaluation on large-scale networks

The main barrier that hinders SCC to handle large-scale directed networks is the SVD computation. Now we compare the computational time of six approximate SVD methods (implementations) including the random-projection-based and random-sampling-based SVD studied in this paper, denoted respectively by RP and RS for short, the deterministic iterative methods svds (Calvetti et al., 1994; Qiu and Mei, 2019) and irlba (Baglama and Reichel, 2005; Baglama et al., 2019), the randomized block Krylov method RB-Krylov (Musco and Musco, 2015), and also the random-projection-based method (Halko et al., 2011) implemented in R package rsvd (Erichson et al., 2019)), denoted by rsvd. To the best of our knowledge, the fast implementation of RB-Krylov is lacked. For convenience and fair comparisons, we provide an efficient implementation of RB-Krylov in R package RandClust.

We examine five real networks with their number of nodes ranging from more than seventy thousands to more than two millions. Table 3 summarizes the basic information of each network, where the target rank means the number of singular vectors to be computed after we examine the (approximate) eigen-gap of the corresponding adjacency matrices.

Table 3: A summary of the five real large-scale networks.

Data	**No. of nodes**	**No. of edges**	**Target rank**
Epinions social network (Richardson et al., 2003)	75,877	508,836	3
Slashdot social network (Leskovec et al., 2009)	77,360	905,468	5
Berkeley-Stanford web network (Leskovec et al., 2009)	654,782	7,499,425	4
Wikipedia top categories network(Yin et al., 2017)	1,791,489	28,511,807	5
Wikipedia talk network (Leskovec et al., 2010)	2,388,953	5,018,445	3

For svds and irlba, the tolerance parameter is set to be 10^{-5}. For RP, the power parameter is 1 and the oversampling parameter is 5, which are adequate to improve the approximation quality (Halko et al., 2011). For RS, the sampling parameter is 0.7, and irlba is used after the sampling procedure. For RB-Krylov, the power parameter is 1. For rsvd, the power and oversampling parameters are both the same with those in RP. A machine with Intel Core i9-9900K CPU 3.60GHz, 32GB memory, and 64-bit WS operating-system, and R version 4.2.2 is used for all computations. Table 4 shows the median time (milliseconds) of each method for computing the SVD of the corresponding adjacency matrices of five real networks over 20 replications. For RS, we report the time including and excluding the sampling procedure, respectively.

It should be noted that the full SVD failed in all five cases. Among the compared six approximate methods for partial SVD, RP is faster than other methods considered on all five networks. In particular, RP is faster than the random-projection-based methods RB-Krylov and rsvd. RS is generally faster than rsvd and the baseline deterministic iterative method irlba. In particular, RS is comparable to or faster than irlba even when the sampling time is included. This provides evidence for the computational superiority of randomized methods.

Apart from computational time, we also compare the ARI between each pair of methods, evaluating their similarity of clustering results. Note that the silhouette method and other related methods for selecting the target number of clusters often fail on large-scale networks. Thus we choose the target number of sending and receiving clusters as the target rank roughly. Also note that we have provided evidence that RP-SsCC perform more simi-

Table 4: Median time (milliseconds) of each method for computing the SVD of the corresponding adjacency matrix of five real networks over 20 replications, where for RS, the time with the sampling time included and excluded (shown in the parentheses) are reported, respectively.

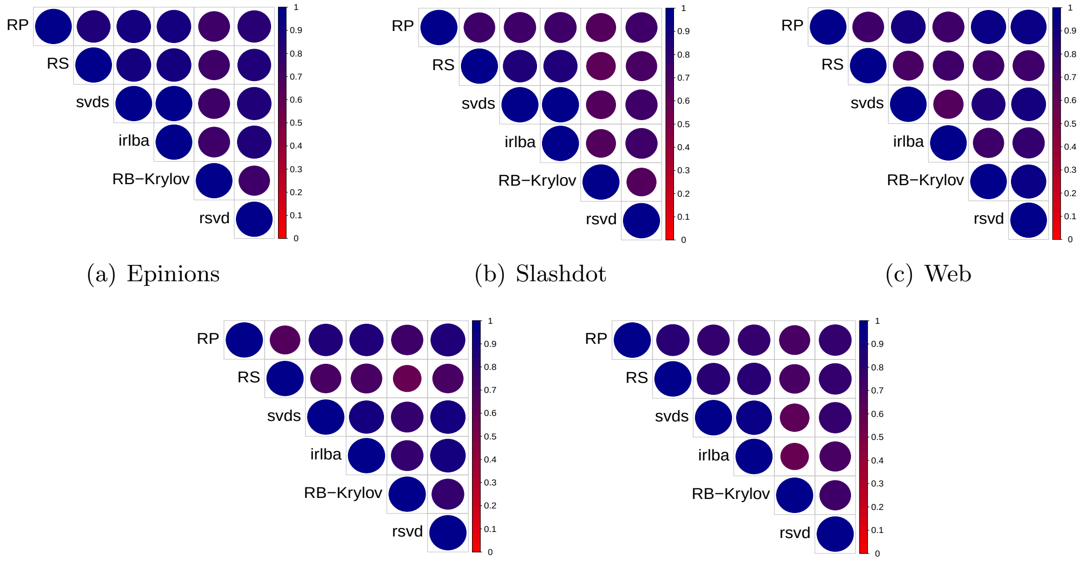
Data	RP	RS	svds	irlba	RB-Krylov	rsvd
Epinions social network	27.44	76.57(71.19)	43.71	79.00	30.72	138.19
Slashdot social network	50.16	110.01(100.94)	73.38	111.40	62.42	205.50
Berkeley-Stanford web network	431.55	844.82(770.56)	693.08	830.23	516.13	1506.53
Wikipedia top categories network	2844.67	4605.77(4327.97)	3593.60	4865.36	3121.93	7928.02
Wikipedia talk network	1067.70	1721.48(1658.13)	1593.60	1678.38	1249.12	4833.77

larly to SsCC than the other SsCC-based methods (RS-SsCC, svds, irlba and RB-Krylov) on small-scale networks. And rsvd and RP-SsCC are both based on Halko et al. (2011). Hence we here focus on comparing the SCC-based methods, denoted for short by RP, RS, svds, irlba, RB-Krylov and rsvd. Figure 14 and 15 show the averaged ARI and the corresponding standard deviations associated with sending clusters over 20 replications, respectively. The results corresponding to the receiving clusters are relegated to Section E of the Appendix. We observe that our randomized methods, especially RP, perform similarly to the baseline deterministic iterative methods irlba and svds with small standard deviations. Among the random-projection-based methods, RP performs similarly to rsvd and better than RB-Krylov, if one focus on the relative ARI with the baseline deterministic methods irlba and svds.

Overall, RP and RS show computational superiority while maintaining satisfactory clustering performance on tested real networks. In real applications, one can balance the accuracy and efficiency via changing the hyper-parameters according to the problems faced.

7. Conclusion

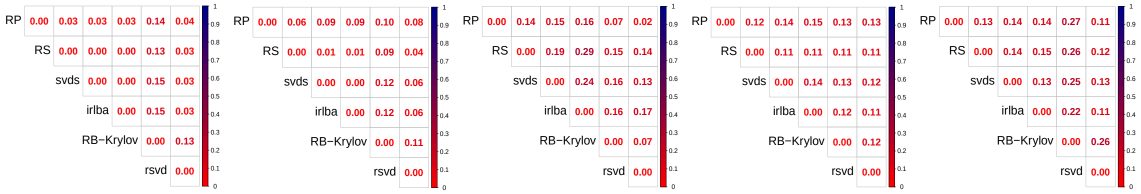
In this paper, we studied how randomization can be used to speed up the spectral co-clustering algorithms for co-clustering large-scale directed networks and how well the resulting algorithms perform under specific network models. In particular, the random-projection-based and random-sampling-based spectral co-clustering algorithms were derived. The clustering performance of these two algorithms was studied under the ScBMs and the DC-ScBMs, respectively. The theoretical bounds are high-dimensional in nature and easy to interpret. The theoretical optimality and possible extensions of the models were also discussed. We numerically compared our algorithms with fast iterative methods and other two randomized methods (implementations) for computing the SVD. It turns out that the our algorithms are faster than or comparable to the compared methods at the same time maintaining satisfactory clustering performance. We developed a publicly available R package RandClust for better usability of the proposed methods.



(a) Epinions (b) Slashdot (c) Web

(d) Wiki-cat (e) Wiki-talk

Figure 14: The pairwise comparison of the sending clusters of six methods on five large-scale networks. The relative clustering performance are measured by ARI. Larger ARI, i.e., larger circles in the figure, indicates that the clustering results of the two associated methods are more close.



(a) Epinions

RP	0.00	0.03	0.03	0.03	0.14	0.04
RS		0.00	0.00	0.00	0.13	0.03
svds			0.00	0.00	0.15	0.03
irlba				0.00	0.15	0.03
RB−Krylov					0.00	0.13
rsvd						0.00

(b) Slashdot

RP	0.00	0.06	0.09	0.09	0.10	0.08
RS		0.00	0.01	0.01	0.09	0.04
svds			0.00	0.00	0.12	0.06
irlba				0.00	0.12	0.06
RB−Krylov					0.00	0.11
rsvd						0.00

(c) Web

RP	0.00	0.14	0.15	0.16	0.07	0.02
RS		0.00	0.19	0.29	0.15	0.14
svds			0.00	0.24	0.16	0.13
irlba				0.00	0.16	0.17
RB−Krylov					0.00	0.07
rsvd						0.00

(d) Wiki-cat

RP	0.00	0.12	0.14	0.15	0.13	0.13
RS		0.00	0.11	0.11	0.11	0.11
svds			0.00	0.14	0.13	0.12
irlba				0.00	0.12	0.11
RB−Krylov					0.00	0.12
rsvd						0.00

(e) Wiki-talk

RP	0.00	0.13	0.14	0.14	0.27	0.11
RS		0.00	0.14	0.15	0.26	0.12
svds			0.00	0.13	0.25	0.13
irlba				0.00	0.22	0.11
RB−Krylov					0.00	0.26
rsvd						0.00

Figure 15: The standard deviations corresponding to the pairwise ARI of row clusters.

In this work, we focused on the pure spectral clustering without regularization or other refinements. The current theoretical results might be further improved if one uses refined spectral clustering as the starting algorithm and control its time complexity simultaneously. See Qin and Rohe (2013); Gao et al. (2017) for example. Note that the numbers of clusters were assumed to be known in the theoretical analysis. It would be important to study the selection of target cluster numbers (Ma et al., 2021), especially in an efficient way. In addition, it would be interesting to generalize the current framework to bipartite networks (Zhou and Amini, 2019), multi-layer networks (Lei et al., 2020), etc. Finally, we used the adjacency matrix as the input of spectral co-clustering. It is interesting to study how to generalize the techniques to Laplacian-matrix-based clustering (Rohe et al., 2016).

Acknowledgments

The authors are grateful to the editor and reviewers for their helpful comments and suggestions. Our research is partially supported by National Natural Science Foundation of China (No. 72122018, No. U1811461, No. 12101389, No. 12326615 and No. 12301384), Shaanxi Province Natural Science Foundation (No. 2021JQ-429 and No. 2021JC-01), Shanghai Pujiang Program (21PJC056), the MOE Project of Key Research Institute of Humanities and Social Sciences (22JJD110001) and Shanghai Research Center for Data Science and Decision Technology.

Appendix A. Proofs for ScBMs

This sections includes the proofs with respect to ScBMs.

A.1 Proof of Lemma 2

Recall $\Delta_y = \mathrm{diag}(\sqrt{n_1^y}, ..., \sqrt{n_{K^y}^y})$ and $\Delta_z = \mathrm{diag}(\sqrt{n_1^z}, ..., \sqrt{n_{K^z}^z})$. Then we can write P as

$$P = YBZ^\intercal = Y\Delta_y^{-1}\Delta_y B\Delta_z\Delta_z^{-1}Z^\intercal, \tag{A.1}$$

where $Y\Delta_y^{-1}$ and $Z\Delta_z^{-1}$ are both column orthogonal matrices. Recall that the SVD of $\Delta_y B\Delta_z$ is denoted by $L_{K^y\times K^y}D_{K^y\times K^y}R^\intercal_{K^y\times K^z}$, then (A.1) implies

$$P = YBZ^\intercal = Y\Delta_y^{-1}LDR^\intercal\Delta_z^{-1}Z^\intercal. \tag{A.2}$$

Note that L, R, $Y\Delta_y^{-1}$ and $Z\Delta_z^{-1}$ are all orthonormal matrices and recall that the SVD of P is $\bar{U}\bar{\Sigma}\bar{V}^\intercal$, and then we have $\bar{\Sigma} = D$,

$$\bar{U} = Y\Delta_y^{-1}L, \tag{A.3}$$

and

$$\bar{V} = Z\Delta_z^{-1}R. \tag{A.4}$$

For $\bar{U}$, since $\Delta_y^{-1}L$ is invertible, $Y_{i*} = Y_{j*}$ if and only if $\bar{U}_{i*} = \bar{U}_{j*}$. In addition, we can easily verify that the rows of $\Delta_y^{-1}L$ are perpendicular to each other and the kth row has length $\sqrt{1/n_k^y}$, therefore we have

$$\|\bar{U}_{i*} - \bar{U}_{j*}\|_2 = \sqrt{(n^y_{g_i^y})^{-1} + (n^y_{g_j^y})^{-1}},$$

if $g_i^y \neq g_j^y$. The argument (1) follows.

For $\bar{V}$, it is obvious that $Z_{i*} = Z_{j*}$ can imply $\bar{V}_{i*} = \bar{V}_{j*}$. While if $Z_{i*} \neq Z_{j*}$, then by (A.4), we have

$$\|\bar{V}_{i*} - \bar{V}_{j*}\|_2 = \|\frac{R_{g_i^z*}}{\sqrt{n^z_{g_i^z}}} - \frac{R_{g_j^z*}}{\sqrt{n^z_{g_j^z}}}\|_2 \geq \xi_n > 0,$$

where the last inequality follows from assumption (C1). The argument (2) follows. ■

A.2 Proof of Lemma 3

In (A.4), we already observe $\bar{V} = Z\Delta_z^{-1}R$. Note that R is not invertible, to facilitate further analysis, we reformulate $\bar{V}$ as

$$\bar{V} = Z\Delta_z^{-1}R = ZB^\intercal \Delta_y (L^{-1})^\intercal D^{-1},$$

using $\Delta_y B \Delta_z = L_{K^y \times K^y} D_{K^y \times K^y} R^\intercal_{K^y \times K^z}$. Without loss of generality, we assume $g_i^z = k, g_j^z = l (l \neq k)$. Then, we have

$$\begin{aligned}\|\bar{V}_{i*} - \bar{V}_{j*}\|_2 &= \|(Z_{i*} - Z_{j*})B^\intercal \Delta_y (L^{-1})^\intercal D^{-1}\|_2 \\ &= \|(B_{*k} - B_{*l})^\intercal \Delta_y (L^{-1})^\intercal D^{-1}\|_2, \\ &\geq \|B_{*k} - B_{*l}\|_2 \|\Delta_y (L^{-1})^\intercal\|_m \|D^{-1}\|_m, \\ &\geq \mu_n \cdot \min_{k=1,\dots,K^y} (n_k^y)^{1/2} / \sigma_n, \qquad (A.5)\end{aligned}$$

where $\|M\|_m := \min_{x:\|x\|_2=1} \|Mx\|_2$, and the last inequality is implied by our condition and the following facts, $\|D^{-1}\|_m = 1/\sigma_n$ and $\|\Delta_y (L^{-1})^\intercal\|_m \geq \|\Delta_y\|_m \|L^{-1}\|_m = \min_{k=1,\dots,K^y} (n_k^y)^{1/2}$ by the definition of Δ_y and the orthogonality of L. The proof is completed. ■

A.3 Proof of Theorem 4

To begin with, we notice that

$$\begin{aligned}\|A^{\rm rp} - P\|_2 &= \|QQ^\intercal ATT^\intercal - P\| \\ &\leq \|A - P\|_2 + \|QQ^\intercal ATT^\intercal - A\|_2 \\ &=: \mathcal{I}_1 + \mathcal{I}_2. \qquad (A.6)\end{aligned}$$

In the sequel, we discuss $\mathcal{I}_1$ and $\mathcal{I}_2$, respectively.

To bound $\mathcal{I}_1$, namely, the deviation of adjacency matrix from its population, we use the results in Lei and Rinaldo (2015). Specifically, under condition (C2), there exists a constant $c = c(\epsilon, c_0)$ such that

$$\|A - P\|_2 \leq c\sqrt{n\alpha_n}. \qquad (A.7)$$

with probability at least $1 - n^{-\epsilon}$ for any $\epsilon > 0$.

