Journal of Machine Learning Research 21 (2020) 1-45 Submitted 8/18; Revised 3/20; Published 7/20

Spectral Algorithms for Community Detection in Directed Networks

Zhe Wang
Department of Electrical and Computer Engineering
The Ohio State University
Columbus, OH 43202, USA
WANG.10982@OSU.EDU

Yingbin Liang
Department of Electrical and Computer Engineering
The Ohio State University
Columbus, OH 43202, USA
LIANG.889@OSU.EDU

Pengsheng Ji
Department of Statistics
University of Georgia
Athens, GA 30602, USA
PSJI@UGA.EDU

Editor: Francois Caron

Abstract

Community detection in large social networks is affected by degree heterogeneity of nodes. The D-SCORE algorithm for directed networks was introduced to reduce this effect by taking the element-wise ratios of the singular vectors of the adjacency matrix before clustering. Meaningful results were obtained for the statistician citation network, but rigorous analysis on its performance was missing. First, this paper establishes theoretical guarantee for this algorithm and its variants for the directed degree-corrected block model (Directed-DCBM). Second, this paper provides significant improvements for the original D-SCORE algorithms by attaching the nodes outside of the community cores using the information of the original network instead of the singular vectors.

Keywords: directed networks, community detection, clustering, degree-corrected block model, k-means, principle component analysis

1. Introduction

Social platforms have become increasingly important in our modern life since they provide fast and easy path to make new friends, maintain relationship and share moments. Due to the highly interactive activities in social platforms (e.g., Facebook, Wechat, Twitter, Line), people have generated a huge amount of data which is highly rich in social information. Various algorithms have been developed to extract useful information from these big social data sets, and community detection or clustering is one of the major tools to uncover the community information from big data.

The basic community detection problem has a simple form: given an n-node graph $\mathcal{N} = (\mathcal{V}, \mathcal{E})$ where $\mathcal{V} = \{1, 2 \cdots n\}$ is the set of nodes and $\mathcal{E}$ is the set of edges, the goal is to divide n nodes into K disjoint communities. It is believed that nodes within the communities share much more edges than those across communities. In order to formulate the problem more formally and facilitate the design and analysis of algorithms, some network models have been proposed. As one of the classic models, the *stochastic block model* (SBM) assumes that nodes in the same community have the same statistical edge pattern, i.e., they are stochastically equivalent as pointed out in Holland et al. (1983). While SBM is useful to capture the community character and easy to analyze, it implies that the distribution of degrees within the community is Poisson, in contrast to the empirical observation that in many natural networks, the degrees follow approximately a power-law distribution (Goldenberg et al., 2009). To overcome this shortcoming, *degree corrected* block model (DCBM) was proposed by Karrer and Newman (2011) to characterize the personality of each node with a heterogeneous parameter. DCBM is more realistic than SBM in terms of the degree distribution, but is usually impossible to fit due to the huge amount of heterogeneous parameters .

In reality, there exists a lot of **directed** networks such as citation networks, protein-protein interaction networks, the hyperlink network of websites. Such directed networks are more complex in that there are two types of information involved, namely starting links or receiving links, citing others or being cited, etc, which are not captured by SBM and DCBM. Thus, this paper explores a *directed* degree-corrected block model (Directed-DCBM) (see Section 2 for more details), which associates different degree parameters with two edge directions for individual nodes in order to model directed networks.

Many community detection algorithms have been proposed in recent years. Among these algorithms, we focus on spectral clustering algorithms for their efficiency and popularity. In this paper, we provide theoretical analysis of two spectral algorithms for the Directed-DCBM. The first one is D-SCORE algorithm proposed by Ji and Jin (2016) to analyze the statistician citation networks, but no rigorous analysis on its performance was provided. The second one is D-SCORE$_q$ which is a generalization of the row normalization technique. For $q = 2$, it becomes the row normalization technique which is commonly used in spectral clustering algorithms (Jin, 2015; Rohe et al., 2016) before clustering.

1.1. Contribution

In theory, this paper provides rigorous analysis of the D-SCORE algorithm for Directed-DCBM. The error bound is in the form of pure heterogeneous parameters, and shows clearly how heterogeneous parameters affect the clustering result and when consistency can be achieved. This paper also provides unified theoretical analysis of the D-SCORE$_q$ algorithm for the Directed-DCBM. Through the rigorous proof, we show that row normalization for the singular vectors using any ℓ_q-norm also reduces the effects of heterogeneous parameters and improves the algorithm performance.

The analytical techniques in this paper differ significantly from the previous work in the following aspects. First, the techniques in Jin (2015) for analyzing undirected networks can not be adapted to directed networks. Instead, we manage to use the Davis-Kahan theorem and take a more direct and general approach, and our techniques are potentially very useful for general network modeling such multi-layer networks and node-attributed networks. Sec-

ond, our way to deal with the asymmetric matrix is different from Rohe et al. (2016) who constructed a symmetric matrix by extending the adjacency $\mathbf{A}$ to $\begin{bmatrix} 0, & \mathbf{A} \\ \mathbf{A}^T, & 0 \end{bmatrix}$, whereas we use $\mathbf{A}\mathbf{A}^T$ and $\mathbf{A}^T\mathbf{A}$ that are naturally symmetric matrices and correspond to meaningful networks. Furthermore, our results are directly in the form of heterogeneous parameters which provide explicit insights about the impact of the heterogeneous parameters on the performance of the algorithm, unlike Rohe et al. (2016).

Furthermore, we identify possible issues with the original D-SCORE algorithms for large networks and improve these algorithms using the intersection-with-attachment technique. Specifically, we run the spectral algorithms on the graph core (intersection) and then attach the remaining nodes to the communities, instead of running the spectral algorithms directly on the entire graph as in Ji and Jin (2016). The rationale is presented carefully in text and then further demonstrated using real world data and simulations.

1.2. Related Work

We discuss the related work in view of different models as well as algorithms proposed for these models. Due to the extremely intensive studies on community detection, we focus on only algorithms which have theoretical consistency promise and are highly relevant to our study here. There are roughly three kinds of such algorithms that come with theoretical promise, namely the modularity method, spectral clustering, and optimization relaxation.

SBM was introduced by Holland et al. (1983), and various algorithms have been proposed for solving the community detection problem under SBM. In particular, the modularity method includes profile likelihood modularity (Bickel and Chen, 2009a), Erdos-Renyi modularity (Zhao et al., 2012), etc, and Zhao et al. (2012) provided the consistency proof for these two methods. Spectral clustering mainly has two kinds of methods: spectral clustering with normalized Laplacian matrix (Rohe et al., 2011), and spectral clustering with adjacency matrix (Sussman et al., 2012). In addition, regularization technique has been used to concentrate the eigenvector and improve the algorithm performance, where the details can be found in Joseph and Yu (2016). For the optimization method, objective functions were constructed, which were either inspired by the maximum likelihood estimation or by the insight that there should be more edges inside the community than those outside the community. Solutions to these optimization problems were obtained typically by relaxation, such as SDP relaxation (Amini and Levina, 2018) or convex relaxation (Demaine and Immorlica, 2003; Chen et al., 2012). It is of general interest to characterize sufficient and necessary conditions that guarantee the consistency of community detection. For example, Mossel et al. (2016, 2017) provided the if and only if conditions for consistent community detection for the case with $K = 2$ communities for the planted partition model, which is a special case of SBM. Moreover, Abbe and Sandon (2017) provided the characterization and new insights for consistent clustering for the case with $K \geqslant 3$.

DCBM was proposed by Karrer and Newman (2011) and various community detection algorithms were studied for DCBM. For modularity methods, Karrer and Newman (2011) provided an interpretation of Newman-Girvan modularity method (Newman and Girvan, 2004) under DCBM setting and further proposed a profile likelihood modularity method for DCBM. Zhao et al. (2012) provided the consistency proof for these two modularity methods. Furthermore, Newman (2016) showed that the Newman-Girvan modularity method under

DCBM is equivalent with the profile likelihood method in degree-corrected planted partition model with known block parameters. For spectral clustering methods, Lei and Rinaldo (2015) analyzed the performance of spectral clustering and Gulikers et al. (2017) proposed a spectral algorithm that does not need the knowledge of the number of communities. In addition, the SCORE algorithm (Jin, 2015) and the row-normalization technique (Qin and Rohe, 2013) were used to alleviate the effect of the heterogeneous parameters. For the optimization methods, Chen et al. (2018) proposed and analyzed a convexified modularity maximization approach under DCBM.