To bound $\mathcal{I}_2$, we first notice that

$$\begin{aligned}\mathcal{I}_2 &= \|A - QQ^\intercal A + QQ^\intercal A - QQ^\intercal ATT^\intercal\|_2 \\ &\leq \|A - QQ^\intercal A\|_2 + \|A - ATT^\intercal\|_2, \qquad (A.8)\end{aligned}$$

where in the last inequality we used the facts that $\|AB\|_2 \leq \|A\|_2\|B\|_2$ for any matrices A and B, and $\|QQ^\intercal\|_2 \leq 1$. When $r \geq 4$, $r\text{log}r \leq n$ and $q \geq 1$, the Corollary 10.9 and Theorem 9.2 of Halko et al. (2011) indicate that the following inequality holds with probability at least $1 - 6r^{-r}$,

$$\|A - QQ^\intercal A\|_2 \leq \sigma_{K^y+1}(A)(1 + 11\sqrt{K^y + r} \cdot \sqrt{n})^{\frac{1}{2q+1}}, \qquad (A.9)$$

where $\sigma_{K^y+1}(\cdot)$ denotes the $K+1$th largest singular value of a matrix. In particular, by Weyl's inequality,

$$\sigma_{K^y+1}(A) = \sigma_{K^y+1}(A) - \sigma_{K^y+1}(P) \le \|A-P\|_2. \tag{A.10}$$

Hence, with probability at least $1-6r^{-r}-n^{-\epsilon}$,

$$\|A-QQ^\intercal A\|_2 \le c'\sqrt{n\alpha_n}(1+11\sqrt{K^y+r}\cdot\sqrt{n})^{\frac{1}{2q+1}} \le c''\sqrt{n\alpha_n}(\sqrt{K^y+r}\cdot\sqrt{n})^{\frac{1}{2q+1}} \le c\sqrt{n\alpha_n}, \tag{A.11}$$

where the last inequality follows from the fact that $(\sqrt{K^y+r}\cdot\sqrt{n})^{\frac{1}{2q+1}} = O(1)$ provided that $q = cn^{1/\tau}$ for any $\tau > 0$ and n goes to infinity, and note that we usually use c, c', c'' to denote constants and they may be different from place to place. Similarly, under condition (C3), we have with probability at least $1-6s^{-s}-n^{-\epsilon}$ that

$$\|A-ATT^\intercal\|_2 \le c'\sqrt{n\alpha_n}. \tag{A.12}$$

As a result, with probability larger than $1-6r^{-r}-6s^{-s}-2n^{-\epsilon}$,

$$\mathcal{I}_2 \le c\sqrt{n\alpha_n}. \tag{A.13}$$

Finally, combining the bounds for $\mathcal{I}_1$ and $\mathcal{I}_2$, we have with probability larger than $1-6r^{-r}-6s^{-s}-3n^{-\epsilon}$ that

$$\|A^{\mathrm{rp}}-P\|_2 \le c\sqrt{n\alpha_n}. \tag{A.14}$$

The proof is completed. ■

A.4 Proof of Theorem 5

Generally, we will first bound the perturbation of estimated eigenvectors, and then bound the size for nodes corresponding to a large eigenvector perturbation. At last, we use Lemma 2 to show the remaining nodes are clustered properly. To fix ideas, we now recall and introduce some notation. $\bar{U}$ and $\bar{V}$ denote the left and right K^y leading eigenvectors of P, respectively. Accordingly, U^{rp} and V^{rp} denote the left and right K^y leading eigenvectors of A^{rp}. Likewise, $\tilde{U}^{\mathrm{rp}} := Y^{\mathrm{rp}}X_y^{\mathrm{rp}}$ and $\tilde{V}^{\mathrm{rp}} := Z^{\mathrm{rp}}X_z^{\mathrm{rp}}$ denote the output of RP-SCC, where $Y^{\mathrm{rp}} \in \mathbb{M}_{n,K^y}$ and $Z^{\mathrm{rp}} \in \mathbb{M}_{n,K^z}$ denote the estimated membership matrices, and X_y^{rp} and X_z^{rp} denote the centriods. Next, we discuss the performance of two types of clusters, respectively.

(1) The left side. First, by the modified Davis-Kahan-Wedin sine theorem (O'Rourke et al., 2018) (See Lemma 17), there exists a $K^y \times K^y$ orthogonal matrix O such that,

$$\|U^{\mathrm{rp}}-\bar{U}O\|_{\mathrm{F}} \le \frac{2\sqrt{2K^y}}{\gamma_n}\|A^{\mathrm{rp}}-P\|_2. \tag{A.15}$$

And note that

$$\begin{aligned}\|\tilde{U}^{\mathrm{rp}}-\bar{U}O\|_{\mathrm{F}}^2 &= \|\tilde{U}^{\mathrm{rp}}-U^{\mathrm{rp}}+U^{\mathrm{rp}}-\bar{U}O\|_{\mathrm{F}}^2\\ &\le 2\|\bar{U}O-U^{\mathrm{rp}}\|_{\mathrm{F}}^2+2\|U^{\mathrm{rp}}-\bar{U}O\|_{\mathrm{F}}^2\\ &= 4\|U^{\mathrm{rp}}-\bar{U}O\|_{\mathrm{F}}^2,\end{aligned} \tag{A.16}$$

where the first inequality follows because we assume that $\tilde{U}^{\mathrm{rp}}$ is the global solution minimum of the following k-means objective and $\bar{U}O$ is a feasible solution,

$$(Y^{\mathrm{rp}}, X^{\mathrm{rp}}) = \underset{Y\in\mathbb{M}_{n,K^y}, X\in\mathbb{R}^{K^y\times K^y}}{\arg\min} \|YX - U^{\mathrm{rp}}\|_{\mathrm{F}}^2.$$

So combining (A.16) with (A.15) and the bound of $\|A^{\mathrm{rp}} - P\|_2$ in Theorem 4, we have with probability larger than $1 - 6r^{-r} - 6s^{-s} - 3n^{-\epsilon}$ that

$$\|\tilde{U}^{\mathrm{rp}} - \bar{U}O\|_{\mathrm{F}} \le \frac{c_2\sqrt{K^y n\alpha_n}}{\gamma_n}. \tag{A.17}$$

For notational convenience, we denote the RHS of (A.17) as $\mathrm{err}(K^y, n, c_2, \alpha_n, \gamma_n)$ in what follows.

Now, we begin to bound the fraction of misclustered nodes. Recall

$$\tau = \min\nolimits_{l\neq k} \sqrt{(n_k^y)^{-1} + (n_l^y)^{-1}}, \tag{A.18}$$

and define

$$M^y = \{i \in \{1, ..., n\} : \|\tilde{U}_{i*}^{\mathrm{rp}} - (\bar{U}O)_{i*}\|_{\mathrm{F}} \ge \frac{\tau}{2}\}, \tag{A.19}$$

where M^y is actually the number of misclustered nodes up to permutations as we will see soon. By the definition of M^y, we can see obviously that

$$|M^y| \le \frac{4\|\tilde{U}^{\mathrm{rp}} - \bar{U}O\|_{\mathrm{F}}^2}{\tau^2} \le \frac{4\cdot \mathrm{err}^2(K^y, n, c_1, \alpha_n, \gamma_n)}{\tau^2}. \tag{A.20}$$

Further,

$$\frac{|M^y|}{n} \le \frac{4\|\tilde{U}^{\mathrm{rp}} - \bar{U}O\|_{\mathrm{F}}^2}{\tau^2 n} \le \frac{4\cdot \mathrm{err}^2(K^y, n, c_1, \alpha_n, \gamma_n)}{\tau^2 n}. \tag{A.21}$$

At last, we show that the nodes outside M^y are correctly clustered. First, we have $|M^y| < n_k$ for any k by condition. Define $T_k^y \equiv G_k^y \backslash M^y$, where G_k^y denotes the set of nodes within the true cluster k. Then T_k^y is not an empty set. Let $T^y = \cup_{k=1}^{K^y} T_k^y$. Essentially, the rows in $(\bar{U}O)_{T^y*}$ has a one to one correspondence with those in $\tilde{U}_{T^y*}^{\mathrm{rp}}$. On the one hand, for $i \in T_k^y$ and $j \in T_l^y$ with $l \neq k$, $\tilde{U}_{i*}^{\mathrm{rp}} \neq \tilde{U}_{j*}^{\mathrm{rp}}$, otherwise the following contradiction follows

$$\begin{aligned}\tau &\le \|(\bar{U}O)_{i*} - (\bar{U}O)_{j*}\|_2 \\ &\le \|(\bar{U}O)_{i*} - \tilde{U}_{i*}^{\mathrm{rp}}\|_2 + \|(\bar{U}O)_{j*} - \tilde{U}_{j*}^{\mathrm{rp}}\|_2 \\ &< \frac{\tau}{2} + \frac{\tau}{2},\end{aligned} \tag{A.22}$$

where the first and last inequality follows from the Lemma 2(1) and the definition of M_k^y in (A.19), respectively. On the other hand, for $i, j \in T_k^y$, $\tilde{U}_{i*}^{\mathrm{rp}} = \tilde{U}_{j*}^{\mathrm{rp}}$, because otherwise $\tilde{U}_{T*}$ has more than K^y distinct rows which is contradict with the fact that the output size for the left side cluster is K^y.

As a result, we have arrived at the conclusion (1) of Theorem 5.

(2) The right side. First, follow the same lines as in (1), we have with probability larger than $1-6r^{-r}-6s^{-s}-3n^{-\epsilon}$ that

$$\|\tilde{V}^{\rm rp}-\bar{V}O'\|_{\rm F}\leq \frac{c_2\sqrt{K^y n\alpha_n}}{\gamma_n}={\rm err}(K^y,n,c_2,\alpha_n,\gamma_n), \tag{A.23}$$

where O' is an orthogonal matrix. Here we want to emphasize that $\bar{V}$, $V^{\rm rp}$, and $\tilde{V}^{\rm rp}$ are all $n\times K^y$, but the population cluster size and target cluster size are both K^z. This brings different performance of the right side clusters compared to that of the left side counterpart.

Now we begin to see how the fraction of misclustered nodes corresponding to the right side differs from that corresponding to left side. Denote

$$\delta=\min_{1\leq k\neq l\leq K^z}\|\frac{R_{k*}}{\sqrt{n_k^z}}-\frac{R_{l*}}{\sqrt{n_l^z}}\|_2, \tag{A.24}$$

and define

$$M^z=\{i\in\{1,...,n\}:\ \|(\tilde{V})^{\rm rp}_{i*}-(\bar{V}O')_{i*}\|_{\rm F}\geq\frac{\delta}{2}\}, \tag{A.25}$$

where M^z is actually the number of misclustered nodes up to permutations as we will see soon. By the definition of M^z, it is easy to see that

$$|M^z|\leq 4\frac{\|\tilde{V}^{\rm rp}_{i*}-(\bar{V}O')_{i*}\|_{\rm F}^2}{\delta^2}. \tag{A.26}$$

Moreover, we have

$$\frac{|M^z|}{n}\leq 4\frac{\|\tilde{V}^{\rm rp}_{i*}-(\bar{V}O')_{i*}\|_{\rm F}^2}{\delta^2 n}. \tag{A.27}$$

Finally, we show that the nodes outside M_k^z are correctly clustered up to some permutations. As the left side case, we have $|M_k^z|<n_k^z$ by condition. Define $T_k^z\equiv G_k^z\backslash M_k^z$. Then T_k^z is not an empty set. Let $T^z=\cup_{k=1}^{K^z}T_k^z$. Then follow the same lines as those in (1) and note the results in Lemma 2(2), we can easily show the rows in $(\bar{V}O')_{T^z*}$ has a one to one correspondence with those in $\tilde{V}^{\rm rp}_{T^z*}$. Hence the corresponding nodes are correctly clustered.

Till now, we have proved the results in Theorem 5. ■

A.5 Proof of Theorem 6

Let G be the adjacency matrix of an Erodös-Renyi graph with each edge probability being $0<p<1$, then it is easy to see that $A^{\rm rs}=\frac{1}{p}G\circ A$, where $\circ$ denotes the entry-wise multiplication. Note that

$$\begin{aligned}\|A^{\rm rs}-P\|_2&=\|\frac{1}{p}G\circ A-P\|_2\\&=\|\frac{1}{p}G\circ(A-P)+\frac{1}{p}G\circ P-P\|_2\\&\leq\|\frac{1}{p}G\circ(A-P)\|_2+\|\frac{1}{p}G\circ P-P\|_2,\\&=\mathcal{I}_1+\mathcal{I}_2.\end{aligned} \tag{A.28}$$

In the sequel, we discuss $\mathcal{I}_1$ and $\mathcal{I}_2$, respectively.

First, We bound $\mathcal{I}_1$ using Lemma 18, which provides the a spectral-norm bound of a random matrix with independent and bounded entries. In particular, we proceed by conditioning on $A - P \equiv W$. Write $(G \circ W)_{ij} = g_{ij} W_{ij}$, where $g_{ij} \sim$ Bernoulli(p). By simple calculations, we have,

$$\begin{aligned}\sigma_1 := \max_i \sqrt{\mathbb{E}(\sum_j g_{ij}^2 W_{ij}^2 | W)} &= \max_i \sqrt{\sum_j W_{ij}^2 \mathbb{E}(g_{ij}^2 | W)} \\ \leq \max_i \sqrt{p} \sqrt{\|W_{i*}\|_2^2} &\leq \sqrt{p} \|W\|_2. \end{aligned} \tag{A.29}$$

Analogously, (A.29) also holds for

$$\sigma_2 := \max_j \sqrt{\mathbb{E}(\sum_i g_{ij}^2 W_{ij}^2 | W)}.$$

With these bounds, we have by Lemma 18 that with probability $1 - n^{\nu}$, there exists constant $c(\nu)$ such that,

$$\mathcal{I}_1 \leq \frac{1}{p} c \max(\sqrt{p}\|W\|_2, \sqrt{\log n}). \tag{A.30}$$

Further, by the concentration bound $\|A - P\|_2$ in Lei and Rinaldo (2015), we have by condition (C2) that,

$$\|W\|_2 = \|A - P\|_2 \leq c' \sqrt{n\alpha_n}, \tag{A.31}$$

with probability larger than $1 - n^{-\nu}$. Note that we use c, c', c'' to represent the generic constants and they may be different from line to line. Combining (A.31) with (A.30), we have with probability larger than $1 - 2n^{-\nu}$ that,

$$\mathcal{I}_1 \leq c'' \max(\sqrt{\frac{n\alpha_n}{p}}, \frac{\sqrt{\log n}}{p}). \tag{A.32}$$

Second, we bound $\mathcal{I}_2$. We will use Lemma 19 which provides bounds on the spectral deviation of a random matrix from its expectation. Specifically, B and X in Lemma 19 correspond to P and $\frac{1}{p} G \circ P$ in our case. It is easy to see that $\mathbb{E}(X) = B$ and $\max_{jk} |X_{jk}| \leq \alpha_n / p$. Moreover, we have

$$\mathrm{Var} X_{jk} \leq P_{jk}^2 / p,$$

and

$$\begin{aligned}\mathbb{E}(X_{jk} - P_{jk})^4 &\leq \mathrm{Var} X_{jk} \cdot \|X_{jk} - P_{jk}\|_\infty^2 \\ &\leq \frac{P_{jk}^2}{p} \cdot \max\big(P_{jk}, \frac{P_{jk}}{p} - P_{jk}\big)^2 \\ &= \frac{P_{jk}^4}{p} \cdot \max\big(1, (\frac{1}{p} - 1)\big)^2. \end{aligned} \tag{A.33}$$

Therefore, by Lemma 19 and the fact that $P_{ij} \le \alpha_n$, we have

$$\mathcal{I}_2 \le c\Big(2\alpha_n\sqrt{\frac{n}{p}} + \alpha_n\frac{\sqrt{n}}{p^{1/4}}\max\big(1,\ \sqrt{\frac{1}{p}-1}\big)\Big)$$

$$\le c'\sqrt{\frac{n\alpha_n^2}{p}}\Big(1 + p^{1/4}\cdot\max\big(1,\ \sqrt{\frac{1}{p}-1}\big)\Big), \tag{A.34}$$

with probability larger than $1-\exp\Big(-c''np(1+p^{1/4}\cdot\max\big(1,\ \sqrt{\frac{1}{p}-1}\big)^2\Big)$.

Finally, combining (A.34) with (A.32), we will obtain the conclusion in Theorem 6. ■

Appendix B. Proofs for DC-ScBMs

This section includes the proofs with respect to DC-ScBMs.

B.1 Proof of Lemma 9

Define $\tilde{Y}$ and $\tilde{Z}$ be normalized membership matrices such that $\tilde{Y}(i,k) = \tilde{\theta}_i^y$ if $i \in G_k^y$ and $\tilde{Y}(i,k) = 0$ otherwise, and accordingly $\tilde{Z}(i,k) = \tilde{\theta}_i^z$ if $i \in G_k^z$ and $\tilde{Z}(i,k) = 0$ otherwise. Then it is easy to see $\tilde{Y}^\intercal\tilde{Y} = I$ and $\tilde{Z}^\intercal\tilde{Z} = I$. Let $\Psi^y = \mathrm{diag}(\|\phi_1^y\|_2, ..., \|\phi_{K^y}^y\|_2)$ and $\Psi^z = \mathrm{diag}(\|\phi_1^z\|_2, ..., \|\phi_{K^z}^z\|_2)$. Then after some rearrangements, we can see that

$$\mathrm{diag}(\theta^y)Y = \tilde{Y}\Psi^y, \qquad \mathrm{diag}(\theta^z)Z = \tilde{Z}\Psi^z. \tag{B.1}$$

Thus,

$$P = \mathrm{diag}(\theta^y)YBZ^\intercal\mathrm{diag}(\theta^z) = \tilde{Y}\Psi^yB\Psi^z\tilde{Z}^\intercal. \tag{B.2}$$

Denote the SVD of $\Psi^yB\Psi^z$ as

$$\Psi^yB\Psi^z = H_{K^y\times K^y}D_{K^y\times K^y}J^\intercal_{K^y\times K^z}, \tag{B.3}$$

where H and J have orthonormal columns. Then, (B.2) implies

$$P = \tilde{Y}HDJ^\intercal\tilde{Z}^\intercal. \tag{B.4}$$

By the orthonormality of $\tilde{Y}, \tilde{Z}, H$ and J, we have

$$\bar{U} = \tilde{Y}H, \quad \bar{V} = \tilde{Z}J, \quad \bar{\Sigma} = D.$$

Specifically, $\bar{U}_{i*} = \tilde{\theta}_i^yH_{k*}$ for $i \in G_k^y$, and $\bar{V}_{i*} = \tilde{\theta}_i^zJ_{k*}$ for $i \in G_k^z$. Since H is square matrix with orthonormal columns, $\cos(\bar{U}_{i*}, \bar{U}_{j*}) = 0$ if $g_i^y \ne g_j^y$. Thus the argument (1) follows.

Now we proceed to calculate $\cos(\bar{V}_{i*}, \bar{V}_{j*})$ for $g_i^z \ne g_j^z$. Without loss of generality, we assume $g_i^z = k, g_j^z = l$. First notice that,

$$\cos(\bar{V}_{i*}, \bar{V}_{j*}) = \frac{\bar{V}_{i*}\bar{V}_{j*}^\intercal}{\|\bar{V}_{i*}\|_2\|\bar{V}_{j*}\|_2} = \frac{\tilde{\theta}_i^z\tilde{\theta}_j^zJ_{k*}J_{l*}^\intercal}{\tilde{\theta}_i^z\tilde{\theta}_j^z\|J_{k*}\|_2\|J_{l*}\|_2}, \tag{B.5}$$

where we note that $\bar{V}_{i*}$ and $\bar{V}_{j*}$ are row vectors. We will discuss the numerator and denominator of (B.5), respectively. By (B.3), we have

$$J = \Psi^zB^\intercal\Psi^yHD^{-1} := \tilde{B}^\intercal HD^{-1}, \tag{B.6}$$

where we define $\Psi^y B \Psi^z := \tilde{B}$. Therefore, we obtain

$$\cos(\bar{V}_{i*}, \bar{V}_{j*}) = \cos((\tilde{B}_{*k})^\intercal H \bar{\Sigma}^{-1}, (\tilde{B}_{*l})^\intercal H \bar{\Sigma}^{-1}),$$

where we used the fact that H is orthogonal and hence $H^\intercal = H^{-1}$. The argument (2) holds immediately. ■

B.2 Proof of Lemma 10

First, it is worth noting that for any $1 \le k \le K^z$, $\|(\tilde{B}_{*k})^\intercal (H^{-1})^\intercal \bar{\Sigma}^{-1}\|_2 > 0$, which excludes the trivial case that $\cos((\tilde{B}_{*k})^\intercal (H^{-1})^\intercal \bar{\Sigma}^{-1}, (\tilde{B}_{*l})^\intercal (H^{-1})^\intercal \bar{\Sigma}^{-1}) = 1$ for $k \neq l$, where we have denoted $g_i^z = k, g_j^z = l (l \neq k)$. In fact, by the orthogonality of H, the invertibility of $\bar{\Sigma}$, and our condition that $\min_k \|\tilde{B}_{*k}\|_2 > 0$, we have

$$\|(\tilde{B}_{*k})^\intercal H \bar{\Sigma}^{-1}\|_2 \ge \|\tilde{B}_{*k}\|_2 \sigma_{\min}(H) \sigma_{\min}(\bar{\Sigma}^{-1}) > 0.$$

Second, we proceed to show that under provided conditions, for any λ and any $1 \le k < l \le K^z$,

$$\|(\tilde{B}_{*k})^\intercal H \bar{\Sigma}^{-1} - \lambda (\tilde{B}_{*l})^\intercal H \bar{\Sigma}^{-1}\|_2^2 > 0,$$

which implies that

$$\cos((\tilde{B}_{*k})^\intercal H \bar{\Sigma}^{-1}, (\tilde{B}_{*l})^\intercal H \bar{\Sigma}^{-1}) < 1.$$

In particular, we have

$$\begin{aligned}
\|(\tilde{B}_{*k})^\intercal H \bar{\Sigma}^{-1} - \lambda (\tilde{B}_{*l})^\intercal H \bar{\Sigma}^{-1}\|_2^2 &= \|(\tilde{B}_{*k}^\intercal - \lambda \tilde{B}_{*l}^\intercal) \cdot H \cdot \bar{\Sigma}^{-1}\|_2^2 \\
&\ge \sigma_{\min}^2(H) \sigma_{\min}^2(\bar{\Sigma}^{-1}) \|\tilde{B}_{*k}^\intercal - \lambda \tilde{B}_{*l}^\intercal\|_2^2 \\
&= \sigma_{\min}^2(H) \sigma_{\min}^2(\bar{\Sigma}^{-1}) \cdot (\lambda^2 \|\tilde{B}_{*l}\|_2^2 - 2\lambda \tilde{B}_{*k}^\intercal \tilde{B}_{*l} + \|\tilde{B}_{*k}\|_2^2) \\
&:= \sigma_{\min}^2(H) \sigma_{\min}^2(\bar{\Sigma}^{-1}) \cdot (a\lambda^2 + b\lambda + c), \qquad \text{(B.7)}
\end{aligned}$$

where $a := \|\tilde{B}_{*l}\|_2^2$, $b := -2\tilde{B}_{*k}^\intercal \tilde{B}_{*l}$, and $c := \|\tilde{B}_{*k}\|_2^2$. Note that the parabola of the form $a\lambda^2 + b\lambda + c$ is always larger than 0 if the discriminant $b^2 - 4ac := 4(\tilde{B}_{*k}^\intercal \tilde{B}_{*l})^2 - 4\|\tilde{B}_{*l}\|_2^2 \|\tilde{B}_{*k}\|_2^2 < 0$, which is equivalent to our condition that $\cos(\tilde{B}_{*k}, \tilde{B}_{*l}) < 1$.

Finally, we provide an explicit upper bound for $\cos((\tilde{B}_{*k})^\intercal H \bar{\Sigma}^{-1}, (\tilde{B}_{*l})^\intercal H \bar{\Sigma}^{-1})$. Selecting $\lambda = -\frac{2a}{b}$, we observe that

$$\begin{aligned}
\lambda^2 \|\tilde{B}_{*l}\|_2^2 - 2\lambda \tilde{B}_{*k}^\intercal \tilde{B}_{*l} + \|\tilde{B}_{*k}\|_2^2 \ge \frac{-b^2 + 4ac}{4a} &:= \frac{4(-\tilde{B}_{*k}^\intercal \tilde{B}_{*l})^2 + 4\|\tilde{B}_{*l}\|_2^2 \|\tilde{B}_{*k}\|_2^2}{4\|\tilde{B}_{*l}\|_2^2} \\
&= \|\tilde{B}_{*k}\|_2^2 \left(-\frac{(\tilde{B}_{*k}^\intercal \tilde{B}_{*l})^2}{\|\tilde{B}_{*l}\|_2^2 \|\tilde{B}_{*k}\|_2^2} + 1\right) \\
&\ge \underline{\iota}_n^2 (1 - \zeta_n^2),
\end{aligned}$$

where the last inequality is impled by our assumptions. Combining this with (B.7), we have for any λ that,

$$\|(\tilde{B}_{*k})^\intercal H \bar{\Sigma}^{-1} - \lambda (\tilde{B}_{*l})^\intercal H \bar{\Sigma}^{-1}\|_2^2 \ge \sigma_{\min}^2(H) \sigma_{\min}^2(\bar{\Sigma}^{-1}) \underline{\iota}_n^2 (1 - \zeta_n^2). \qquad \text{(B.8)}$$

Denote $\mathcal{B}_k := (\tilde{B}_{*k})^\intercal H\bar{\Sigma}^{-1}$ and $\mathcal{B}_l := (\tilde{B}_{*l})^\intercal H\bar{\Sigma}^{-1}$ and choose $\lambda = \frac{\mathcal{B}_k\mathcal{B}_l^\intercal}{\|\mathcal{B}_l\|_2^2}$ in the LHS of (B.8), we thus have

$$\frac{-(\mathcal{B}_k\mathcal{B}_l^\intercal)^2 + \|\mathcal{B}_l\|_2^2\|\mathcal{B}_k\|_2^2}{\|\mathcal{B}_l\|_2^2} \geq \sigma_{\min}^2(H)\sigma_{\min}^2(\bar{\Sigma}^{-1})\underline{\iota}_n^2(1-\zeta_n^2),$$

which indicates that

$$\cos((\tilde{B}_{*k})^\intercal H\bar{\Sigma}^{-1}, (\tilde{B}_{*l})^\intercal H\bar{\Sigma}^{-1}) := \cos(\mathcal{B}_k, \mathcal{B}_l) \leq \sqrt{1 - \frac{\sigma_{\min}^2(H)\sigma_{\min}^2(\bar{\Sigma}^{-1})\underline{\iota}_n^2(1-\zeta_n^2)}{\|\mathcal{B}_k\|_2^2}}.$$

At last, by our condition,

$$\|\mathcal{B}_k\|_2^2 \leq \sigma_n^2(H)\sigma_n^2(\bar{\Sigma}^{-1})\bar{\iota}_n^2.$$

Consequently,

$$\cos((\tilde{B}_{*k})^\intercal H\bar{\Sigma}^{-1}, (\tilde{B}_{*l})^\intercal H\bar{\Sigma}^{-1}) \leq \sqrt{1 - \frac{\sigma_{\min}^2(H)\sigma_{\min}^2(\bar{\Sigma})\underline{\iota}_n^2(1-\zeta_n^2)}{\sigma_n^2(H)\sigma_n^2(\bar{\Sigma})\bar{\iota}_n^2}}.$$

The proof is completed. ■

B.3 Proof of Theorem 11

To fix ideas, we now recall and introduce some notation. $\bar{U}$ and $\bar{V}$ denote the left and right K^y leading eigenvectors of P, respectively. Accordingly, $U^{\rm rp}$ and $V^{\rm rp}$ denote the left and right K^y leading eigenvectors of $A^{\rm rp}$. Note that the rows of $\bar{U}$ and $\bar{V}$ are all non-zero, but the rows of $U^{\rm rp}$ and $V^{\rm rp}$ might be zero. Define $\bar{U}'$ and $\bar{V}'$ be the row-normalized version of $\bar{U}$ and $\bar{V}$, respectively. Define $(U^{\rm rp})'$ and $(V^{\rm rp})'$ be the row-normalized version of $U^{\rm rp}$ and $V^{\rm rp}$ with their zero rows remained the same. $\tilde{U}^{\rm rp}$ and $\tilde{V}^{\rm rp}$ denote the output of the randomized spherical spectral clustering, namely, the k-means solution of $(U^{\rm rp})'$ and $(V^{\rm rp})'$, respectively. In the sequel, we discuss the performance of two types of clusters, respectively.

(1) The left side. First, by the modified Davis-Kahan-Wedin sine theorem (Theorem 19 in O'Rourke et al. (2018)), there exists a $K^y \times K^y$ orthogonal matrix O such that,

$$\|U^{\rm rp} - \bar{U}O\|_{\rm F} \leq \frac{2\sqrt{2K^y}}{\gamma_n}\|A^{\rm rp} - P\|_2. \tag{B.9}$$

Combining (B.9) with the results in Theorem 4, we have

$$\|U^{\rm rp} - \bar{U}O\|_{\rm F} \leq c\frac{2\sqrt{2K^y}}{\gamma_n}\sqrt{n\alpha_n}, \tag{B.10}$$

with probability $1 - 6r^{-r} - 6s^{-s} - 3n^{-\epsilon}$ for any $\epsilon > 0$ and some constant $c > 0$. Note that the constant c may be different from line to line in this proof. And without loss of generality, we will assume the orthogonal matrix O is the identity matrix I in the following proof.

Then, we bound $\|\tilde{U}^{\rm rp} - \bar{U}'\|_{\rm F}$. We first notice that for any vectors a and b,

$$\|\frac{a}{\|a\|_2} - \frac{b}{\|b\|_2}\|_2 \leq 2\frac{\|a-b\|_2}{\max(\|a\|_2, \|b\|_2)}$$

holds, and for any $a = 0$, $\|0 - \frac{b}{\|b\|_2}\|_2 \leq 2\frac{\|0-b\|_2}{\|b\|_2}$ holds trivially. Thus we have

$$\|(U^{\mathrm{rp}})' - \bar{U}'\|_{\mathrm{F}}^2 \leq c\sum_{i=1}^{n} \frac{\|(\bar{U}^{\mathrm{rp}})_{i*} - \bar{U}_{i*}\|_2^2}{\|\bar{U}_{i*}\|_2^2} \leq c\frac{\|(U^{\mathrm{rp}}) - \bar{U}\|_{\mathrm{F}}^2}{\min_i \|U_{i*}\|_2^2} \leq c\frac{K^y n\alpha_n \kappa^y}{\gamma_n^2}, \tag{B.11}$$

where the last inequality follows from (B.10) and the definition of κ^y coupled with the fact that $\|\bar{U}_{i*}\|_2^2 = |\tilde{\theta}_i^y|^2$ (see the proof of Lemma 9 for details). Further, since $\tilde{U}^{\mathrm{rp}}$ is the k-means solution of $(U^{\mathrm{rp}})'$, we can obtain

$$\|\tilde{U}^{\mathrm{rp}} - \bar{U}'\|_{\mathrm{F}}^2 \leq \|\tilde{U}^{\mathrm{rp}} - (U^{\mathrm{rp}})'\|_{\mathrm{F}}^2 + \|(U^{\mathrm{rp}})' - \bar{U}'\|_{\mathrm{F}}^2 \leq 2\|(U^{\mathrm{rp}})' - \bar{U}'\|_{\mathrm{F}}^2 \leq c\frac{K^y n\alpha_n \kappa^y}{\gamma_n^2}. \tag{B.12}$$

Next, we bound the number of misclustered nodes. Define

$$M^y = \{i : \|\tilde{U}_{i*}^{\mathrm{rp}} - \bar{U}'_{i*}\|_2 \geq \frac{1}{\sqrt{2}}\}. \tag{B.13}$$

By the definition of M^y and (B.12), we have

$$|M^y| \leq 2\|\tilde{U}^{\mathrm{rp}} - \bar{U}'\|_{\mathrm{F}}^2. \tag{B.14}$$

Combining this with (B.12), we have

$$\frac{|M^y|}{n} \leq c\frac{K^y \alpha_n \kappa^y}{\gamma_n^2}. \tag{B.15}$$

By the condition, we know that $|M^y| < n_k^y$ for all k. Hence the nodes outside those indexed by M^y but within each true cluster are not empty. That is, $G_k^y \cap (\{1, ..., n\} \backslash M^y) \neq \emptyset$, where G_k^y denotes the set of nodes within the true cluster k. Now we show that these nodes are clustered correctly. On the one hand, suppose $i, j \in \{1, ..., n\} \backslash M^y$ are in different clusters, then their estimated clusters are also different. Otherwise we have

$$\begin{aligned} \|\bar{U}'_{i*} - \bar{U}'_{j*}\|_2 &\leq \|\bar{U}'_{i*} - \tilde{U}_{i*}^{\mathrm{rp}}\|_2 + \|\tilde{U}_{j*}^{\mathrm{rp}} - \bar{U}'_{j*}\|_2 \\ &< \sqrt{2}, \end{aligned} \tag{B.16}$$

where the last inequality follows from the definition of M^y. Since $\bar{U}'_{i*}$ and $\bar{U}'_{j*}$ are normalized vectors and by Lemma 9 we know that they are orthogonal with each other, the LHS of (B.16) is $\sqrt{2}$, which contradicts with the RHS of (B.16). On the other hand, if $i, j \in \{1, ..., n\} \backslash M^y$ are in the same cluster, then their estimated clusters are also identical. Otherwise $\tilde{U}^{\mathrm{rp}}$ has more than K^y distinct rows, which violates the fact that the output cluster size is K^y.