Some *directed* network models (where the edges have directions) have been proposed to model directed networks (Wang and Wong, 1987; Reichardt and White, 2007; Yang et al., 2010) and details can be found in Malliaros and Vazirgiannis (2013). We mainly focus on directed-DCBM. For such a model, Ji and Jin (2016) extended DCBM to directed-DCBM, and adapted the SCORE algorithm designed for DCBM to the D-SCORE algorithm which is applicable for directed-DCBM. Rohe et al. (2016) introduced the stochastic co-block model that combined the idea of DCBM and bi-clustering and developed the spectral co-clustering algorithm called DI-SIM for such a model.

Another important issue of community detection is the estimation of the number K of communities in the graph. Various techniques have been proposed to determine the number of communities in the graph. For example, Zhao et al. (2011) proposed to extract one community at a time, and then decided whether the reminder of the graph contains multiple communities by comparing the reminder of the graph with the Erdos-Renyi graph. Bickel and Sarkar (2016) proposed to recursively split the graph into two parts until each part contains only one community. Chen and Lei (2017) proposed a network cross-validation approach and Saldaña et al. (2017) proposed a likelihood-based method to determine the number of communities. More details and other methods can be found in these papers and the references therein.

2. Network Models

2.1. Directed-DCBM

In this section, we introduce the directed-DCBM. We consider a directed network $\mathcal{N}$, in which there are totally n nodes and we use $\mathcal{V} = \{1, 2, \cdots, n\}$ to denote the set of the indices of these nodes. We assume that the nodes in the network are connected by directional edges. We introduce an $n \times n$ adjacency matrix $\mathbf{A}$ of the network $\mathcal{N}$, and the entries of $\mathbf{A}$ take values either 1 or 0. For each entry, $\mathbf{A}(i, j) = 1$ if there is a directional edge from node i to node j, and $\mathbf{A}(i, j) = 0$ otherwise.

We assume that the nodes in the network are divided into K disjoint communities, and we use $\mathcal{V}^{(k)}$ for $k = 1, \ldots, K$ to represent the set that contains the indices of the nodes in community k. Thus, $\mathcal{V} = \mathcal{V}^{(1)} \cup \mathcal{V}^{(2)} \cdots \cup \mathcal{V}^{(K)}$. We let n_k denote the total number of nodes in community k, i.e., $n_k = |\mathcal{V}^{(k)}|$ for $1 \leqslant k \leqslant K$. Thus, $\sum_{k=1}^{K} n_k = n$.

We assume that the connectivity behavior of each node is captured by both the common connectivity parameters shared among all nodes in the same community as well as the connectivity parameters of each node. We use a $K \times K$ matrix $\mathbf{B}$ to model the community connectivity behavior. Here, each entry $\mathbf{B}(k, l)$ represents the chance that there exists a directional edge from a node in community k to a node in community l, for $k, l = 1, \ldots, K$.

For each node i, we assign two parameters denoted by $\boldsymbol{\theta}(i)$ and $\boldsymbol{\delta}(i)$, where $\boldsymbol{\theta}(i)$ captures how likely node i points edges to other nodes, and $\boldsymbol{\delta}(i)$ captures how likely node i receives edges from other nodes. Hence, $\boldsymbol{\theta}(i)$ and $\boldsymbol{\delta}(i)$ for $i = 1, \ldots, n$ model connectivity properties for individual nodes, and are referred to as *heterogeneous parameters*.

We model the entries of the adjacency matrix $\mathbf{A}$ as independent Bernoulli random variables, with each entry $\mathbf{A}(i, j) = 1$ having the following probability

$$P(\mathbf{A}(i, j) = 1) = \boldsymbol{\theta}(i)\mathbf{B}(c_i, c_j)\boldsymbol{\delta}(j), \quad \text{for } i, j = 1, \ldots, n, \tag{2.1}$$

where c_i for $i = 1, \ldots, n$ denotes the index of the community that node i belongs to. Note that $\mathbf{A}(i, j) = 1$ represents that there exists a directional edge from node i to node j. As can be observed from eq. (2.1), the probability that there exists such an edge depends on both the community connectivity parameters $\mathbf{B}(c_i, c_j)$ and heterogeneous parameters $\boldsymbol{\theta}(i)$ and $\boldsymbol{\delta}(j)$ of the individual nodes i and j.

Since the network contains *directional* edges, the directed-DCBM consists of the following three aspects of asymmetry, which distinguishes the directed-DCBM significantly from the typical DCBM. (i) The matrix $\mathbf{B}$ can be asymmetric, i.e., $\mathbf{B}(k, l) \neq \mathbf{B}(l, k)$, which implies that the connectivity parameter from community k to community l can be different from that from community l to community k. (ii) The two heterogeneous parameters for each node can be unequal, i.e., $\boldsymbol{\theta}(i) \neq \boldsymbol{\delta}(i)$, which implies that the chance for one node to point edges to other nodes is generally different from that for one node to receive edges from other nodes. (iii) The random adjacency matrix $\mathbf{A}$ is also asymmetric, where $\mathbf{A}(i, j)$ represents the existence of an edge from node i to node j, while $\mathbf{A}(j, i)$ represents the existence of an edge from node j to node i. And they also take different Bernoulli distribution parameters. As can be seen in eq. (2.1), the asymmetries of $\mathbf{B}(c_i, c_j)$ and that of $\boldsymbol{\theta}(i)$ and $\boldsymbol{\delta}(i)$ yield asymmetric parameters for $P(\mathbf{A}(i, j) = 1)$.

Let $\boldsymbol{\Omega} = \mathrm{E}[\mathbf{A}]$, where $E[\mathbf{A}]$ is the expectation of the $n \times n$ matrix $\mathbf{A}$. Further let

$$\mathbf{W} \equiv \mathbf{A} - E[\mathbf{A}] = \mathbf{A} - \boldsymbol{\Omega}. \tag{2.2}$$

Note that the entries in matrix $\mathbf{W}$ are independently centered Bernoulli random variables.

2.2. Notations

We take the following general notations in this paper. For a vector $\mathbf{v}$ and fixed $q > 0$, $\|\mathbf{v}\|_q$ denotes its ℓ_q-norm. We drop the subscript if $q = 2$. For a matrix $\mathbf{M}$, $\mathbf{M}^T$ denotes the transpose of the matrix $\mathbf{M}$, $\|\mathbf{M}\|$ denotes the spectral norm, and $\|\mathbf{M}\|_F$ denotes the Frobenius norm. We let $\|\mathbf{M}\|_{\min}$ denote the smallest singular value of the matrix $\mathbf{M}$. Let $\sigma_i(\mathbf{M})$ denote the i-th largest singular value of matrix $\mathbf{M}$, and $\lambda_i(\mathbf{M})$ denote the i-th largest eigenvalue of the matrix $\mathbf{M}$ ordered by the magnitude. In addition, we use $\mathbf{M}_{\bar{i}}$ to denote the i-th row of the matrix M (a bar over the subscript i) and $\mathbf{M}(i, j)$ to denote the (i, j)th entry of matrix $\mathbf{M}$. For integer $i, j > 0$, let $\mathbf{M}_{i\sim j}$ denote the matrix that is formed by extracting the i-th to j-th columns of the matrix $\mathbf{M}$.

For two positive sequences $\{a_n\}_{n=1}^{\infty}$ and $\{b_n\}_{n=1}^{\infty}$, we say $a_n \asymp b_n$ if there exists a constant C such that $b_n/C \leqslant a_n \leqslant Cb_n$ for sufficiently large n, i.e., a_n and b_n are in the same order. For a set $\mathcal{V}, |\mathcal{V}|$ denotes its cardinality.

2.3. Assumptions

In this subsection, we describe the assumptions about the matrix $\mathbf{B}$ and the heterogeneous parameters $\boldsymbol{\theta}(i)$ and $\boldsymbol{\delta}(i)$ for $i = 1, \ldots, n$, which we make throughout this paper. For brevity we drop them in our propositions and lemmas.

Assumption 1 *The matrix* $\mathbf{B}$ *satisfies*

$$0 \leqslant \mathbf{B}(i,j) \leqslant 1 \quad \textit{for} \quad 1 \leqslant i,j \leqslant K, \tag{2.3}$$

$$\mathbf{B}\mathbf{B}^T \textit{ and } \mathbf{B}^T\mathbf{B} \textit{ are non-singular, non-negative and irreducible.} \tag{2.4}$$

As we observe later, the non-singularity, non-negativity and irreducibility guarantee that the first leading left and right singular vectors (corresponding to the largest singular value) of B are nonzero so that they can ensure the denominator is nonzero in the D-SCORE and D-SCORE$_q$ algorithms.

To describe our assumptions for the heterogeneous parameters, we first define some simplified notations. We collect $\boldsymbol{\theta}(i)$ for $i = 1, \ldots, n$ into a vector denoted by $\boldsymbol{\theta}$, and collect $\boldsymbol{\delta}(i)$ for $i = 1, \ldots, n$ into a vector denoted by $\boldsymbol{\delta}$. We define n-dimensional vectors $\boldsymbol{\theta}^{(k)}$ and $\boldsymbol{\delta}^{(k)}$ for $1 \leqslant k \leqslant K$ as

$$\boldsymbol{\theta}^{(k)}(i) = \begin{cases} \boldsymbol{\theta}(i) & \text{if} \quad c_i = k \\ 0 & \text{if} \quad c_i \neq k \end{cases} \quad \text{and} \quad \boldsymbol{\delta}^{(k)}(i) = \begin{cases} \boldsymbol{\delta}(i) & \text{if} \quad c_i = k \\ 0 & \text{if} \quad c_i \neq k \end{cases},$$

where c_i denotes the index of the community that node i belongs to. We further define $\boldsymbol{\theta}_{\min} \equiv \min_{1\leqslant i\leqslant n} \boldsymbol{\theta}(i)$, $\boldsymbol{\theta}_{\max} \equiv \max_{1\leqslant i\leqslant n} \boldsymbol{\theta}(i)$, $\boldsymbol{\delta}_{\min} \equiv \min_{1\leqslant i\leqslant n} \boldsymbol{\delta}(i)$, and $\boldsymbol{\delta}_{\max} \equiv \max_{1\leqslant i\leqslant n} \boldsymbol{\delta}(i)$. We also define the following quantity

$$Z \equiv \max\left(\boldsymbol{\theta}_{\max}, \boldsymbol{\delta}_{\max}\right) \max\left(\|\boldsymbol{\theta}\|_1, \|\boldsymbol{\delta}\|_1\right), \tag{2.5}$$

which appears many times in our analysis.

In this paper, we assume that the heterogeneous parameter vectors $\boldsymbol{\theta}$ and $\boldsymbol{\delta}$ can scale with the network size n, and hence the asymptotic properties in the following assumptions are all with respect to n. For notational simplicity, we do not express these parameters explicitly as a function of n.

Assumption 2 *The heterogeneity parameters* $\boldsymbol{\theta}$ *and* $\boldsymbol{\delta}$ *satisfy*

$$0 < \boldsymbol{\theta}_{\min} \leqslant \boldsymbol{\theta}_{\max} \leqslant 1, \quad 0 < \boldsymbol{\delta}_{\min} \leqslant \boldsymbol{\delta}_{\max} \leqslant 1, \tag{2.6}$$

$$\|\boldsymbol{\theta}^{(k)}\| \asymp \|\boldsymbol{\theta}^{(l)}\|, \quad \|\boldsymbol{\delta}^{(k)}\| \asymp \|\boldsymbol{\delta}^{(l)}\| \quad \textit{for } 1 \leqslant k,l \leqslant K, \tag{2.7}$$

$$\lim_{n\to\infty} \frac{\log(n) Z}{\boldsymbol{\theta}_{\min}\boldsymbol{\delta}_{\min}\|\boldsymbol{\theta}\|_1\|\boldsymbol{\delta}\|_1} = 0. \tag{2.8}$$

To further explain these assumptions, eq. (2.7) requires that the ℓ_2-norm of the heterogeneous parameter vectors, i.e., $\|\boldsymbol{\theta}^{(k)}\|$, are in the same order across all communities. Intuitively, $\|\boldsymbol{\theta}^{(k)}\|$ captures the number of edges that community k points to other communities in total. Then eq. (2.7) implies that the total number of edges that each community points out are in the same order. To explain eq. (2.8), $\|\boldsymbol{\theta}\|_1$ and $\|\boldsymbol{\delta}\|_1$ capture the degrees (i.e.,

the numbers of edges) that each node respectively receives and points out in total. Then eq. (2.8) essentially requires that the total degree scales faster than $\log n$.

We next present a few properties that follow directly from Assumption 2. Since $\|\boldsymbol{\theta}\|^2 = \sum_{k=1}^K \|\boldsymbol{\theta}^{(k)}\|^2$ and $\|\boldsymbol{\delta}\|^2 = \sum_{k=1}^K \|\boldsymbol{\delta}^{(k)}\|^2$, eq. (2.7) implies

$$\|\boldsymbol{\theta}^{(i)}\| \asymp \|\boldsymbol{\theta}\| \text{ and } \|\boldsymbol{\delta}^{(i)}\| \asymp \|\boldsymbol{\delta}\| \quad \text{for } 1 \leqslant i, j \leqslant K. \tag{2.9}$$

To interpret eq. (2.9), for all $1 \leqslant i \leqslant K$, $\|\boldsymbol{\theta}^{(i)}\| \asymp \|\boldsymbol{\theta}\|$ implies that $\|\boldsymbol{\theta}^{(i)}\|$ has the same order as the total degree norm $\|\boldsymbol{\theta}\|$. The similar interpretation holds for $\|\boldsymbol{\delta}^{(i)}\| \asymp \|\boldsymbol{\delta}\|$.

Furthermore, since $\boldsymbol{\theta}_{\min}\|\boldsymbol{\theta}\|_1 \leqslant \|\boldsymbol{\theta}\|^2$ and $\boldsymbol{\delta}_{\min}\|\boldsymbol{\delta}\|_1 \leqslant \|\boldsymbol{\delta}\|^2$, by eq. (2.8) we have

$$\lim_{n\to\infty} \frac{\log(n)Z}{\|\boldsymbol{\theta}\|^2\|\boldsymbol{\delta}\|^2} \leqslant \lim_{n\to\infty} \frac{\log(n)Z}{\boldsymbol{\theta}_{\min}\boldsymbol{\delta}_{\min}\|\boldsymbol{\theta}\|_1\|\boldsymbol{\delta}\|_1} = 0. \tag{2.10}$$

Since $\frac{\log(n)Z}{\|\boldsymbol{\theta}\|^2\|\boldsymbol{\delta}\|^2} \geqslant 0$ holds for all $n \geqslant 0$, we conclude that

$$\lim_{n\to\infty} \frac{\log(n)Z}{\|\boldsymbol{\theta}\|^2\|\boldsymbol{\delta}\|^2} = 0. \tag{2.11}$$

Since the definition of Z suggests that $Z \geqslant \boldsymbol{\theta}_{\min}\|\boldsymbol{\theta}\|_1$ and $Z \geqslant \boldsymbol{\delta}_{\min}\|\boldsymbol{\delta}\|_1$, combining with eq. (2.8) we have

$$\lim_{n\to\infty} \frac{\log(n)}{Z} = \lim_{n\to\infty} \frac{\log(n)Z}{Z^2} \leqslant \lim_{n\to\infty} \frac{\log(n)Z}{\boldsymbol{\theta}_{\min}\boldsymbol{\delta}_{\min}\|\boldsymbol{\theta}\|_1\|\boldsymbol{\delta}\|_1} = 0. \tag{2.12}$$

Since $\lim_{n\to\infty} \frac{\log(n)}{Z} \geqslant 0$ holds for all $n > 0$, we conclude that

$$\lim_{n\to\infty} \frac{\log(n)}{Z} = 0. \tag{2.13}$$

3. Algorithms

In this section, we describe the two community detection algorithms D-SCORE and D-SCORE$_q$ that we analyze in this paper. We also provide an improved algorithm, i.e., Algorithm 3, which is more suitable to deal with real data.