As a result, we have arrived the conclusion in (1).

(2) The right side. Following the same proof strategy with that in (1), we can show that

$$\|V^{\mathrm{rp}} - \bar{V}\|_{\mathrm{F}} \leq c\frac{2\sqrt{2K^y}}{\gamma_n}\sqrt{n\alpha_n}, \tag{B.17}$$

and

$$\|\tilde{V}^{\rm rp}-\bar{V}'\|_{\rm F}^2 \le \|\tilde{V}^{\rm rp}-(V^{\rm rp})'\|_{\rm F}^2+\|(V^{\rm rp})'-\bar{V}'\|_{\rm F}^2 \le 2\|(V^{\rm rp})'-\bar{V}'\|_{\rm F}^2 \le c\frac{K^y n\alpha_n\kappa^z}{\gamma_n^2}, \tag{B.18}$$

where the last inequality follows from the fact that $\bar{V}_{i*}=\tilde{\theta}_i^z(\tilde{B}_{*g_i})^\intercal HD^{-1}$ and the definition of κ^z.

Next, we bound the number of misclustered nodes. Define

$$M^z=\{i:\|\tilde{V}_{i*}^{\rm rp}-\bar{V}_{i*}'\|_2 \ge \frac{\sqrt{1-\eta(P)}}{\sqrt{2}}\}. \tag{B.19}$$

By the definition of M^z,

$$\frac{|M^z|}{n} \le c\frac{K^y\alpha_n\kappa^z}{(1-\eta(P))\gamma_n^2}. \tag{B.20}$$

By the condition, we know that $|M^z|<n_k^z$ for all k. Hence the nodes outside those indexed by M^z but within each true cluster are not empty. That is, $G_k^z\cap(\{1,...,n\}\backslash M^z)\neq\emptyset$, where G_k^z denotes the set of nodes within the true cluster k. Now we show that these nodes are clustered correctly. On the one hand, if $i,j\in\{1,...,n\}\backslash M^z$ are in different clusters, then their estimated clusters are also different. Otherwise we have

$$\begin{aligned}\|\bar{V}_{i*}'-\bar{V}_{j*}'\|_2 &\le \|\bar{V}_{i*}'-(\tilde{V}^{\rm rp})_{i*}\|_2+\|(\tilde{V}^{\rm rp})_{j*}-\bar{V}_{j*}'\|_2\\ &<\sqrt{2(1-\eta(P))},\end{aligned} \tag{B.21}$$

where the last inequality follows from the definition of S^z. Since $\bar{V}_{i*}'$ and $\bar{V}_{j*}'$ are normalized vectors and by Lemma 9 and the definition of $\eta(P)$, we have

$$\|\bar{V}_{i*}'-\bar{V}_{j*}'\|_2=\sqrt{1+1-2{\rm cos}(\bar{V}_{i*}',\bar{V}_{j*}')}\ge\sqrt{2(1-\eta(P))}, \tag{B.22}$$

which contradicts with the RHS of (B.21). On the other hand, if $i,j\in\{1,...,n\}\backslash M^z$ are in the same cluster, then their estimated clusters are the also identical. Otherwise $\tilde{V}^{\rm rp}$ has more than K^z distinct rows, which contradicts the fact that the output cluster size is K^z.

As a result, we have arrived the conclusion in (2). ■

Appendix C. Proofs for auxiliary theorems

This section includes proofs for auxiliary theorems.

C.1 Proofs of Theorem 14

Recall that the SVD of $\Delta_y B\Delta_z$ is $L_{K^y\times K'}D_{K'\times K'}R_{K'\times K^z}^\intercal$, and $\bar{B}=B\Delta_z$, we thus have

$$(\Delta_y B\Delta_z)(\Delta_y B\Delta_z)^\intercal=\Delta_y B\Delta_z^2B^\intercal\Delta_y=\Delta_y\bar{B}\bar{B}^\intercal\Delta_y=LD^2L^\intercal.$$

Without loss of generality, suppose $g_i^y = k$ and $g_j^y = l$ $(l \neq k)$, we then have

$$\begin{aligned}
\mu_n^2 \|U_{i*} - U_{j*}\|_2^2 &= \sum_{k_1=1}^{K'} \mu_n^2 (\frac{L_{kk_1}}{\sqrt{n_k^y}} - \frac{L_{lk_1}}{\sqrt{n_l^y}})^2 \\
&\geq \sum_{k_1=1}^{K'} D_{k_1k_1}^2 (\frac{L_{kk_1}}{\sqrt{n_k^y}} - \frac{L_{lk_1}}{\sqrt{n_l^y}})^2 \\
&= \sum_{k_1=1}^{K'} D_{k_1k_1}^2 (\frac{L_{kk_1}}{\sqrt{n_k^y}})^2 + \sum_{k_1=1}^{K'} D_{k_1k_1}^2 (\frac{L_{lk_1}}{\sqrt{n_l^y}})^2 - 2\sum_{k_1=1}^{K'} D_{k_1k_1}^2 \frac{L_{kk_1}L_{lk_1}}{\sqrt{n_k^y n_l^y}} \\
&= (\bar{B}\bar{B}^\intercal)_{kk} + (\bar{B}\bar{B}^\intercal)_{ll} - 2(\bar{B}\bar{B}^\intercal)_{kl} \\
&= \|\bar{B}_{k*}\|_2^2 + \|\bar{B}_{l*}\|_2^2 - 2\bar{B}_{l*}(\bar{B}_{k*})^\intercal \\
&= \|\bar{B}_{k*} - \bar{B}_{l*}\|_2^2 \\
&\geq \nu_n^2,
\end{aligned}$$

where the first and last inequalities follow from our conditions. As a result, for $\Theta_{i*} \neq \Theta_{j*}$, we obtain $\|U_{i*} - U_{j*}\|_2 \geq \frac{\nu_n}{\mu_n}$. ■

C.2 Proofs of Theorem 13

The proof follows that of Theorem 3.2 in Abbe et al. (2020) closely. Before moving on, we here first provide important results which would be used to prove Theorem 13.

Theorem 15 (Simplification of Theorem 2.1 of Abbe et al. (2020)) *Suppose P is the population adjacency matrix of* $\mathrm{SBM}(n, a\frac{\mathrm{log}n}{n}, b\frac{\mathrm{log}n}{n}, J)$. *Consider a symmetric matrix $\tilde{A}$ which satisfies $\mathbb{E}(\tilde{A}) = P$. u_2 and u_2^* are eigenvectors associated with the second largest eigenvalues of $\tilde{A}$ and P, respectively. Note that the two non-zero eigenvalues of P are $\lambda_1^* = (a+b)\mathrm{log}n/2$ and $\lambda_2^* = (a-b)\mathrm{log}n/2$, respectively. Define $\Delta^* := (\lambda_1^* - \lambda_2^*) \wedge \lambda_1^* \wedge \lambda_2^* = (b \wedge \frac{a-b}{2})\mathrm{log}n$ and $\kappa := \lambda_1^*/\Delta^*$. Suppose the following* **A1-A4** *hold,*

A1 *There exists $\gamma > 0$ such that $\|P\|_{2\to\infty} \leq \gamma\Delta^*$.*

A2 *For any $m \in [n]$, the entries in the mth row and column of $\tilde{A}$ are independent with others, i.e., $\{\tilde{A}_{ij}, i = m \text{ or } j = m\}$ are independent of $\{\tilde{A}_{ij} : i \neq m, j \neq m\}$.*

A3 *For some $\delta_0 \in (0,1)$, $\mathbb{P}(\|\tilde{A} - P\|_2 \leq \gamma\Delta^*) \geq 1 - \delta_0$.*

A4 *Suppose $\phi(x)$ is continuous and non-decreasing in $\mathbb{R}_+$ with $\phi(0) = 0$, $\phi(x)/x$ is non-decreasing in $\mathbb{R}_+$ and $\delta_1 \in (0,1)$. For any $m \in [n]$ and $w = (w_i) \in \mathbb{R}^n$,*

$$\mathbb{P}\left(\|\sum_{i=1}^n w_i(\tilde{A} - P)_{i*}\|_2 \leq \Delta^*\|w\|_\infty \phi\left(\frac{\|w\|_2}{\sqrt{n}\|w\|_\infty}\right)\right) \geq 1 - \frac{\delta_1}{n},$$

and $32\kappa\mathrm{max}\{\gamma, \phi(\gamma)\} \leq 1$.

Then with probability at least $1-\delta_0-2\delta_1$,

$$\min_{s\in\{\pm 1\}}\|u_2 - s\tilde{A}u_2^*/\lambda_2^*\|_\infty \le \kappa(\kappa+\phi(1))(\gamma+\phi(\gamma))\|u_2^*\|_\infty. \tag{C.1}$$

■

Theorem 15 is a direct simplification of Theorem 2.1 in Abbe et al. (2020) provided that the population matrix P is the two-block model $\mathrm{SBM}(n, a\frac{\log n}{n}, b\frac{\log n}{n}, J)$, and thus we omit its proof. In the next corollary, we show that **A1-A4** actually hold for $\tilde{A} = A^{\mathrm{rs}}$ under certain parameters and thus the specified bound for (C.1) is obtained.

Corollary 16 *Suppose P is the population adjacency matrix of* $\mathrm{SBM}(n, a\frac{\log n}{n}, b\frac{\log n}{n}, J)$*, A is one realization from* $\mathrm{SBM}(n, a\frac{\log n}{n}, b\frac{\log n}{n}, J)$ *and A^{rs} (see 2.1) is a sparsified adjacency matrix from A with the sampling probability being p. If $p > 1/2$, then with probability larger than* $1-2n^{-\nu}-\exp\Big(-c_6np\big(1+p^{1/4}\cdot\max(1, \sqrt{\frac{1}{p}-1})^2\big)\Big)-4n^{-4p+1}$

$$\min_{s\in\{\pm 1\}}\|u_2 - sA^{\mathrm{rs}}u_2^*/\lambda_2^*\|_\infty \le \frac{C}{\sqrt{n}\log\log n^{p^2}}, \tag{C.2}$$

where u_2 and u_2^* are eigenvectors associated with the second largest eigenvalues of A^{rs} and P, respectively, ν and c_6 are the same with those in Theorem 6, and C is some constant depending on a, b and p.

Proof of Corollary 16 We would use Theorem 13 to prove. Take $\gamma = [(b\wedge\frac{a-b}{2})\sqrt{\log n}]^{-1}c_5/p$, where c_5 is the same constant as that in Theorem 6, and take $\phi(x) = \frac{2a+4}{(b\wedge\frac{a-b}{2})}(1\vee(\log(1/x)))^{-1}$. Note that $\|P\|_{2\to\infty} = \frac{\log n}{\sqrt{n}}\sqrt{\frac{a^2+b^2}{2}}$, then **A1** is satisfied when n is sufficiently large. **A2** is trivially satisfied. By Theorem 6 and the assumption that $p > 1/2$, **A3** is satisfied with $\delta_0 = 2n^{-\nu}-\exp\Big(-c_6np\big(1+p^{1/4}\cdot\max(1, \sqrt{\frac{1}{p}-1})^2\big)\Big)$. Note that pA^{rs} can be regarded as one realization from $\mathrm{SBM}(n, ap\frac{\log n}{n}, bp\frac{\log n}{n}, J)$ with population matrix being pP. Therefore, taking $\bar{p} = pa\log n/n$, $\alpha = 4/a$ and $X_i = pA^{\mathrm{rs}}_{i*}$ in Lemma 20 and after rearranging, we can obtain

$$\mathbb{P}\left(|(A-A^*)_{i*}w| \le \frac{(2a+4)\log n}{1\vee\log(\frac{\sqrt{n}\|w\|_\infty}{\|w\|_2})}\|w\|_\infty\right) \ge 1-2n^{-4p}.$$

As a result, **A4** is satisfied. And the result follows from Theorem 13. ■

Proof of Theorem 13 The proof follows that of Theorem 3.2 in Abbe et al. (2020) closely.

(1) As $\sqrt{a}-\sqrt{b} > \sqrt{2/p}$, we can select $\varepsilon = \varepsilon(a,b,p) > 0$ such that $(\sqrt{ap}-\sqrt{bp})^2/2 - \varepsilon\log(a/b)/2 > 1$. Let $s\in\{\pm1\}$ be the one that minimizes $\|u_2 - sA^{\mathrm{rs}}u_2^*/\lambda_2^*\|_\infty$. By Corollary 16, we have with probability $1-o(1)$ that,

$$\sqrt{n}\min_{i\in[n]} sz_i(u_2)_i \ge \sqrt{n}\min_{i\in[n]} s^2z_i(A^{\mathrm{rs}}u_2^*)_i/\lambda_2^* - C(\log\log n^{p^2})^{-1},$$

where C is defined in Corollary 16. Note that $s^2 = 1$ and

$$\sqrt{n} z_i (pA^{\mathrm{rs}} u_2^*)_i / \lambda_2^* = \frac{2}{(a-b)\mathrm{log}n} (\sum_{i\sim j} pA^{\mathrm{rs}}_{ij} - \sum_{i\nsim j} pA^{\mathrm{rs}}_{ij}).$$

Hence, applying Lemma 21, we can obtain

$$\mathbb{P}\left(\sqrt{n} \min_{i\in[n]} s^2 z_i (A^{\mathrm{rs}} u_2^*)_i / \lambda_2^* \le \frac{2\varepsilon}{p(a-b)}\right) \le n^{-(\sqrt{ap}-\sqrt{bp})^2/2-\varepsilon\mathrm{log}(a/b)/2} = o(n^{-1}).$$

By the union bound, we further obtain that with probability $1 - o(1)$,

$$\sqrt{n} \min_{i\in[n]} s z_i (u_2)_i \ge \frac{2\varepsilon}{p(a-b)} - o(1) \ge \frac{\varepsilon}{p(a-b)}.$$

Choosing $\eta = \frac{\varepsilon}{p(a-b)}$, we then arrive the conclusion of part (1).

(2) Fix any $\varepsilon_0 > 0$ and take $\eta_0 = [(ap - bp)\mathrm{log}(a/b)/2]^{-1}\varepsilon_0$. Let $s_0 \in \{\pm 1\}$ be the one that minimizes $\|u_2 - sA^{\mathrm{rs}}u_2^*/\lambda_2^*\|_\infty$. Let B_n be the event that (C.2) holds. Let $C(a, b, p)$ be the constant in Corollary 16. When n is large enough, we can have $C(a, b, p) \le \eta_0 \mathrm{loglog} n^{p^2}$, and thus under B_n, we have $\|u_2 - s_0 A^{\mathrm{rs}} u_2^*/\lambda_2^*\|_\infty \le \eta_0/\sqrt{n}$. For all $i \in [n]$, we have the following observations,

$$\{\hat{z}_i \ne s_0 z_i\} \subseteq \{s_0 z_i (u_2)_i \le 0\} \subseteq B_n^c \cup \{s_0^2 z_i (A^{\mathrm{rs}} u_2^*)_i/\lambda_2^* \le \eta_0/\sqrt{n}\}.$$

As mentioned before, $s^2 = 1$ and

$$z_i (pA^{\mathrm{rs}} u_2^*)_i / \lambda_2^* = \frac{2}{(a-b)\mathrm{log}n\sqrt{n}} (\sum_{i\sim j} pA^{\mathrm{rs}}_{ij} - \sum_{i\nsim j} pA^{\mathrm{rs}}_{ij}).$$

Applying Lemma 21 again and using the fact that $\eta_0 = [(ap - bp)\mathrm{log}(a/b)/2]^{-1}\varepsilon_0$, we can obtain

$$\mathbb{P}(z_i (A^{\mathrm{rs}} u_2^*)_i/\lambda_2^* \le \eta_0/\sqrt{n}) \le n^{-\frac{(\sqrt{ap}-\sqrt{bp})^2}{2} + \frac{\varepsilon_0}{2}}.$$

Therefore, the expectation of misclustering rate can be bounded as follows,

$$\begin{aligned}
\mathbb{E}(\hat{z}, z) &= \frac{1}{n}\sum_{i=1}^n \mathbb{P}(B_n^c \cup \{z_i (A^{\mathrm{rs}} u_2^*)_i/\lambda_2^* \le \eta_0/\sqrt{n}\}) \\
&\le \mathbb{P}(B_n^c) + \frac{1}{n}\sum_{i=1}^n \mathbb{P}(\{z_i (A^{\mathrm{rs}} u_2^*)_i/\lambda_2^* \le \eta_0/\sqrt{n}\}) \\
&\le \mathbb{P}(B_n^c) + n^{-\frac{(\sqrt{ap}-\sqrt{bp})^2}{2} + \frac{\varepsilon_0}{2}}.
\end{aligned}$$

It is not hard to observe that for sufficiently large ν, $\mathbb{P}(B_n^c)$ is of smaller order $n^{-\frac{(\sqrt{ap}-\sqrt{bp})^2}{2}}$ as $\sqrt{ap} - \sqrt{bp} \in (0, \sqrt{2}]$. Consequently, taking small enough ε_0, we arrive at the conclusion of part (2). ∎

Appendix D. Auxiliary lemmas

This section includes the auxiliary lemmas that are used for proving the theorems in the paper.

Lemma 17 (Theorem 19 in O'Rourke et al. (2018)) *Consider two matrices B and C with the same dimensions. Suppose matrix B has rank $r(B)$, and denote the jth largest singular value of B as $\sigma_j(B)$. For integer $1 \le j \le r(B)$, suppose matrix V and V' consist of the first j singular vectors of B and C, respectively. Then*

$$\sin(V, V') \le 2\frac{\|B-C\|_2}{\sigma_j(B) - \sigma_{j+1}(B)},$$

where $\sin(V, V') := \|VV^\intercal - V'(V')^\intercal\|_2$.

■

It can be shown that

$$\|VV^\intercal - V'(V')^\intercal\|_2 \ge \frac{\sqrt{2}}{2}\text{inf}_{O\in\mathbb{O}_j}\|V - V'O\|_2,$$

where $\mathbb{O}_j$ denotes the set consisting of orthogonal square matrices with dimension j. Hence we further have

$$\text{inf}_{O\in\mathbb{O}_j}\|V - V'O\|_2 \le 2\sqrt{2}\frac{\|B-C\|_2}{\sigma_j(B) - \sigma_{j+1}(B)},$$

which is actually used in this paper.

Lemma 18 (Proposition 13 in Klopp (2015)) *Let X be an $n \times n$ random matrix with each entry X_{ij} being independent and bounded such that $\max_{ij}|X_{ij}| \le \sigma$. Define*

$$\sigma_1 = \text{max}_i\sqrt{\mathbb{E}\sum_j X_{ij}^2} \quad \text{and} \quad \sigma_2 = \text{max}_j\sqrt{\mathbb{E}\sum_i X_{ij}^2}.$$

Then, for any $\nu > 0$, there exists constant $c = c(\sigma, \nu) > 0$ such that,

$$\|X\|_2 \le c\max(\sigma_1, \sigma_2, \sqrt{\text{log}n}),$$

with probability larger than $1 - n^{-\nu}$.

■

Lemma 19 (Corollary 4 and Theorem 5 in Gittens and Tropp (2009)) *Suppose B is a fixed matrix, and let X be a random matrix with each entry X_{jk} being independent and bounded such that $\max_{jk}|X_{jk}| \le \frac{D}{2}$ almost surely, for which $\mathbb{E}(X) = B$. Then for all $\delta > 0$,*

$$\|X - B\|_2 \le (1+\delta)\mathbb{E}\|B - X\|_2,$$

with probability larger than $1 - \exp^{-\delta^2(\mathbb{E}\|X-B\|_2)^2/4D^2}$. Further,

$$\begin{aligned}\mathbb{E}\|X - B\|_2 \le c\,\Big(&\max_j\big(\sum_k \text{Var}(X_{jk})\big)^{1/2} \\ &+ \max_k\big(\sum_j \text{Var}(X_{jk})\big)^{1/2} + \big(\sum_{jk}\mathbb{E}(X_{jk} - b_{jk})^4\big)^{1/4}\Big).\end{aligned}$$

■

Lemma 20 (Lemma 7 in Abbe et al. (2020)) *Let $w \in \mathbb{R}^n$ be a fixed vector, $\{X_i\}_{i=1}^n$ be independent random variables where $X_i \sim \text{Bernoulli}(p_i)$. Suppose $\bar{p} \geq \max_i p_i$ and $\alpha \geq 0$. Then,*

$$\mathbb{P}\left(\left|\sum_{i=1}^n w_i(X_i - \mathbb{E}(X_i))\right| \geq \frac{(2+\alpha\bar{p}n)}{1 \vee \log(\frac{\sqrt{n}\|w\|_\infty}{\|w\|_2})}\|w\|_\infty\right) \leq e^{-\alpha n\bar{p}}.$$

■

Lemma 21 (Lemma 8 in Abbe et al. (2020)) *Suppose $a > b$, $\{W_i\}_{i=1}^{n/2}$ are i.i.d.* Bernoulli($\frac{a\log n}{n}$)*, and $\{Z_i\}_{i=1}^{n/2}$ are i.i.d.* Bernoulli($\frac{b\log n}{n}$)*, independent of $\{W_i\}_{i=1}^{n/2}$. For any $\varepsilon \in \mathbb{R}$, the following tail bound holds,*

$$\mathbb{P}\left(\sum_{i=1}^{n/2} W_i - \sum_{i=1}^{n/2} Z_i \leq \varepsilon \log n\right) \leq n^{-(\sqrt{a}-\sqrt{b})^2/2+\varepsilon\log(a/b)/2}.$$

■

Appendix E. Additional experimental results

This section provides the additional experimental details and results that are not shown in the main text.

E.1 Additional simulation experimental results

The following provides the model set-up 3-8 in the simulation experiments.

Model set-up 3 (Networks with core-periphery structure) In such networks, there exists a set of nodes (termed as core nodes) which send and receive edges with each other, and they also receive edges from another set of nodes (termed as periphery nodes) which has no incoming edges. Thus, the sending and receiving clusters are different. In this model-set up, $K^y = 1$ and $K^z = 2$, and we consider the following two cases for the link probability matrix B:

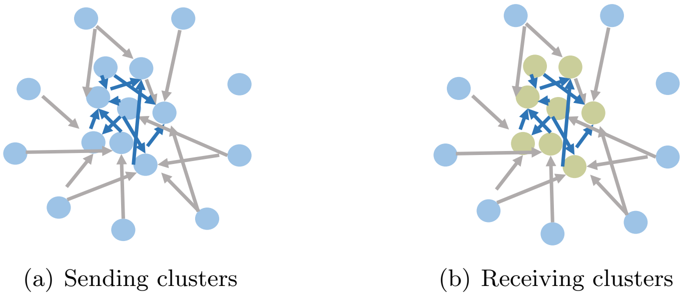
$$B_1 := \begin{bmatrix} 0.05 & 0 \end{bmatrix}, \quad B_2 := \begin{bmatrix} \frac{1.8\log n}{n} & 0 \end{bmatrix}.$$

Figure 16 shows the topological example of model set-up 3.

Model set-up 4 (Multi-layer networks) In such networks, nodes with in each "layer" share similar sending *and* receiving patterns, while nodes across different layers might also share similar sending *or* receiving patterns. Thus considering the directions of edges, the sending and receiving clusters may be different. In this model-set up, $K^y = 2$ and $K^z = 3$, and we consider the following two cases for the link probability matrix B:

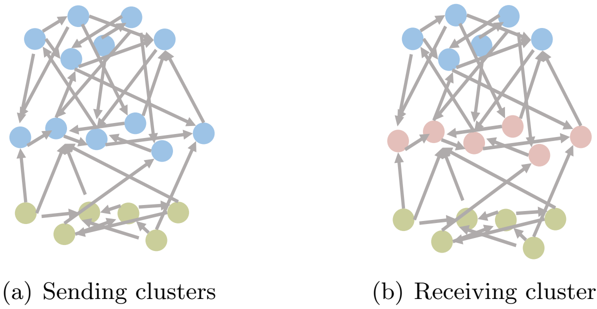
$$B_1 := \begin{bmatrix} 0.1 & 0.1 & 0 \\ 0 & 0.04 & 0.1 \end{bmatrix}, \quad B_2 := \begin{bmatrix} \frac{1}{\sqrt{n}} & \frac{1}{\sqrt{n}} & 0 \\ 0 & \frac{1}{\sqrt{n}} & \frac{1}{\sqrt{n}} \end{bmatrix}.$$

Figure 17 shows the topological example of model set-up 4.



(a) Sending clusters (b) Receiving clusters

Figure 16: Illustration for networks under model set-up 3. Colors indicate clusters.



(a) Sending clusters (b) Receiving clusters

Figure 17: Illustration for networks under model set-up 4. Colors indicate clusters.

Model set-up 5 (Networks with 'message loop') In such networks, edges only exist between nodes belonging to different clusters. The information flow from one cluster to another and finally forms a 'message loop'. The sending and receiving clusters are the same but as we can imagine, clustering based on symmetrized adjacency matrix would not perform well. In this model-set up, $K^y = K^z = 3$, and we consider the following two cases for the link probability matrix B:

$$B_1 := \begin{bmatrix} 0 & 0.05 & 0 \\ 0 & 0 & 0.05 \\ 0.1 & 0 & 0 \end{bmatrix}, \quad B_2 := \begin{bmatrix} 0 & \frac{1}{2\sqrt{n}} & 0 \\ 0 & 0 & \frac{1}{2\sqrt{n}} \\ \frac{1}{2\sqrt{n}} & 0 & 0 \end{bmatrix}.$$

Figure 18 shows the topological example of model set-up 5.

Model set-up 6 (Networks of 'message passing') The network structure is similar to that in model set-up 4 except that the edges across different layers start from upper layers down to lower layers, just like passing messages. The sending and receiving clusters are the same but as we can imagine, treating these networks as undirected networks does not work well. In this model-set up, $K^y = K^z = 4$, and we consider the following two cases for the

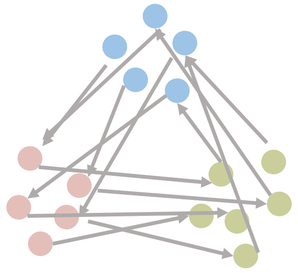


Figure 18: Illustration for networks under model set-up 5. Sending and receiving clusters are the same. Colors indicate clusters.

link probability matrix B:

$$B_1 := \begin{bmatrix} 0.2 & 0.1 & 0.05 & 0.01 \\ 0 & 0.2 & 0.1 & 0.05 \\ 0 & 0 & 0.2 & 0.1 \\ 0 & 0 & 0 & 0.2 \end{bmatrix}, \quad B_2 := \begin{bmatrix} \frac{2\mathrm{log}n}{n} & \frac{\mathrm{log}n}{2n} & \frac{\mathrm{log}n}{2n} & \frac{\mathrm{log}n}{2n} \\ 0 & \frac{2\mathrm{log}n}{n} & \frac{\mathrm{log}n}{2n} & \frac{\mathrm{log}n}{2n} \\ 0 & 0 & \frac{2\mathrm{log}n}{n} & \frac{\mathrm{log}n}{2n} \\ 0 & 0 & 0 & \frac{2\mathrm{log}n}{n} \end{bmatrix}.$$

Figure 19 shows the topological example of model set-up 6.

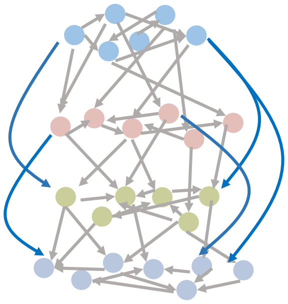


Figure 19: Illustration for networks under model set-up 6. Sending and receiving clusters are the same. Colors indicate clusters.

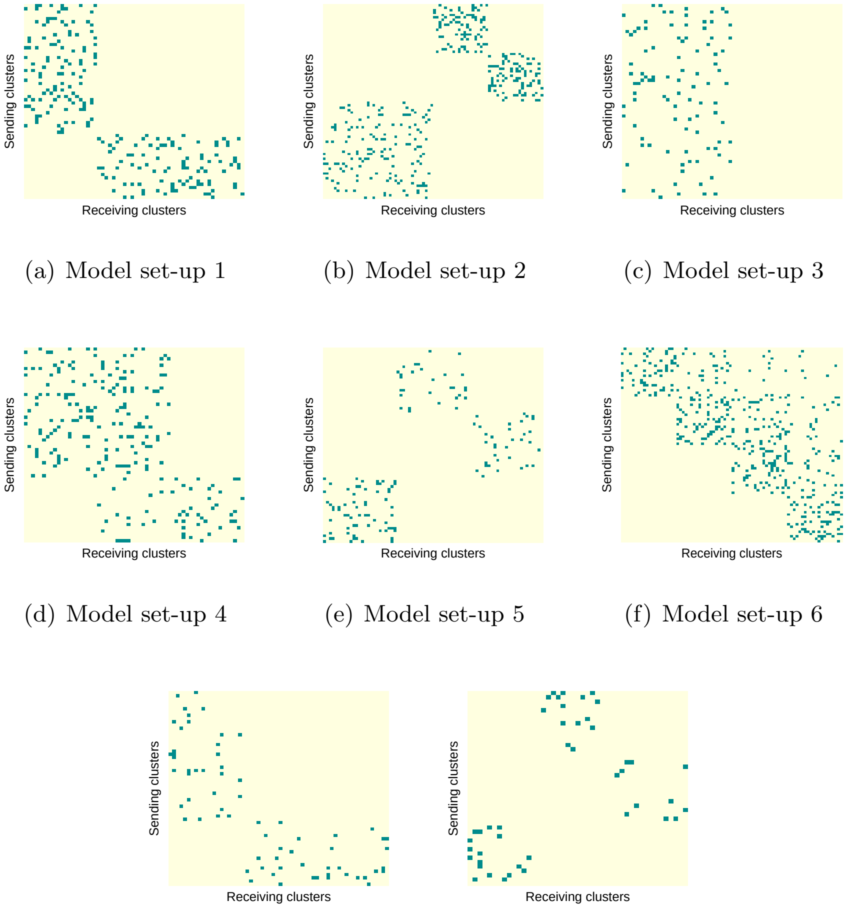
Model set-up 7 (Networks in model set-up 1 with degree heterogeneity) The basic set-up is similar to model 1 except that we also incorporate the degree heterogeneity. See also Figure 1 for the topological structure example. In this model-set up, $K^y = K^z = 2$, each θ_i^y is generated i.i.d. to be 0.3 w.h.p. 0.8 and 1 w.h.p. 0.2, and θ_i^z is generated in a similar way. We consider the following two cases for the link probability matrix B:

$$B_1 := \begin{bmatrix} 0.3 & 0 \\ 0 & 0.3 \end{bmatrix}, \quad B_2 := \begin{bmatrix} \frac{3}{\sqrt{n}} & 0 \\ 0 & \frac{3}{\sqrt{n}} \end{bmatrix}.$$

Model set-up 8 (Networks in model set-up 5 with degree heterogeneity) The basic set-up is similar to model 5 except that the degree heterogeneity is also encoded. See also Figure 18 for the topological structure example. In this model-set up, $K^y = K^z = 3$, the entries of θ^y are 0.5 except the first three ones, and $\theta^z = \theta^y$. We consider the following two cases for the link probability matrix B:

$$B_1 := \begin{bmatrix} 0 & 0.2 & 0 \\ 0 & 0 & 0.2 \\ 0.2 & 0 & 0 \end{bmatrix}, \quad B_2 := \begin{bmatrix} 0 & \frac{2}{\sqrt{n}} & 0 \\ 0 & 0 & \frac{2}{\sqrt{n}} \\ \frac{2}{\sqrt{n}} & 0 & 0 \end{bmatrix}.$$

Figure 20 show the matrix representations of the eight model set-ups (the first two models are in the main text) in the simulation. Figure 21-34 show the experimental results that not displayed in the main text.