D-SCORE (see Algorithm 1) was proposed in Ji and Jin (2016) for directed-DCBM, as an adapted version of SCORE proposed in Jin (2015) for community detection for DCBM with undirected edges. SCORE is a type of spectral clustering algorithm and can deal with the model with nodes having heterogeneous parameters to capture their individual connectivity behavior. The central idea of SCORE is to first collect the first K leading eigenvectors of the adjacency matrix into a new matrix, and then divide each row of such a matrix by its first entry. The effect of heterogeneous parameters can be reduced dramatically, and hence the standard clustering approaches can be applied. SCORE handles network models with undirected edges, but cannot handle networks with directed edges.

D-SCORE adapts SCORE to network models with *directed edges*, where the adjacency matrix is usually *asymmetric*. Thus D-SCORE uses the left and right singular vectors for spectral clustering as opposed to SCORE that uses eigenvectors due to the *symmetry* of

Algorithm 1: D-SCORE($\hat{\mathbf{U}}, \hat{\mathbf{V}}, K$)

Input : The number K of communities, the $n \times K$ (unit-norm) leading left and right singular vector matrices of the adjacency matrix $\mathbf{A}$ denoted by $\hat{\mathbf{U}} = [\hat{\mathbf{U}}_1, \ldots, \hat{\mathbf{U}}_K]$ and $\hat{\mathbf{V}} = [\hat{\mathbf{V}}_1, \ldots, \hat{\mathbf{V}}_K]$.

1 Fix a threshold $T_n = \log n$ (used to avoid zero denominator), define the $n \times (K-1)$ ratio matrices $\mathbf{R}_{\hat{\mathbf{U}}}$ and $\mathbf{R}_{\hat{\mathbf{V}}}$, such that for $1 \leqslant i \leqslant n, 1 \leqslant k \leqslant (K-1)$,

$$\mathbf{R}_{\hat{\mathbf{U}}}(i,k) = \begin{cases} T_n & \text{if} \quad \frac{\hat{\mathbf{U}}_{k+1}(i)}{\hat{\mathbf{U}}_1(i)} > T_n \\ \frac{\hat{\mathbf{U}}_{k+1}(i)}{\hat{\mathbf{U}}_1(i)} & \text{if} \quad \left|\frac{\hat{\mathbf{U}}_{k+1}(i)}{\hat{\mathbf{U}}_1(i)}\right| \leqslant T_n \\ -T_n & \text{if} \quad \frac{\hat{\mathbf{U}}_{k+1}(i)}{\hat{\mathbf{U}}_1(i)} < -T_n \end{cases}, \mathbf{R}_{\hat{\mathbf{V}}}(i,k) = \begin{cases} T_n & \text{if} \quad \frac{\hat{\mathbf{V}}_{k+1}(i)}{\hat{\mathbf{V}}_1(i)} > T_n \\ \frac{\hat{\mathbf{V}}_{k+1}(i)}{\hat{\mathbf{V}}_1(i)} & \text{if} \quad \left|\frac{\hat{\mathbf{V}}_{k+1}(i)}{\hat{\mathbf{V}}_1(i)}\right| \leqslant T_n \\ -T_n & \text{if} \quad \frac{\hat{\mathbf{V}}_{k+1}(i)}{\hat{\mathbf{V}}_1(i)} < -T_n \end{cases} \tag{3.1}$$

2 Put $\mathbf{R}_{\hat{\mathbf{U}}}$ and $\mathbf{R}_{\hat{\mathbf{V}}}$ together to form an $n \times (2K-2)$ ratio matrix $\hat{\mathbf{R}}$, i.e., $\hat{\mathbf{R}} = [\mathbf{R}_{\hat{\mathbf{U}}}, \mathbf{R}_{\hat{\mathbf{V}}}]$. Then run k-means on $\hat{\mathbf{R}}$, i.e., find the solution to the following optimization problem:

$$\mathbf{M}^* = \underset{\mathbf{M} \in \mathbf{M}_{n,2k-2,K}}{\operatorname{argmin}} \left\| \mathbf{M} - \hat{\mathbf{R}} \right\|_F^2,$$

where $\mathcal{M}_{n,2K-2,K}$ denotes the set of $n \times (2K-2)$ matrices with only K different rows.

3 Use M^* to assign membership.

Output: The community labels of the nodes.

Algorithm 2: D-SCORE$_q(\hat{\mathbf{U}}, \hat{\mathbf{V}}, K)$

Input : The number K of communities, the $n \times K$ (unit-norm) leading left and right singular vector matrices of the adjacency matrix $\mathbf{A}$ denoted by $\hat{\mathbf{U}} = [\hat{\mathbf{U}}_1, \ldots, \hat{\mathbf{U}}_K]$ and $\hat{\mathbf{V}} = [\hat{\mathbf{V}}_1, \ldots, \hat{\mathbf{V}}_K]$.

1 Fix a threshold $T_n = \log n$ (used to avoid zero denominator), define two $n \times K$ ratio matrices $\mathbf{R}_{\hat{\mathbf{U}}}$ and $\mathbf{R}_{\hat{\mathbf{V}}}$, such that for $1 \leqslant i \leqslant n, 1 \leqslant k \leqslant K$,

$$\mathbf{R}_{\hat{\mathbf{U}}}(i,k) = \begin{cases} T_n & \text{if } \frac{\hat{\mathbf{U}}_k(i)}{\|\hat{\mathbf{U}}_{\bar{i}}\|_q} > T_n \\ \frac{\hat{\mathbf{U}}_k(i)}{\|\hat{\mathbf{U}}_{\bar{i}}\|_q} & \text{if } \left|\frac{\hat{\mathbf{U}}_k(i)}{\|\hat{\mathbf{U}}_{\bar{i}}\|_q}\right| \leqslant T_n \\ -T_n & \text{if } \frac{\hat{\mathbf{U}}_k(i)}{\|\hat{\mathbf{U}}_{\bar{i}}\|_q} < -T_n \end{cases}, \mathbf{R}_{\hat{\mathbf{V}}}(i,k) = \begin{cases} T_n & \text{if } \frac{\hat{\mathbf{V}}_k(i)}{\|\hat{\mathbf{V}}_{\bar{i}}\|_q} > T_n \\ \frac{\hat{\mathbf{V}}_k(i)}{\|\hat{\mathbf{V}}_{\bar{i}}\|_q} & \text{if } \left|\frac{\hat{\mathbf{V}}_k(i)}{\|\hat{\mathbf{V}}_{\bar{i}}\|_q}\right| \leqslant T_n \\ -T_n & \text{if } \frac{\hat{\mathbf{V}}_k(i)}{\|\hat{\mathbf{V}}_{\bar{i}}\|_q} < -T_n \end{cases} \tag{3.2}$$

2 Put $\mathbf{R}_{\hat{\mathbf{U}}}$ and $\mathbf{R}_{\hat{\mathbf{V}}}$ together to form an $n \times 2K$ ratio matrix $\hat{\mathbf{R}}$, i.e., $\hat{\mathbf{R}} = [\mathbf{R}_{\hat{\mathbf{U}}}, \mathbf{R}_{\hat{\mathbf{V}}}]$. Then run k-means on $\hat{\mathbf{R}}$, i.e., find the solution to the following optimization problem:

$$\mathbf{M}^* = \underset{\mathbf{M} \in \mathcal{M}_{n,2K,K}}{\operatorname{argmin}} \left\| \mathbf{M} - \hat{\mathbf{R}} \right\|_F^2,$$

where $\mathcal{M}_{n,2k,K}$ denotes the set of $n \times 2K$ matrices with K different rows.