(a) Model set-up 1

(b) Model set-up 2

(c) Model set-up 3

(d) Model set-up 4

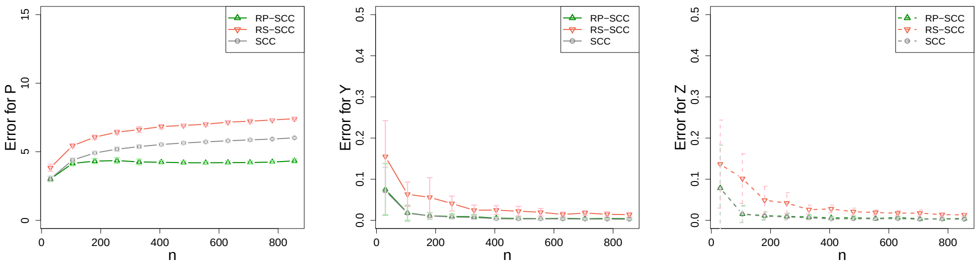
(e) Model set-up 5

(f) Model set-up 6

(g) Model set-up 7

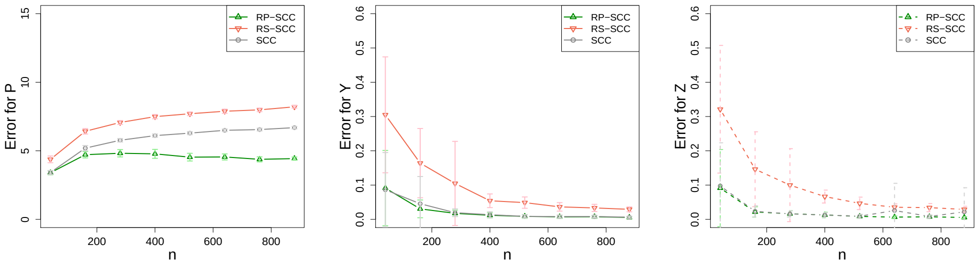
(h) Model set-up 8

Figure 20: Matrix representation of eight model set-ups considered in simulations. Darker entries are 1's and lighter entries are 0's. The column-wise and row-wise block structures reveal the sending and receiving clusters, respectively.



(a) (b) (c)

Figure 21: Simulation results of case 2 under model set-up 1.



(a) (b) (c)

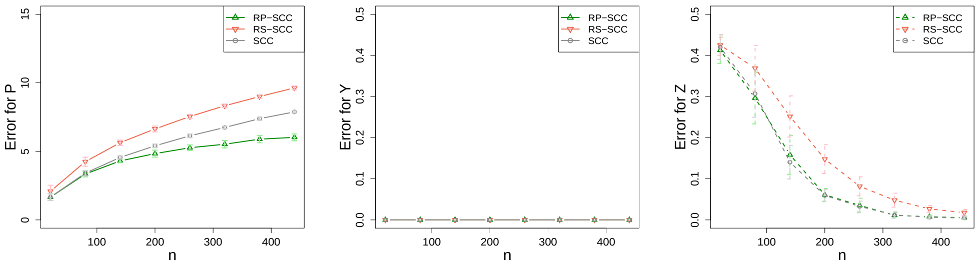
Figure 22: Simulation results of case 2 under model set-up 2.

E.2 Additional real data experimental results

Figure 35 shows receiving clusters for the statistical citation network detected by SCC and five SCC-based approximate algorithm, where the embedding of nodes corresponding to the three right singular vectors are shown and colors indicate receiving clusters.

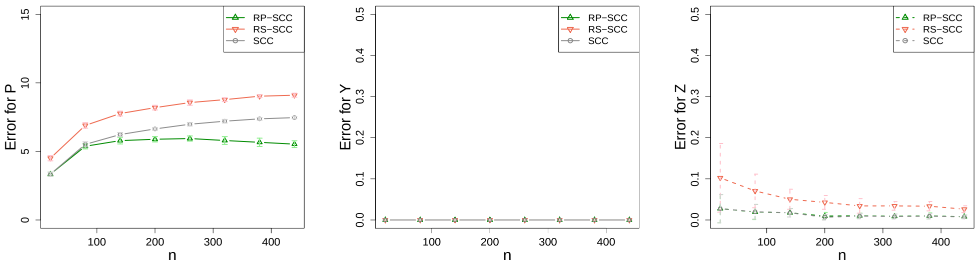
The results corresponding to the email dataset are shown in the following figures. Figure 36 shows the plot of singular values. Figure 37 shows the histogram of movement scores of different methods. Figure 38 (Figure 41) displays the optimal cluster number selected by the average silhouette method based on the SCC (SsCC). Figure 39 (Figure 43) shows the box plot of the relative ARI between the SCC (SsCC) and SCC-based (SsCC-based) five approximate methods. Figure 40 and 42 show that sending and receiving clusters detected by SCC and five SCC-based approximate algorithms.

Figure 44 and 45 display the additional results for the five large-scale real networks, where the averaged ARI and the corresponding standard deviations associated with receiving clusters over 20 replications are shown, respectively.



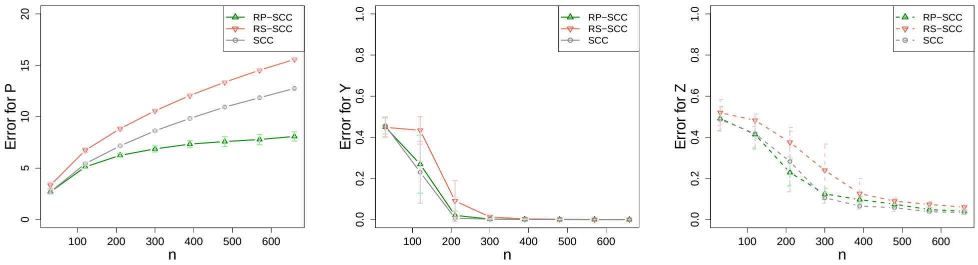
(a) (b) (c)

Figure 23: Simulation results of case 1 under model set-up 3.



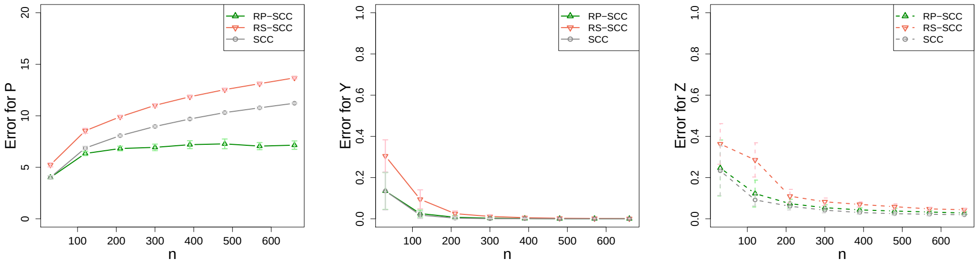
(a) (b) (c)

Figure 24: Simulation results of case 2 under model set-up 3.



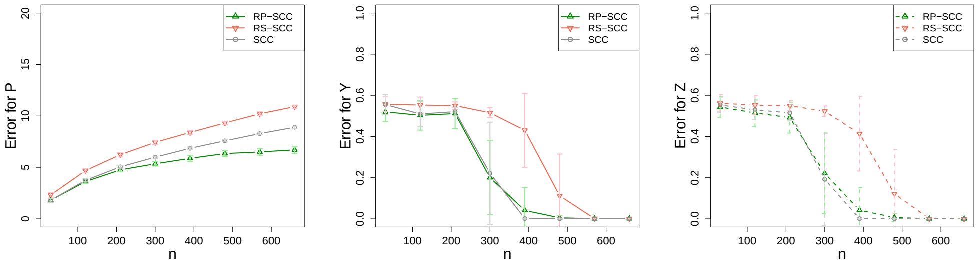
(a) (b) (c)

Figure 25: Simulation results of case 1 under model set-up 4.



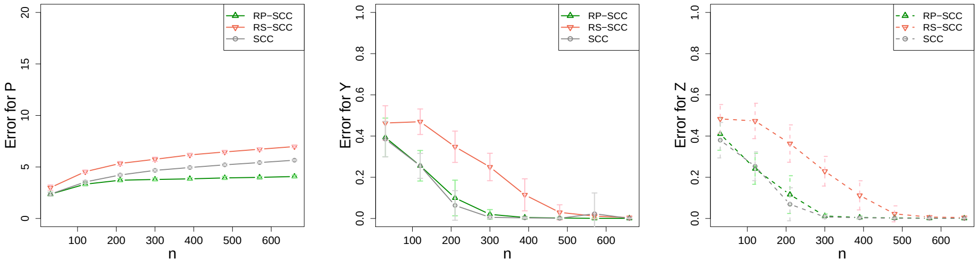
(a) (b) (c)

Figure 26: Simulation results of case 2 under model set-up 4.



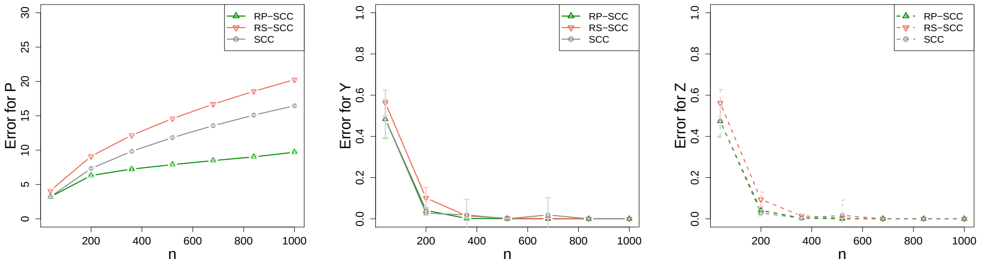
(a) (b) (c)

Figure 27: Simulation results of case 1 under model set-up 5.



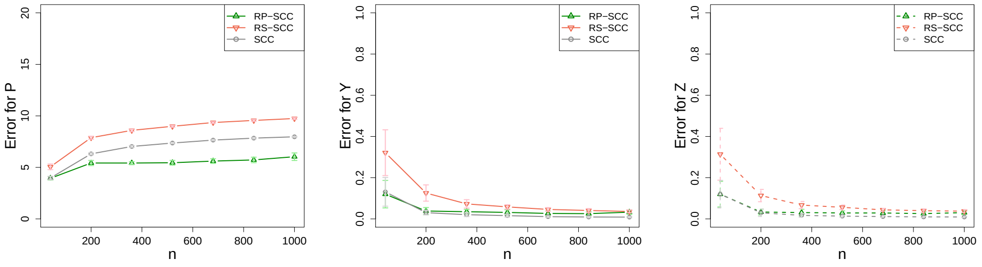
(a) (b) (c)

Figure 28: Simulation results of case 2 under model set-up 5.



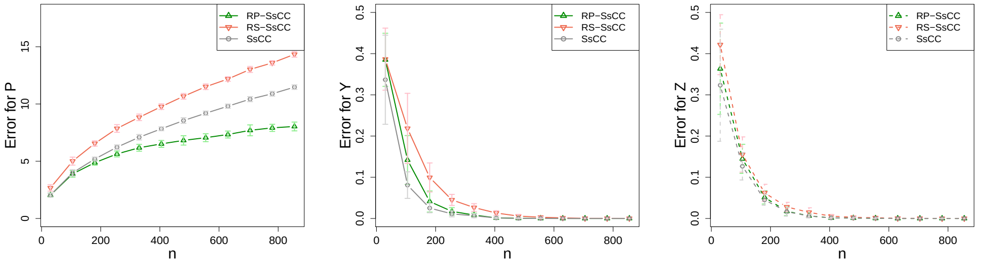
(a) (b) (c)

Figure 29: Simulation results of case 1 under model set-up 6.



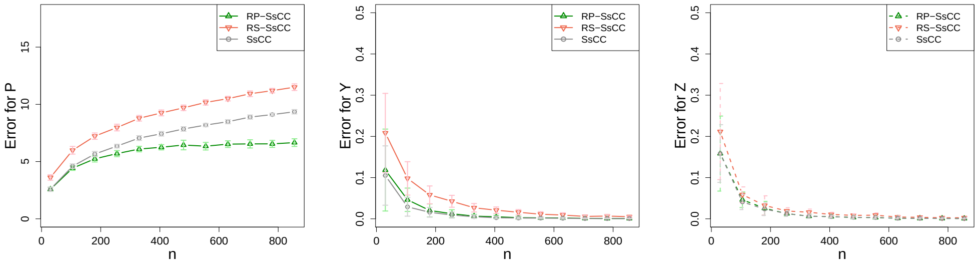
(a) (b) (c)

Figure 30: Simulation results of case 2 under model set-up 6.



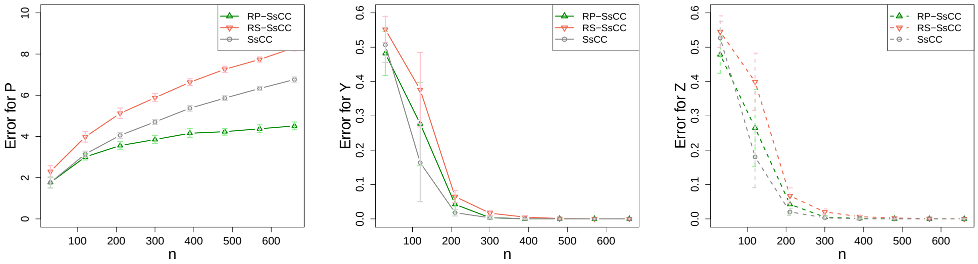
(a) (b) (c)

Figure 31: Simulation results of case 1 under model set-up 7.



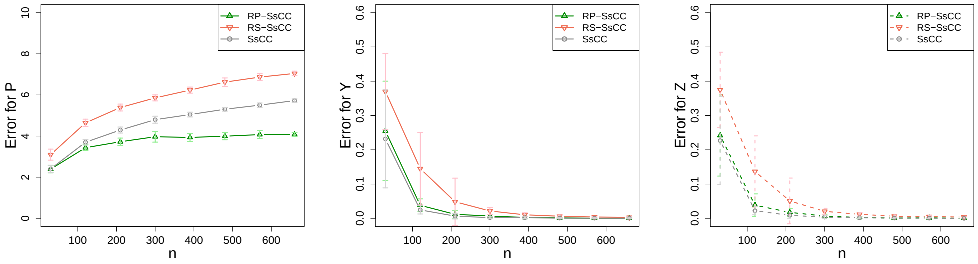
(a) (b) (c)

Figure 32: Simulation results of case 2 under model set-up 7.



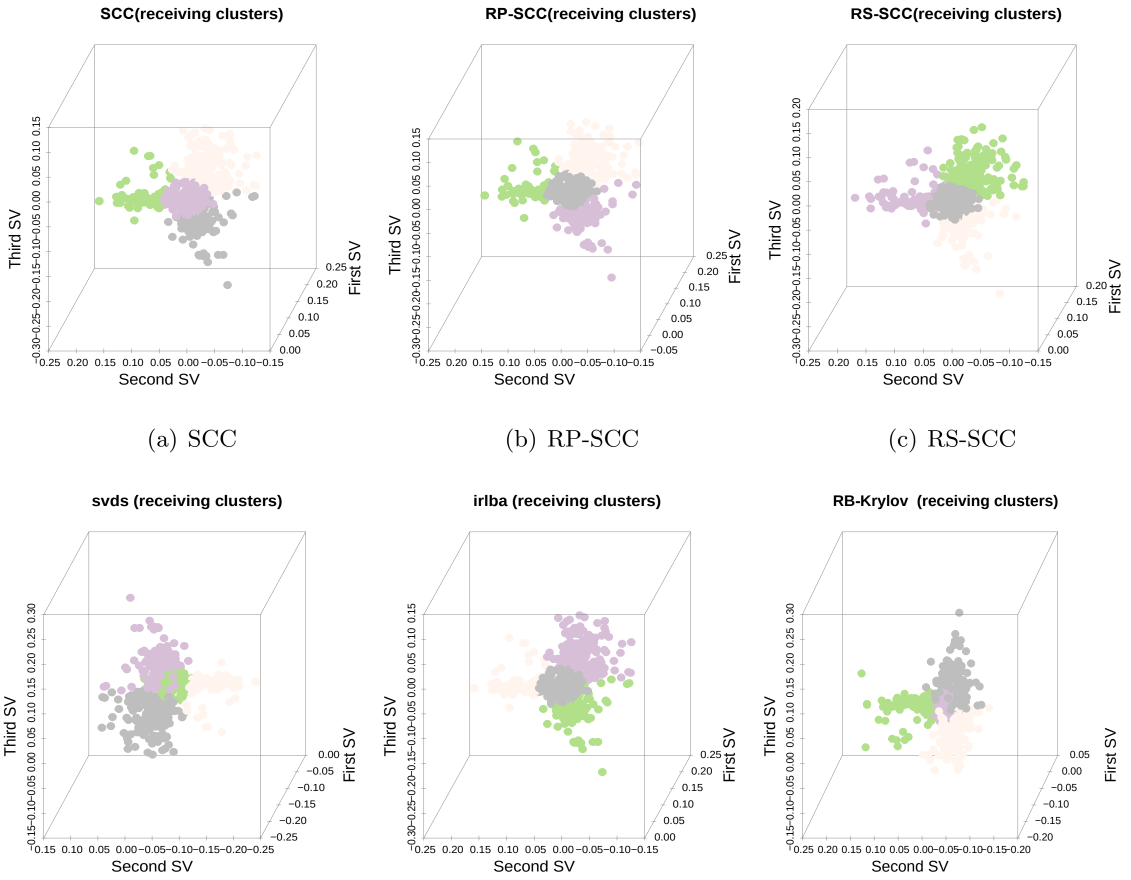
(a) (b) (c)

Figure 33: Simulation results of case 1 under model set-up 8.