3 Use $\mathbf{M}^*$ to assign membership.

Output: The community labels of the nodes.

the adjacency matrix. More specifically, D-SCORE first collects the first K leading left and right singular vectors into two matrices, and then divides each row of these two matrices by its first entry. In this way, the effect caused by the heterogeneous parameters can also be eliminated. D-SCORE then combines these two matrices together and applies standard approaches for clustering. D-SCORE was shown to have good empirical performance when it was applied to analyze data of a co-authorship and a citation network for statisticians in Ji and Jin (2016). However, the performance guarantee for D-SCORE was not established. In Section 4, we provide such performance analysis.

We then propose an alternative algorithm, i.e., D-SCORE$_q$ (see Algorithm 2), for directed-DCBM, which is an adapted version of the SCORE$_q$ algorithm proposed in Jin (2015) for community detection for DCBM with undirected edges. SCORE$_q$ differs from SCORE in that SCORE$_q$ divides each row of the matrix by the ℓ_q norm rather than the first entry of the corresponding row in SCORE to eliminate the effect caused by the heterogeneous parameters. Note that both SCORE$_q$ and SCORE are designed for networks with *undirected edges*. D-SCORE$_q$ differs from D-SCORE in the same way as SCORE$_q$ differs from SCORE, i.e., D-SCORE$_q$ divides each row of the matrix of singular vectors by the ℓ_q norm of the corresponding row. Both D-SCORE and D-SCORE$_q$ are designed for networks with *directed edges*. In Section 4, we provide the performance guarantee for D-SCORE$_q$ for any integer $q > 0$.

Algorithm 3: Improved D-SCORE$_q(K, A)$ using intersection-with-attachment

Input : The number K of communities and the adjacency matrix A.

1 Compute the K largest (unit-norm) leading left and right singular vectors of the adjacency matrix A to form two $n \times K$ singular vector matrices denoted by $U = [\, U_1, \ldots, U_K]$ and $V = [V_1, \ldots, V_K]$. Denote the set of the nodes by S.

2 Extract the largest connected components of matrices AA^T and A^TA, and denote S_l and S_r respectively as the sets of nodes in the two connected components.

3 Select the rows of U and V corresponding to $S_l \cap S_r$ to form two $|S_l \cap S_r| \times K$ matrices $\hat{U} = [\hat{U}_1, \hat{U}_2, \ldots, \hat{U}_K]$ and $\hat{V} = [\hat{V}_1, \hat{V}_2, \ldots, \hat{V}_K]$.

4 Run D-SCORE$(\hat{U}, \hat{V}, K)$ or D-SCOREq$(\hat{U}, \hat{V}, K)$ to assign the community labels to the nodes in $S_l \cap S_r$.

5 Attach these nodes outside $S_l \cap S_r$, i.e., $i \in S \backslash (S_l \cap S_r)$, by the following optimization step.

$$c_i = \max_{c \in 1, \cdots, K} \sum_{j=1}^{n} \big(A_{ij} + A_{ji}\big) 1_{\{c_j\}}(c), \tag{3.3}$$

where $1_{\{c_j\}}(\cdot)$ equals one if $c = c_j$ and equals zero otherwise.

Output: Community labels of the nodes.

We further propose an algorithm based on the intersection graph with attachment (see Algorithm 3) to improve the performance of D-SCORE and D-SCORE$_q$. In order for D-SCORE and D-SCORE$_q$ to perform well, it requires that the weighted graphs defined by $\mathbf{A}^T\mathbf{A}$ and $\mathbf{A}\mathbf{A}^T$ are both connected. This connectivity requirement on $\mathbf{A}^T\mathbf{A}$ and $\mathbf{A}\mathbf{A}^T$ can

be violated in real data with large networks. When this happens to either matrix, its leading eigenvector is 0 in theory for all nodes outside of the giant component, but the extremely small numbers (computational errors for 0) appear as the denominators for D-SCORE$_q$ and D-SCORE, causing misclustering errors on these nodes.

To fix this issue, Algorithm 3 is introduced to first extract the intersection of the sets of the nodes respectively corresponding to the largest connected components of $\mathbf{A}^T\mathbf{A}$ and $\mathbf{A}\mathbf{A}^T$ (see step 2 in Algorithm 3). Such an intersection set can be interpreted as the core of the graph. And then we apply D-SCORE$_q$ or D-SCORE over this intersection set (see steps 3 and 4 in Algorithm 3) to assign community labels to nodes in the intersection set. We then assign each node outside the intersection set to the community, to which the node has the most edge connections (including received and pointed out edges). This step, i.e., step 5 in Algorithm 3, is referred to as the attachment step. As demonstrated by our experiments in Section 5 and Section 5.3, the experiments show that the *intersection-with-attachment* technique can greatly improve performance of all the original D-SCORE algorithms.

The intuition behind Algorithm 3 is that nodes outside the intersection set is kind of noise nodes with less information since they do not have a strong connection with the graph, we extract the core of the graph by ignoring the noise nodes, and then attach them with the core graph. This observation can be seen clearly in figs. 2a, 2b, 3a and 3b, nodes in the intersection (the core) have a clear community structure while nodes outside the intersection is kind of mingling with each other. Ignoring noise nodes in the first step gives a clear picture for the underlying community structure, and thus improves the performance of proposed algorithms.

Furthermore, for the robustness consideration, we can replace the k-means step in D-SCORE and D-SCORE$_q$ with k-medoids (Park and Jun, 2009) or other approaches for clustering, which are more robust to outliers.

4. Main Results

In this section, we establish the performance guarantee for D-SCORE and D-SCORE$_q$ in Section 4.1 and Section 4.2, respectively.

4.1. Performance Guarantee for D-SCORE

As a road map to prove the performance guarantee for D-SCORE, we first analyze the property of the matrix that consists of singular vectors of the expected adjacency matrix $\mathbf{\Omega}$ in Proposition 1, and then bound the distance between this matrix and its random version that consists of the singular vectors of the random adjacency matrix $\mathbf{A}$ in Proposition 2. Furthermore, we prove that the ratio matrix generated by the expected adjacency matrix $\mathbf{\Omega}$ has a desired property for spectral clustering in Proposition 3, and then bound the distance between such a ratio matrix and its random version generated by the random adjacency matrix $\mathbf{A}$ in Proposition 4. After that we bound the distance between M^* and the ratio matrix generated by the singular vectors of the expected adjacency matrix $\mathbf{\Omega}$ in Proposition 5. Combining all these five propositions together, we establish our main result in Theorem 1. All the proofs are provided in Appendix A.

First, we analyze the singular vector matrix of the expected matrix $\mathbf{\Omega}$ of the random adjacency matrix $\mathbf{A}$, which captures the key information for clustering. We also anticipate

that the property of $\mathbf{\Omega}$ should well approximate that of $\mathbf{A}$, which we study next. We first define $\mathbf{S} \equiv \mathbf{\Psi_\theta B \Psi_\delta^T}$, where the matrix $\mathbf{B}$ captures the connectivity parameters among communities (see eq. (2.1)), and $\mathbf{\Psi_\theta}$, $\mathbf{\Psi_\delta}$ are the $K \times K$ diagonal matrices such that for $1 \leqslant i \leqslant K$,

$$\Psi_{\boldsymbol{\theta}}(i,i) = \frac{\|\boldsymbol{\theta}^{(i)}\|}{\|\boldsymbol{\theta}\|} \quad \text{and} \quad \Psi_{\boldsymbol{\delta}}(i,i) = \frac{\|\boldsymbol{\delta}^{(i)}\|}{\|\boldsymbol{\theta}\|}. \tag{4.1}$$

Hence, $\mathbf{\Psi_\theta}$, $\mathbf{\Psi_\delta}$ capture the total heterogeneity of each community.

The following proposition provides the singular vector decomposition of $\mathbf{\Omega}$.

Proposition 1 *Let $\mathbf{\Omega} = \mathbf{U \Lambda V}^T$ denote the compact singular value decomposition of $\mathbf{\Omega}$. Then, the singular values of $\mathbf{\Omega}$ are given by*

$$\sigma_i(\mathbf{\Omega}) = \begin{cases} \|\boldsymbol{\theta}\|\|\boldsymbol{\delta}\|\sigma_i(\mathbf{S}) & \text{if } 1 \leqslant i \leqslant K, \\ 0 & \text{if } i > K, \end{cases} \tag{4.2}$$

where $\mathbf{S} \equiv \mathbf{\Psi_\theta B \Psi_\delta^T}$. Let $\mathbf{S} = \mathbf{Y \Lambda_s H}^T$ denote the singular value decomposition of $\mathbf{S}$. The singular vectors of $\mathbf{\Omega}$ in row's form are given by

$$\mathbf{V}_{\bar{i}} = \frac{\boldsymbol{\delta}(i)}{\left\|\boldsymbol{\delta}^{(c_i)}\right\|}\mathbf{H}_{\bar{c}_i} \quad \text{and} \quad \mathbf{U}_{\bar{i}} = \frac{\boldsymbol{\theta}(i)}{\left\|\boldsymbol{\theta}^{(c_i)}\right\|}\mathbf{Y}_{\bar{c}_i} \quad \text{for } 1 \leqslant i \leqslant n, \tag{4.3}$$

and in column's form are given by

$$\mathbf{V}_i = \sum_{k=1}^{K} \frac{\boldsymbol{\delta}^{(k)}}{\|\boldsymbol{\delta}^{(k)}\|}\mathbf{H}_i(k) \quad \text{for } 1 \leqslant i \leqslant K, \tag{4.4}$$

$$\mathbf{U}_i = \sum_{k=1}^{K} \frac{\boldsymbol{\theta}^{(k)}}{\|\boldsymbol{\theta}^{(k)}\|}\mathbf{Y}_i(k) \quad \text{for } 1 \leqslant i \leqslant K. \tag{4.5}$$

Furthermore,

$$\|\mathbf{V}_{\bar{i}}\| \asymp \frac{\boldsymbol{\delta}(i)}{\|\boldsymbol{\delta}\|} \quad \text{and} \quad \|\mathbf{U}_{\bar{i}}\| \asymp \frac{\boldsymbol{\theta}(i)}{\|\boldsymbol{\theta}\|}, \quad \text{for } 1 \leqslant i \leqslant n. \tag{4.6}$$

Proof The proof can be found in Appendix A.1. ■

We note that eq. (4.2) implies that $\mathbf{\Omega}$ has only K non-zero singular values due to the fact that there are in total K disjoint communities. Thus, the *compact* singular value decomposition of $\mathbf{\Omega}$ is written in the form of an $n \times K$ left singular matrix $\mathbf{U}$, an $n \times K$ right singular matrix $\mathbf{V}$, and a $K \times K$ diagonal matrix $\mathbf{\Lambda}$.

To further explain the result of Proposition 1, consider nodes i, j and suppose they are in the same community, i.e., $c_i = c_j = k$. Then by eq. (4.3), the corresponding rows of nodes i and j in the matrix $\mathbf{V}$ are given by $\mathbf{V}_{\bar{i}} = \frac{\boldsymbol{\delta}(i)}{\left\|\boldsymbol{\delta}^{(k)}\right\|}\mathbf{H}_{\bar{k}}$ and $\mathbf{V}_{\bar{j}} = \frac{\boldsymbol{\delta}(j)}{\left\|\boldsymbol{\delta}^{(k)}\right\|}\mathbf{H}_{\bar{k}}$, respectively. These two row vectors differ only by the individual node parameters $\boldsymbol{\delta}(i)$ and $\boldsymbol{\delta}(j)$. In fact, the step (3.1) in the Algorithm 1 exactly eliminates these heterogeneous parameters to make

the corresponding vectors become the same if nodes are in the same community. On the other hand, if nodes i, j are in the different communities, i.e., $c_i \neq c_j$, their corresponding row vectors $\mathbf{V}_{\bar{i}}$ and $\mathbf{V}_{\bar{j}}$ are very different. The same argument is applicable to the row vectors in the left singular vector matrix $\mathbf{U}$. This observation intuitively justifies why the singular vector matrices can be used for recovering the community labels of the nodes.

Next, we bound the distance between the singular vectors of the random adjacency matrix $\mathbf{A}$ and those of $\mathbf{\Omega}$. The central idea of the proof is the proper application of Davis-Kahan inequality.

Proposition 2 *Let the first K leading left and right singular vectors of $\mathbf{A}$ be denoted by $\hat{\mathbf{V}}_1 \cdots \hat{\mathbf{V}}_K$ and $\hat{\mathbf{U}}_1 \cdots \hat{\mathbf{U}}_K$, and the first K leading left and right singular vectors of $\mathbf{\Omega}$ be denoted by $\mathbf{V}_1 \cdots \mathbf{V}_K$ and $\mathbf{U}_1 \cdots \mathbf{U}_K$. Then there exist two constants C_V and C_U with absolute value 1 and two orthogonal $(K-1) \times (K-1)$ matrices $\mathbf{O_V}$ and $\mathbf{O_U}$, such that for n large enough, with probability at least $1 - O(n^{-4})$, the following bounds hold*

$$\|\hat{\mathbf{V}}_1 - \mathbf{V}_1 C_V\|_F \leqslant C\frac{\sqrt{\log(n)Z}}{\|\boldsymbol{\theta}\|\|\boldsymbol{\delta}\|}, \qquad \|\hat{\mathbf{V}}_{2\sim K} - \mathbf{V}_{2\sim K}\mathbf{O_V}\|_F \leqslant C\frac{\sqrt{\log(n)Z}}{\|\boldsymbol{\theta}\|\|\boldsymbol{\delta}\|},$$
$$\|\hat{\mathbf{U}}_1 - \mathbf{U}_1 C_U\|_F \leqslant C\frac{\sqrt{\log(n)Z}}{\|\boldsymbol{\theta}\|\|\boldsymbol{\delta}\|}, \qquad \|\hat{\mathbf{U}}_{2\sim K} - \mathbf{U}_{2\sim K}\mathbf{O_U}\|_F \leqslant C\frac{\sqrt{\log(n)Z}}{\|\boldsymbol{\theta}\|\|\boldsymbol{\delta}\|},$$

where Z is defined in eq. (2.5).

Proof The proof can be found in Appendix A.2. ■

With Proposition 2, we are ready to explain further the idea of eliminating the effect caused by heterogeneous parameters from the singular vectors in Algorithm 1. *The central idea is to divide each row of the singular vector matrix by its first entry.* To this end, for $i = 1, \cdots n$, we define ratio matrices $\mathbf{R_V}$ and $\mathbf{R_U}$ as

$$(\mathbf{R_V})_{\bar{i}} = \frac{(\mathbf{V}_{2\sim K}\mathbf{O_V})_{\bar{i}}}{C_V\mathbf{V}_1(i)} \quad \text{and} \quad (\mathbf{R_U})_{\bar{i}} = \frac{(\mathbf{U}_{2\sim K}\mathbf{O_U})_{\bar{i}}}{C_U\mathbf{U}_1(i)}. \tag{4.7}$$

Namely we divide each row of the matrix $\mathbf{V}$ by its first entry and then collect the 2nd to Kth columns to form the ratio matrix $\mathbf{R_V}$. The matrix $\mathbf{R_U}$ is similar. Note that

$$(\mathbf{R_V})_{\bar{i}} = \frac{(\mathbf{V}_{2\sim K}O_V)_{\bar{i}}}{C_V\mathbf{V}_1(i)} = \frac{(\mathbf{V}_{2\sim K})_{\bar{i}}\mathbf{O_V}}{C_V\mathbf{V}_1(i)} \overset{\text{(i)}}{=} \frac{\frac{\boldsymbol{\delta}(i)}{\|\boldsymbol{\delta}^{(c_i)}\|}(\mathbf{H}_{2\sim K})_{\bar{c}_i}\mathbf{O_V}}{\frac{\boldsymbol{\delta}(i)}{\|\boldsymbol{\delta}^{(c_i)}\|}C_V\mathbf{H}_1(c_i)} = \frac{(\mathbf{H}_{2\sim K}O_V)_{\bar{c}_i}}{C_V\mathbf{H}_1(c_i)}, \tag{4.8}$$

where (i) follows from eq. (4.3).