(a) (b) (c)

Figure 34: Simulation results of case 2 under model set-up 8.



(a) SCC

(b) RP-SCC

(c) RS-SCC

(d) svds

(e) irlba

(f) RB-Krylov

Figure 35: Receiving clusters of the statisticians citation network detected by SCC and five SCC-based approximate algorithms.

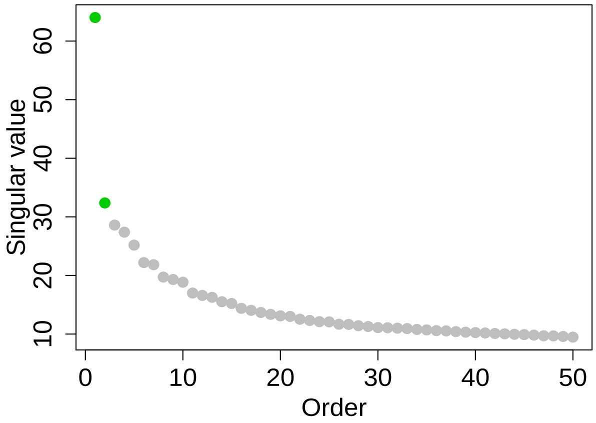
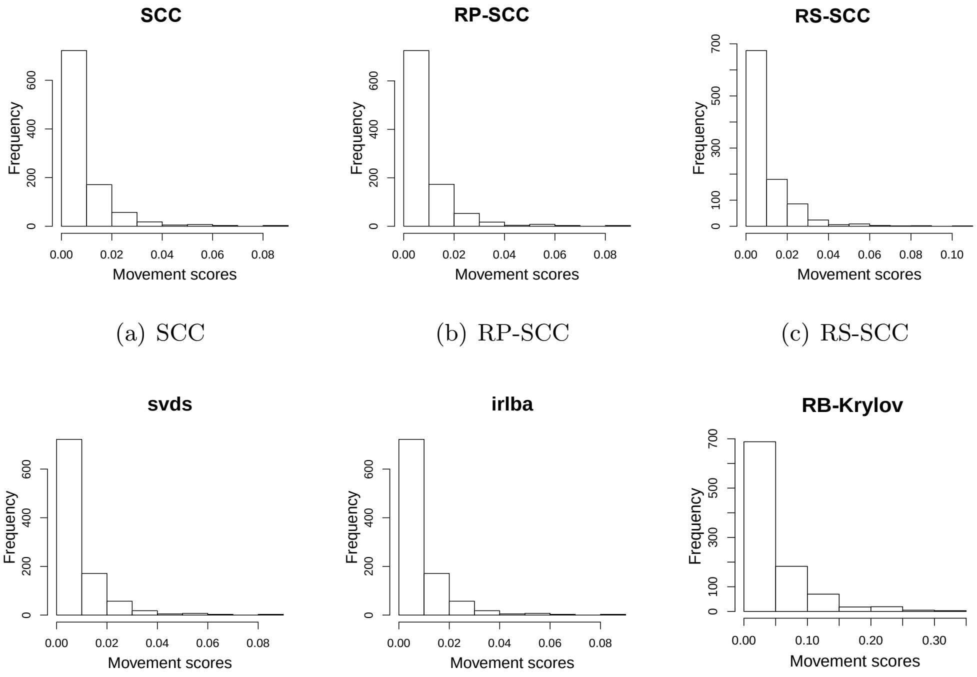


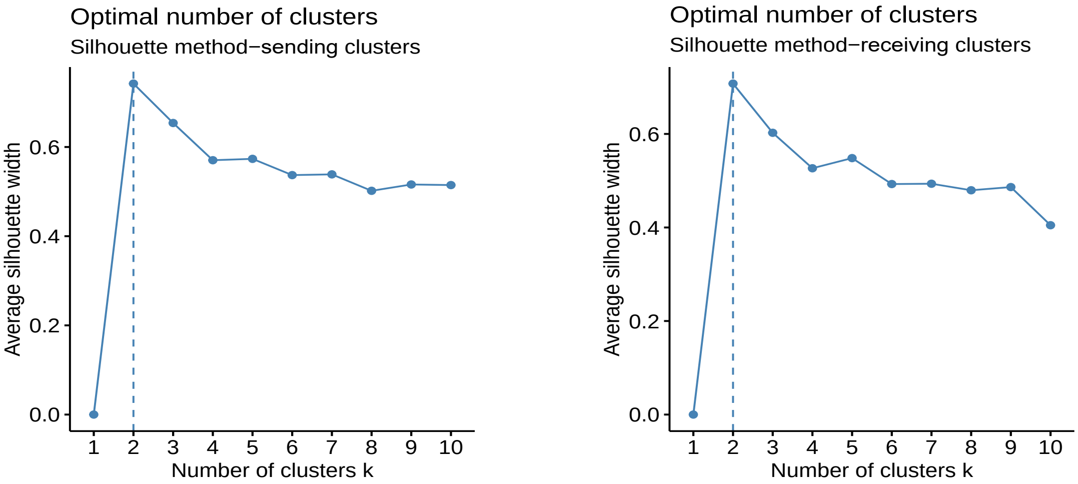
Figure 36: The top 50 singular values of the adjacency matrix of the European email network.



(a) SCC (b) RP-SCC (c) RS-SCC

(d) svds (e) irlba (f) RB-Krylov

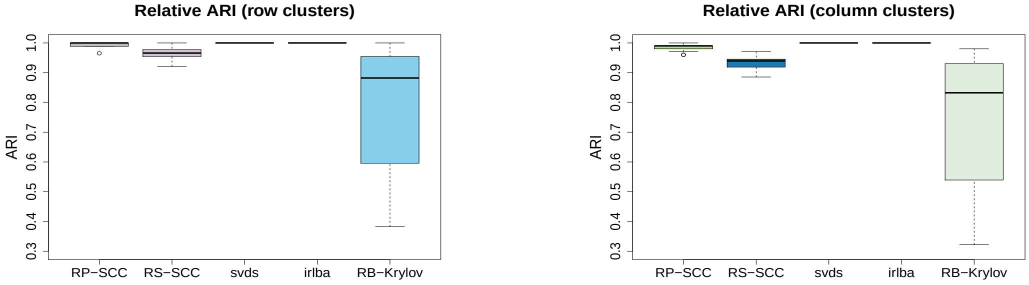
Figure 37: Histogram of movement scores of different methods for the European email network.



(a) Sending clusters

(b) Receiving clusters

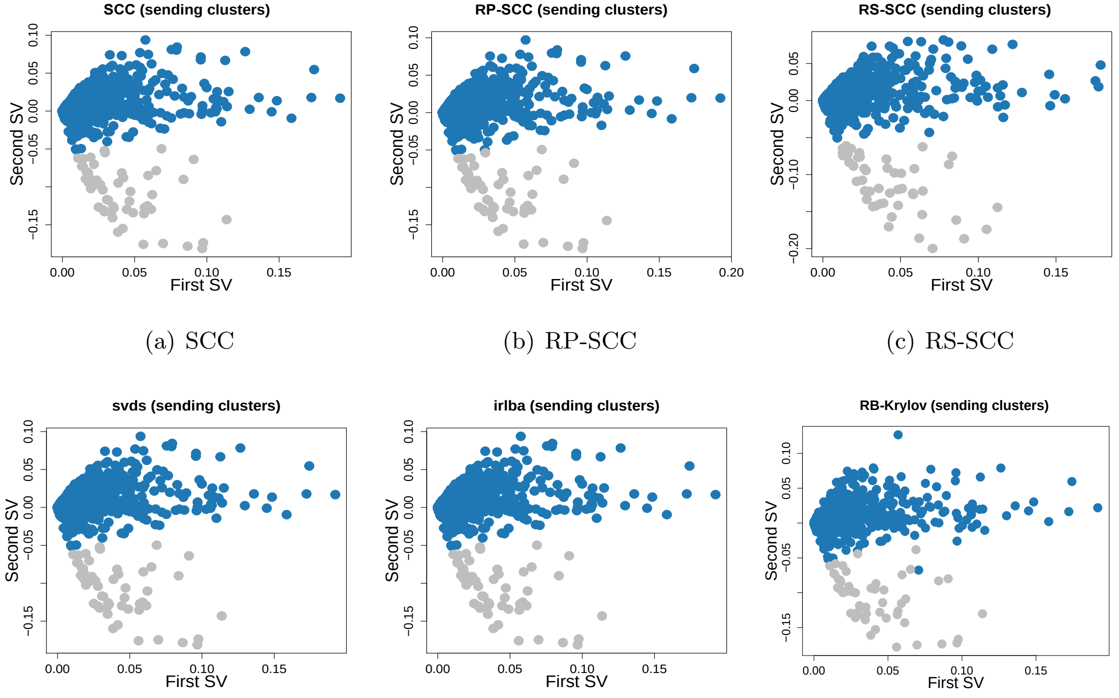
Figure 38: Optimal number of clusters of the European email network selected by the average silhouette method based on the SCC.



(a) Sending clusters

(b) Receiving clusters

Figure 39: Relative ARI between the SCC and SCC-based five approximate methods on the European email network.



(a) SCC

(b) RP-SCC

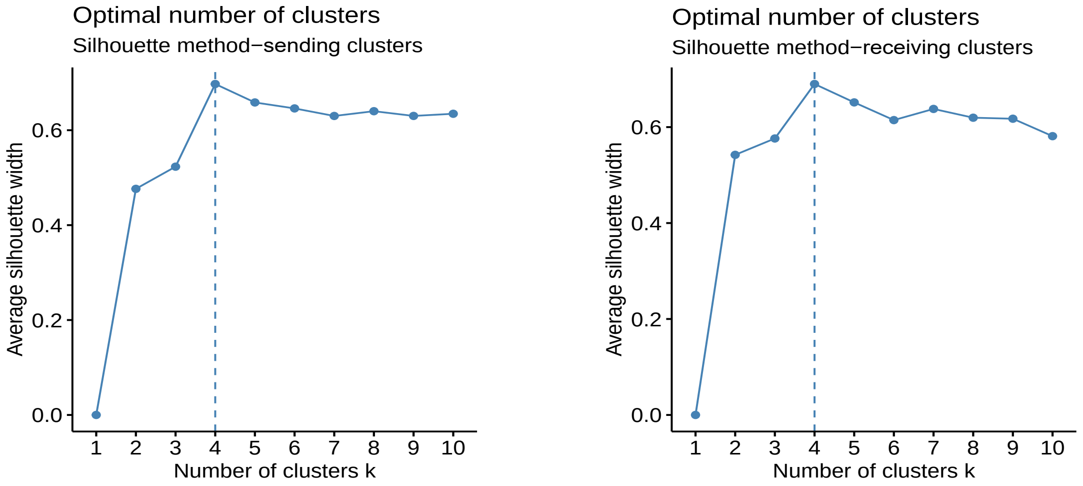
(c) RS-SCC

(d) svds

(e) irlba

(f) RB-Krylov

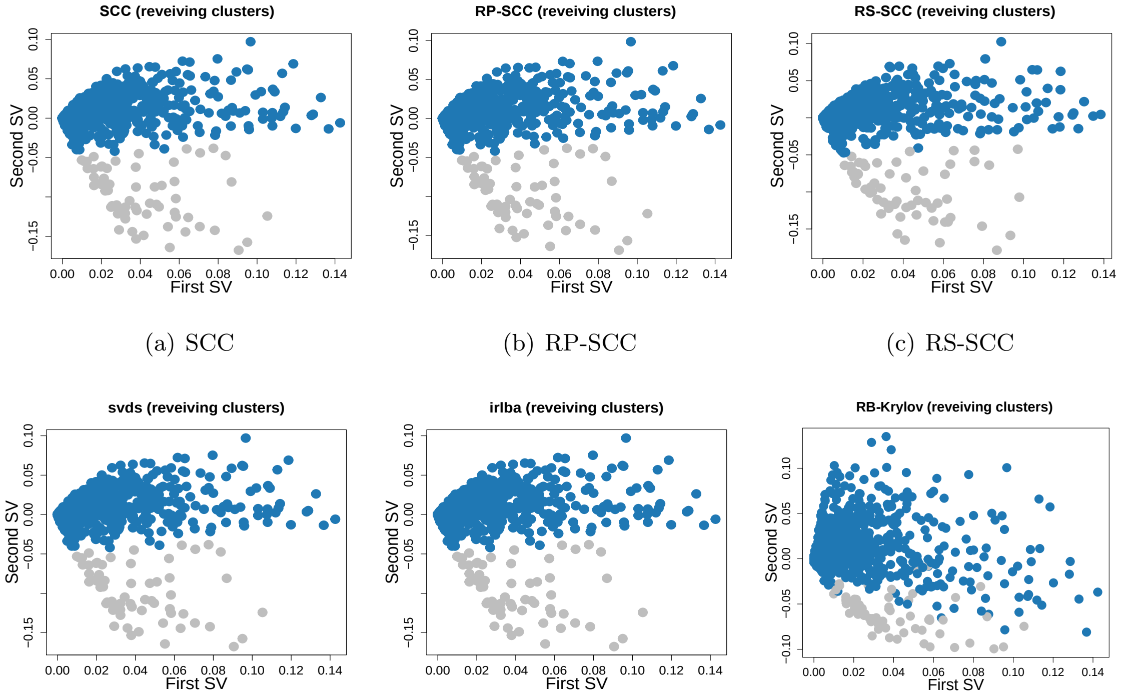
Figure 40: Sending clusters of the European email network detected by SCC and five SCC-based approximate algorithms.



(a) Sending clusters

(b) Receiving clusters

Figure 41: Optimal number of clusters of the European email network selected by the average silhouette method based on the SsCC.



(a) SCC

(b) RP-SCC

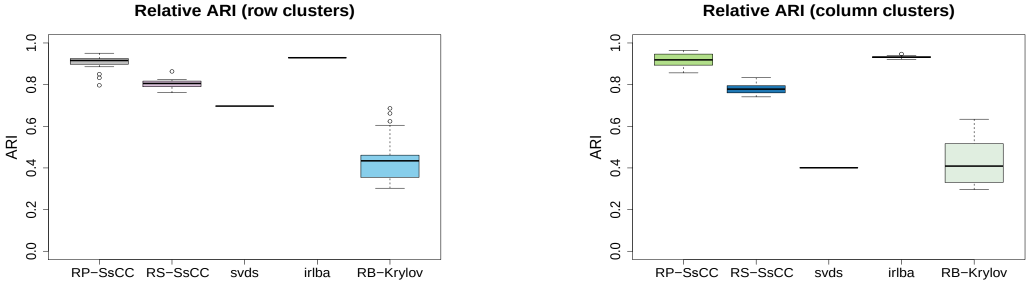
(c) RS-SCC

(d) svds

(e) irlba

(f) RB-Krylov

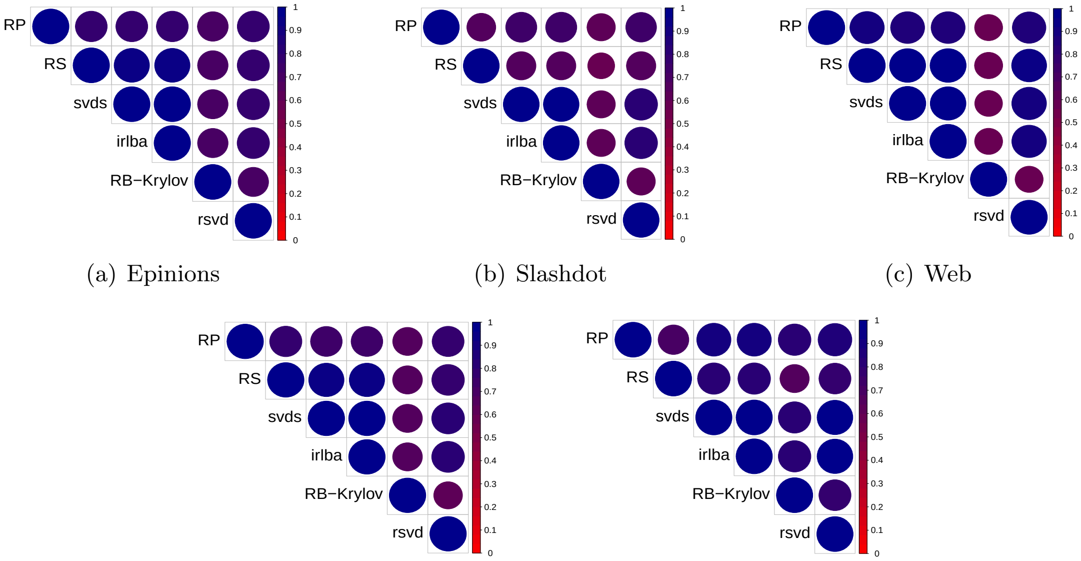
Figure 42: Receiving clusters of the European email network detected by SCC and five SCC-based approximate algorithms.