Comparing eq. (4.8) with $\mathbf{V}_{\bar{i}} = \frac{\boldsymbol{\delta}(i)}{\|\boldsymbol{\delta}^{(c_i)}\|}\mathbf{H}_{\bar{c}_i}$ in eq. (4.3), we observe that the ratio matrix $\mathbf{R_V}$ in eq. (4.8) does not contain the heterogeneous parameters, and the corresponding row of each node i in $\mathbf{R_V}$, i.e., $(\mathbf{R_V})_{\bar{i}}$, is determined only by c_i, which denotes the community that node i belongs to. This implies that if the nodes are in the same community, then their corresponding rows in $\mathbf{R_V}$ are the same. The same argument is also applicable to the ratio matrix $\mathbf{R_U}$. This explains the importance of the ratio step in Algorithm 1. Our next result formally legitimates the ratio matrix $\mathbf{R} \equiv [\mathbf{R_V}, \mathbf{R_U}]$ for clustering.

Proposition 3 *For the ratio matrix* $\mathbf{R} = [\mathbf{R_V}, \mathbf{R_U}]$ *generated by the singular vectors of the matrix* $\mathbf{\Omega}$*, and for* $1 \leqslant i \leqslant n$ *and* $1 \leqslant j \leqslant n$*, the following inequalities hold:*

$$\|\mathbf{R}_{\bar{i}} - \mathbf{R}_{\bar{j}}\| \geqslant 2 \quad \textit{if} \quad c_i \neq c_j, \quad \textit{and} \quad \|\mathbf{R}_{\bar{i}} - \mathbf{R}_{\bar{j}}\| = 0 \quad \textit{if} \quad c_i = c_j.$$

Proof The proof can be found in Appendix A.3. ■

Proposition 3 states that if the nodes are in the same community, then their corresponding rows in $\mathbf{R} \equiv [\mathbf{R_V}, \mathbf{R_U}]$ are the same. Otherwise if they are in different communities, their corresponding rows are sufficiently different. Proposition 3 also implies that the ratio matrix $\mathbf{R}$ has exactly K different rows due to the fact that there are only K communities in the graph. Thus, the ratio matrix $\mathbf{R}$ has the desirable properties for spectral clustering.

We then generate another ratio matrix $\hat{\mathbf{R}} = [\mathbf{R}_{\hat{\mathbf{V}}}, \mathbf{R}_{\hat{\mathbf{U}}}]$, where $\mathbf{R}_{\hat{\mathbf{V}}}$ and $\mathbf{R}_{\hat{\mathbf{U}}}$ are generated from $\hat{\mathbf{V}}$ and $\hat{\mathbf{U}}$ in the way similar to the generation of $\mathbf{R_V}$ and $\mathbf{R_U}$ from $\mathbf{V}$ and $\mathbf{U}$. The exact definitions of $\mathbf{R}_{\hat{\mathbf{V}}}$ and $\mathbf{R}_{\hat{\mathbf{U}}}$ are in eq. (3.1). Note that, $\hat{\mathbf{R}}$ is the ratio matrix generated from the random adjacency matrix $\mathbf{A}$, whereas $\mathbf{R}$ is the ratio matrix generated from the expected matrix of $\mathbf{A}$, i.e., the $\mathbf{\Omega}$.

To bound the distance between the ratio matrices $\mathbf{R}$ and $\hat{\mathbf{R}}$, define a quantity err_n,

$$err_n \equiv \frac{\max\{\boldsymbol{\theta}_{\max}, \boldsymbol{\delta}_{\max}\} \max\{\|\boldsymbol{\theta}\|_1, \|\boldsymbol{\delta}\|_1\}}{\min\{\boldsymbol{\theta}^2_{\min}, \boldsymbol{\delta}^2_{\min}\} \min\{\|\boldsymbol{\theta}\|^2, \|\boldsymbol{\delta}\|^2\}}, \tag{4.9}$$

which characterizes the effect of heterogeneous parameters on the difference between $\mathbf{R}$ and $\hat{\mathbf{R}}$ as shown in Proposition 4.

Proposition 4 *For* $\mathbf{R} = [\mathbf{R_V}, \mathbf{R_U}], \hat{\mathbf{R}} = [\mathbf{R}_{\hat{\mathbf{V}}}, \mathbf{R}_{\hat{\mathbf{U}}}]$*, and* n *large enough, with probability at least* $1 - O(n^{-4})$*, we have*

$$\|\hat{\mathbf{R}} - \mathbf{R}\|_F^2 \leqslant C T_n^2 \log(n) err_n. \tag{4.10}$$

Proof The proof can be found in Appendix A.4. ■

We then analyze the matrix $\mathbf{M}^*$ which is defined as the output matrix of step 2 in Algorithm 1. In fact, $\mathbf{M}^*$ is the matrix with exactly K different rows and nearest to the ratio matrix $\hat{\mathbf{R}}$ in term of Frobenius norm. In the following proposition, we bound the distance of $\mathbf{M}^*$ and the ratio matrix $\mathbf{R}$, so that the properties of $\mathbf{R}$ in Proposition 3 can serve as a good approximation of the properties of $\mathbf{M}^*$. The proof of Proposition 5 is based on Proposition 4 and the definition of $\mathbf{M}^*$.

Proposition 5 *For* n *large enough, with probability at least* $1 - O(n^{-4})$*, we have*

$$\|\mathbf{M}^* - \mathbf{R}\|_F^2 \leqslant T_n^2 \log(n) err_n.$$

Proof The proof can be found in Appendix A.5. ■

In order to present our main theorem for the D-SCORE algorithm, we first define the following notation for convenience. Let $\mathcal{V}$ denote the set of all the nodes in the graph and let $\mathcal{W}$ be the set of nodes that are correctly clustered by the D-SCORE algorithm. Then by definition, $\mathcal{V} \backslash \mathcal{W}$ is the set of incorrectly clustered nodes, i.e., the nodes which are misclustered by the algorithm. Recall that n_i denotes the number of nodes in community i, for $i = 1, \cdots, K$. The following theorem establishes the bound on the number of misclustered notes for D-SCORE.

Theorem 1 (Convergence of D-SCORE) *Consider directed-DCBM, for which Assumption 1 and Assumption 2 hold. Suppose* $|\mathcal{V}\backslash\mathcal{W}| < \min\{n_1, n_2 \cdots n_K\}$. *Let* $\mathcal{W} \equiv \{1 \leqslant i \leqslant n : \|M_{\bar{i}}^* - R_{\bar{i}}\| \leqslant \frac{1}{2}\}$. *Then nodes in the set* $\mathcal{W}$ *are correctly clustered by the D-SCORE algorithm. Furthermore, for* n *large enough, with probability at least* $1 - o(n^{-4})$,

$$|\mathcal{V}\backslash\mathcal{W}| \leqslant CT_n^2 \log(n) err_n. \tag{4.11}$$

Proof The proof can be found in Appendix A.6. ■

We note that the assumption $|\mathcal{V}\backslash\mathcal{W}| < \min\{n_1, n_2 \cdots n_K\}$ in Theorem 1 guarantees that D-SCORE clusters at least one node in each community correctly. A Similar assumption was also made in Jin (2015) to show the performance guarantee for SCORE algorithm.