(a) Sending clusters

(b) Receiving clusters

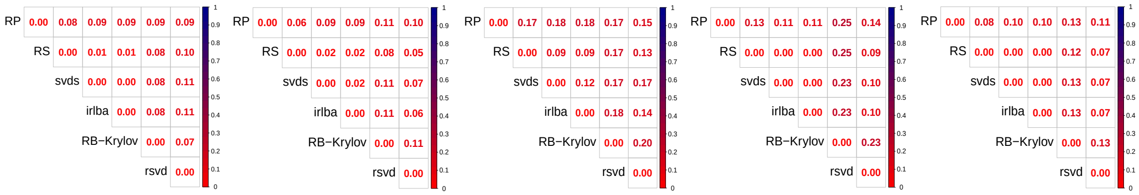
Figure 43: Relative ARI between the SsCC and SsCC-based five approximate methods on the European email network.



(a) Epinions (b) Slashdot (c) Web

(d) Wiki-cat (e) Wiki-talk

Figure 44: The pairwise comparison of the column clusters of six methods on five large-scale networks. The relative clustering performance are measured by ARI. Larger ARI, i.e., larger circles in the figure, indicates that the clustering results of the two associated methods are more close.



(a) Epinions (b) Slashdot (c) Web (d) Wiki-cat (e) Wiki-talk

Figure 45: The standard deviations corresponding to the pairwise ARI of column clusters.

References

Emmanuel Abbe. Community detection and stochastic block models: recent developments. *Journal of Machine Learning Research*, 18(1):6446–6531, 2018.

Emmanuel Abbe, Afonso S Bandeira, and Georgina Hall. Exact recovery in the stochastic block model. *IEEE Transactions on Information Theory*, 62(1):471–487, 2015.

Emmanuel Abbe, Jianqing Fan, Kaizheng Wang, Yiqiao Zhong, et al. Entrywise eigenvector analysis of random matrices with low expected rank. *Annals of Statistics*, 48(3):1452–1474, 2020.

Raman Arora, Andrew Cotter, and Nathan Srebro. Stochastic optimization of PCA with capped msg. *arXiv preprint arXiv:1307.1674*, 2013.

Jesús Arroyo and Elizaveta Levina. Overlapping community detection in networks via sparse spectral decomposition. *Sankhya A*, 84(1):1–35, 2022.

James Baglama and Lothar Reichel. Augmented implicitly restarted lanczos bidiagonalization methods. *SIAM Journal on Scientific Computing*, 27(1):19–42, 2005.

Jim Baglama, Lothar Reichel, and B. W. Lewis. *irlba: Fast Truncated Singular Value Decomposition and Principal Components Analysis for Large Dense and Sparse Matrices*, 2019. URL `https://CRAN.R-project.org/package=irlba`. R package version 2.3.3.

Daniel Boley, Gyan Ranjan, and Zhi-Li Zhang. Commute times for a directed graph using an asymmetric laplacian. *Linear Algebra and its Applications*, 435(2):224–242, 2011.

Daniela Calvetti, Lothar Reichel, and Danny Chris Sorensen. An implicitly restarted lanczos method for large symmetric eigenvalue problems. *Electronic Transactions on Numerical Analysis*, 2:1–21, 1994.

Xiangyu Chang, Danyang Huang, and Hansheng Wang. A popularity scaled latent space model for large-scale directed social network. *Statistica Sinica*, 29:1277–1299, 2019.

Peter Chin, Anup Rao, and Van Vu. Stochastic block model and community detection in sparse graphs: A spectral algorithm with optimal rate of recovery. In *Conference on Learning Theory*, pages 391–423, 2015.

Fan Chung. Laplacians and the Cheeger inequality for directed graphs. *Annals of Combinatorics*, 9(1):1–19, 2005.

Kenneth L Clarkson and David P Woodruff. Low-rank approximation and regression in input sparsity time. *Journal of the ACM (JACM)*, 63(6):1–45, 2017.

Inderjit S Dhillon. Co-clustering documents and words using bipartite spectral graph partitioning. In *Proceedings of the seventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 269–274, 2001.

Petros Drineas and Michael W Mahoney. Randnla: randomized numerical linear algebra. *Communications of the ACM*, 59(6):80–90, 2016.

Petros Drineas, Michael W Mahoney, and Shan Muthukrishnan. Sampling algorithms for ℓ_2 regression and applications. In *Proceedings of the seventeenth Annual ACM-SIAM Symposium on Discrete Algorithm*, pages 1127–1136. Society for Industrial and Applied Mathematics, 2006.

N Benjamin Erichson, Sergey Voronin, Steven L Brunton, and J Nathan Kutz. Randomized matrix decompositions using R. *Journal of Statistical Software*, 89(1):1–48, 2019.

Uriel Feige and Eran Ofek. Spectral techniques applied to sparse random graphs. *Random Structures & Algorithms*, 27(2):251–275, 2005.

Santo Fortunato. Community detection in graphs. *Physics Reports*, 486(3-5):75–174, 2010.

Chao Gao, Zongming Ma, Anderson Y Zhang, and Harrison H Zhou. Achieving optimal misclassification proportion in stochastic block models. *Journal of Machine Learning Research*, 18(1):1980–2024, 2017.

Alex Gittens and Joel A Tropp. Error bounds for random matrix approximation schemes. *arXiv preprint arXiv:0911.4108*, 2009.

Anna Goldenberg, Alice X Zheng, Stephen E Fienberg, Edoardo M Airoldi, et al. A survey of statistical network models. *Foundations and Trends® in Machine Learning*, 2(2):129–233, 2010.

Gene Golub and William Kahan. Calculating the singular values and pseudo-inverse of a matrix. *Journal of the Society for Industrial and Applied Mathematics, Series B: Numerical Analysis*, 2(2):205–224, 1965.

Gene Golub and Christian Reinsch. Singular value decomposition and least squares solutions. numerische mathematik. *Numerische Mathematik*, 14:403–420, 1970.

Nathan Halko, Per-Gunnar Martinsson, and Joel A Tropp. Finding structure with randomness: Probabilistic algorithms for constructing approximate matrix decompositions. *SIAM Review*, 53(2):217–288, 2011.

John A Hartigan. Direct clustering of a data matrix. *Journal of the American Statistical Association*, 67(337):123–129, 1972.

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

Lawrence Hubert and Phipps Arabie. Comparing partitions. *Journal of Classification*, 2 (1):193–218, 1985.

Pengsheng Ji and Jiashun Jin. Coauthorship and citation networks for statisticians. *The Annals of Applied Statistics*, 10(4):1779–1812, 2016.

Zhongxiao Jia and Datian Niu. An implicitly restarted refined bidiagonalization lanczos method for computing a partial singular value decomposition. *SIAM Journal on Matrix Analysis and Applications*, 25(1):246–265, 2003.

Zhongxiao Jia and Datian Niu. A refined harmonic lanczos bidiagonalization method and an implicitly restarted algorithm for computing the smallest singular triplets of large matrices. *SIAM Journal on Scientific Computing*, 32(2):714–744, 2010.

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

Olga Klopp. Matrix completion by singular value thresholding: sharp bounds. *Electronic Journal of Statistics*, 9(2):2348–2369, 2015.

Amit Kumar, Yogish Sabharwal, and Sandeep Sen. A simple linear time (1+ epsilon)-approximation algorithm for k-means clustering in any dimensions. In *Annual Symposium on Foundations of Computer Science*, volume 45, pages 454–462. IEEE Computer Society Press, 2004.

Jing Lei and Alessandro Rinaldo. Consistency of spectral clustering in stochastic block models. *The Annals of Statistics*, 43(1):215–237, 2015.

Jing Lei, Kehui Chen, and Brian Lynch. Consistent community detection in multi-layer network data. *Biometrika*, 107(1):61–73, 2020.

Jure Leskovec, Kevin J Lang, Anirban Dasgupta, and Michael W Mahoney. Community structure in large networks: Natural cluster sizes and the absence of large well-defined clusters. *Internet Mathematics*, 6(1):29–123, 2009.

Jure Leskovec, Daniel Huttenlocher, and Jon Kleinberg. Predicting positive and negative links in online social networks. In *Proceedings of the 19th International Conference on World Wide Web*, pages 641–650, 2010.

Miaoqi Li and Emily L Kang. Randomized algorithms of maximum likelihood estimation with spatial autoregressive models for large-scale networks. *Statistics and Computing*, 29 (5):1165–1179, 2019.

Zhenyu Liao, Romain Couillet, and Michael W Mahoney. Sparse quantized spectral clustering. *arXiv preprint arXiv:2010.01376*, 2020.

Ping Ma, Michael W Mahoney, and Bin Yu. A statistical perspective on algorithmic leveraging. *Journal of Machine Learning Research*, 16(1):861–911, 2015.

Shujie Ma, Liangjun Su, and Yichong Zhang. Determining the number of communities in degree-corrected stochastic block models. *Journal of Machine Learning Research*, 22(1): 3217–3279, 2021.

Michael W Mahoney et al. Randomized algorithms for matrices and data. *Foundations and Trends® in Machine Learning*, 3(2):123–224, 2011.

Christopher Manning, Prabhakar Raghavan, and Hinrich Schütze. Introduction to information retrieval. *Natural Language Engineering*, 16(1):100–103, 2010.

Per-Gunnar Martinsson. Randomized methods for matrix computations. *arXiv preprint arXiv:1607.01649*, 2016.

Xiangrui Meng and Michael W Mahoney. Low-distortion subspace embeddings in input-sparsity time and applications to robust linear regression. In *Proceedings of the forty-fifth Annual ACM Symposium on Theory of Computing*, pages 91–100, 2013.

Elchanan Mossel, Joe Neeman, and Allan Sly. Consistency thresholds for the planted bisection model. In *Proceedings of the forty-seventh Annual ACM symposium on Theory of Computing*, pages 69–75, 2015.

Cameron Musco and Christopher Musco. Randomized block krylov methods for stronger and faster approximate singular value decomposition. *Advances in Neural Information Processing Systems*, 28, 2015.

Jelani Nelson and Huy L Nguyên. Osnap: Faster numerical linear algebra algorithms via sparser subspace embeddings. In *2013 IEEE 54th Annual Symposium on Foundations of Computer Science*, pages 117–126. IEEE, 2013.

Mark Newman. *Networks*. Oxford university press, 2018.

Erkki Oja and Juha Karhunen. On stochastic approximation of the eigenvectors and eigenvalues of the expectation of a random matrix. *Journal of Mathematical Analysis and Applications*, 106(1):69–84, 1985.

Sean O'Rourke, Van Vu, and Ke Wang. Random perturbation of low rank matrices: Improving classical bounds. *Linear Algebra and its Applications*, 540:26–59, 2018.

Mert Pilanci and Martin J Wainwright. Iterative hessian sketch: Fast and accurate solution approximation for constrained least-squares. *Journal of Machine Learning Research*, 17 (1):1842–1879, 2016.

Tai Qin and Karl Rohe. Regularized spectral clustering under the degree-corrected stochastic blockmodel. In *Advances in Neural Information Processing Systems*, pages 3120–3128, 2013.

Yixuan Qiu and Jiali Mei. *RSpectra: Solvers for Large-Scale Eigenvalue and SVD Problems*, 2019. URL `https://CRAN.R-project.org/package=RSpectra`. R package version 0.16-0.

Garvesh Raskutti and Michael W Mahoney. A statistical perspective on randomized sketching for ordinary least-squares. *Journal of Machine Learning Research*, 17(1):7508–7538, 2016.

Matthew Richardson, Rakesh Agrawal, and Pedro Domingos. Trust management for the semantic web. In *International Semantic Web Conference*, pages 351–368. Springer, 2003.

Karl Rohe, Sourav Chatterhee, and Bin Yu. Spectral clustering and the high-dimensional stochastic block model. *The Annals of Statistics*, 39(4):1878–1915, 2011.

Karl Rohe, Tai Qin, and Bin Yu. Co-clustering for directed graphs: the stochastic co-blockmodel and spectral algorithm Di-Sim. *arXiv preprint arXiv:1204.2296*, 2012.

Karl Rohe, Tai Qin, and Bin Yu. Co-clustering directed graphs to discover asymmetries and directional communities. *Proceedings of the National Academy of Sciences*, 113(45): 12679–12684, 2016.

Ohad Shamir. A stochastic PCA and SVD algorithm with an exponential convergence rate. In *International Conference on Machine Learning*, pages 144–152. PMLR, 2015.

Ohad Shamir. Convergence of stochastic gradient descent for PCA. In *International Conference on Machine Learning*, pages 257–265. PMLR, 2016.

Liangjun Su, Wuyi Wang, and Yichong Zhang. Strong consistency of spectral clustering for stochastic block models. *IEEE Transactions on Information Theory*, 66(1):324–338, 2019.

Minh Tang, Joshua Cape, and Carey E Priebe. Asymptotically efficient estimators for stochastic blockmodels: The naive MLE, the rank-constrained MLE, and the spectral estimator. *Bernoulli*, 28(2):1049–1073, 2022.

Nicolas Tremblay and Andreas Loukas. Approximating spectral clustering via sampling: a review. *Sampling Techniques for Supervised or Unsupervised Tasks*, pages 129–183, 2020.

Nicolas Tremblay, Gilles Puy, Rémi Gribonval, and Pierre Vandergheynst. Compressive spectral clustering. In *International Conference on Machine Learning*, pages 1002–1011. PMLR, 2016.

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

Haiying Wang, Min Yang, and John Stufken. Information-based optimal subdata selection for big data linear regression. *Journal of the American Statistical Association*, 114(525): 393–405, 2019.

Shusen Wang, Alex Gittens, and Michael W Mahoney. Sketched ridge regression: Optimization perspective, statistical perspective, and model averaging. *Journal of Machine Learning Research*, 18(1):8039–8088, 2017.

Duncan J Watts and Steven H Strogatz. Collective dynamics of 'small-world'networks. *Nature*, 393(6684):440–442, 1998.

Rafi Witten and Emmanuel Candès. Randomized algorithms for low-rank matrix factorizations: sharp performance bounds. *Algorithmica*, 72(1):264–281, 2015.

David P Woodruff and Taisuke Yasuda. Improved algorithms for low rank approximation from sparsity. In *Proceedings of the 2022 Annual ACM-SIAM Symposium on Discrete Algorithms (SODA)*, pages 2358–2403. SIAM, 2022.

David P Woodruff et al. Sketching as a tool for numerical linear algebra. *Foundations and Trends® in Theoretical Computer Science*, 10(1–2):1–157, 2014.

Lingfei Wu and Andreas Stathopoulos. A preconditioned hybrid SVD method for accurately computing singular triplets of large matrices. *SIAM Journal on Scientific Computing*, 37 (5):S365–S388, 2015.

Lingfei Wu, Eloy Romero, and Andreas Stathopoulos. Primme_svds: A high-performance preconditioned SVD solver for accurate large-scale computations. *SIAM Journal on Scientific Computing*, 39(5):S248–S271, 2017.

Peng Xu, Bryan He, Christopher De Sa, Ioannis Mitliagkas, and Chris Re. Accelerated stochastic power iteration. In *International Conference on Artificial Intelligence and Statistics*, pages 58–67. PMLR, 2018.

Haishan Ye, Yujun Li, Cheng Chen, and Zhihua Zhang. Fast fisher discriminant analysis with randomized algorithms. *Pattern Recognition*, 72:82–92, 2017.

Hao Yin, Austin R Benson, Jure Leskovec, and David F Gleich. Local higher-order graph clustering. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 555–564. ACM, 2017.

Se-Young Yun and Alexandre Proutiere. Optimal cluster recovery in the labeled stochastic block model. In *Advances in Neural Information Processing Systems*, pages 965–973, 2016.

Anderson Y Zhang, Harrison H Zhou, et al. Minimax rates of community detection in stochastic block models. *The Annals of Statistics*, 44(5):2252–2280, 2016.

Hai Zhang, Xiao Guo, and Xiangyu Chang. Randomized spectral clustering in large-scale stochastic block models. *Journal of Computational and Graphical Statistics*, 31(3):887–906, 2022.

Zhixin Zhou and Arash A Amini. Analysis of spectral clustering algorithms for community detection: the general bipartite setting. *Journal of Machine Learning Research*, 20(47): 1–47, 2019.