To further understand Theorem 1, we consider a simple situation, in which the heterogeneous parameters $\boldsymbol{\theta}$ and $\boldsymbol{\delta}$ are bounded by constants, i.e., $0 < \alpha \leqslant \boldsymbol{\theta}, \boldsymbol{\delta} \leqslant \beta \leqslant 1$. (Note that the special case of the stochastic block model Holland et al. (1983) has $\boldsymbol{\theta}$ and $\boldsymbol{\delta}$ to be constant.) In such a case, $err_n \leqslant \frac{\beta^2}{\alpha^4}$, i.e., it is bounded by a constant. Hence, the error bound of Theorem 1 is in the order of $O(T_n^2 \log(n))$. Typically, we take $T_n = \log(n)$, and then the misclustering rate satisfies

$$\lim_{n\to\infty} \frac{|\mathcal{V}\backslash\mathcal{W}|}{n} \leqslant \lim_{n\to\infty} \frac{C\log^3(n)}{n} = 0.$$

4.2. Performance Guarantee for D-SCOREq

The general idea of the analysis of D-SCORE$_q$ is similar to that of D-SCORE with some technical differences. Hence, here we directly present the main theorem for D-SCORE$_q$ below and relegate the technical proof to Appendix B.

With a little abuse of notations, we reuse $\mathbf{R}, \mathbf{R}_{\mathbf{V}}, \mathbf{R}_{\mathbf{U}}$ and $\hat{\mathbf{R}}, \mathbf{R}_{\hat{\mathbf{V}}}, \mathbf{R}_{\hat{\mathbf{U}}}$ for D-SCORE$_q$, which have slightly different meaning as those for D-SCORE as we explain below. The matrices $\mathbf{R}_{\hat{\mathbf{V}}}$ and $\mathbf{R}_{\hat{\mathbf{U}}}$ are defined in eq. (3.2), and $\mathbf{R}_{\mathbf{V}}$ and $\mathbf{R}_{\mathbf{U}}$ are defined as

$$(\mathbf{R}_{\mathbf{V}})_{\bar{i}} = \frac{(\mathbf{V}\mathbf{O}_{\mathbf{V}})_{\bar{i}}}{\|(\mathbf{V}\mathbf{O}_{\mathbf{V}})_{\bar{i}}\|_q} \quad \text{and} \quad (\mathbf{R}_{\mathbf{U}})_{\bar{i}} = \frac{(\mathbf{U}\mathbf{O}_{\mathbf{U}})_{\bar{i}}}{\|(\mathbf{U}\mathbf{O}_{\mathbf{U}})_{\bar{i}}\|_q}, \tag{4.12}$$

for $i = 1, \cdots n$. Thus, we have

$$(\mathbf{R}_{\mathbf{V}})_{\bar{i}} = \frac{(\mathbf{V}\mathbf{O}_{\mathbf{V}})_{\bar{i}}}{\|(\mathbf{V}\mathbf{O}_{\mathbf{V}})_{\bar{i}}\|_q} = \frac{\mathbf{V}_{\bar{i}}\mathbf{O}_{\mathbf{V}}}{\|\mathbf{V}_{\bar{i}}\mathbf{O}_{\mathbf{V}}\|_q} \overset{\text{(i)}}{=} \frac{\frac{\boldsymbol{\delta}(i)}{\|\boldsymbol{\delta}^{(c_i)}\|}\mathbf{H}_{\bar{c}_i}\mathbf{O}_{\mathbf{V}}}{\frac{\boldsymbol{\delta}(i)}{\|\boldsymbol{\delta}^{(c_i)}\|}\|\mathbf{H}_{\bar{c}_i}\mathbf{O}_{\mathbf{V}}\|_q} = \frac{\mathbf{H}_{\bar{c}_i}\mathbf{O}_{\mathbf{V}}}{\|\mathbf{H}_{\bar{c}_i}\mathbf{O}_{\mathbf{V}}\|_q}, \tag{4.13}$$

where (i) follows from eq. (4.3).

Comparing eq. (4.13) with $\mathbf{V}_{\bar{i}} = \frac{\boldsymbol{\delta}(i)}{\|\boldsymbol{\delta}^{(c_i)}\|}\mathbf{H}_{\bar{c}_i}$ in eq. (4.3), we observe that the ratio matrix $\mathbf{R}_{\mathbf{V}}$ in eq. (4.13) does not contain factor $\frac{\boldsymbol{\delta}(i)}{\|\boldsymbol{\delta}^{(c_i)}\|}$ of the heterogeneous parameters, and the corresponding row of each node i in $\mathbf{R}_{\mathbf{V}}$, i.e., $(\mathbf{R}_{\mathbf{V}})_{\bar{i}}$, is determined only by c_i, which denotes the community that node i belongs to. This implies that if these nodes are in the same community, and then their corresponding rows in $\mathbf{R}_{\mathbf{V}}$ are the same. The same

argument is also applicable to the matrix $\mathbf{R_U}$. This explains the importance of the ratio step in Algorithm 2 and also explains why D-SCOREq is as powerful as D-SCORE.

We are now ready to present the main theorem for the D-SCORE$_q$ algorithm as follows.

Theorem 2 (Convergence of D-SCORE$_q$) *Consider the directed-DCBM under Assumption 1 and Assumption 2. Suppose* $|\mathcal{V}\backslash\mathcal{W}| < \min\{n_1, n_2 \cdots n_K\}$. *Let* $\mathcal{W} \equiv \{1 \leqslant i \leqslant n : \|\mathbf{M}^*_{\bar{i}} - \mathbf{R}_{\bar{i}}\| \leqslant \frac{C}{2}\}$. *Then there exists a constant* C, *such that nodes in the set* $\mathcal{W}$ *are correctly clustered by the D-SCORE$_q$ algorithm. Furthermore, for* n *large enough, with probability at least* $1 - o(n^{-4})$,

$$|\mathcal{V}\backslash\mathcal{W}| \leqslant CT_n^2 \log(n) err_n. \tag{4.14}$$

Proof See Appendix B. ■

5. Experiments

In this section, we conduct experimental studies to compare the performance of six spectral clustering algorithms, namely, D-SCORE, D-SCORE$_q$, rD-SCORE, rD-SCORE$_q$, oPCA, rPCA, and two likelihood algorithms APL (Amini et al., 2013) and BCPL (Bickel and Chen, 2009b). We compare these eight algorithms on the web blogs data and the experiments on simulated data.

5.1. Algorithms

Among the algorithms that we compare in the experiments, D-SCORE and D-SCORE$_q$ correspond to Algorithm 1 and Algorithm 2 in this paper. The algorithm oPCA (see Algorithm 4) is the original spectral clustering method, which collects the singular vectors of the adjacency matrix into one matrix and runs K-means on such a matrix. Furthermore, for these algorithms, instead of directly dealing with adjacency matrix $\mathbf{A}$, a pre-processing step called *regularized graph Laplacian* (Rohe et al., 2016; Joseph and Yu, 2016) (see Algorithm 5) can be added to regularize the adjacency matrix $\mathbf{A}$. Hence, correspondingly, rPCA first regularizes the adjacency matrix $\mathbf{A}$ to generate a regularized graph Laplacian $\mathbf{L}$ (as in Algorithm 5), and then applies oPCA to $\mathbf{L}$. Similarly, rD-SCORE first generates a regularized graph Laplacian $\mathbf{L}$ and then applies D-SCORE (Algorithm 1) to $\mathbf{L}$. The rD-SCORE$_q$ follows the similar regularization procedure of rD-SCORE, but applies D-SCORE$_q$ (Algorithm 2) to $\mathbf{L}$ instead of D-SCORE. Specially for $q = 2$, rD-SCORE$_2$ is almost the same as the DI-SIM algorithm in Rohe et al. (2016). The only difference lies in that Rohe et al. (2016) provided a bi-clustering structure, whereas rD-SCORE$_2$ provides a single cluster structure for nodes. Similarly, rD-SCORE$_q$ can be seen as an extension of the DI-SIM algorithm from the ℓ_2-norm to the ℓ_q-norm for any positive integer q.

5.2. Applications to Real Data Sets

5.2.1. Applications to Political Blogs Data

In this subsection, we apply the above mentioned eight algorithms to the web blogs data introduced in Adamic and Glance (2005). The blogs data was collected at 2004 presidential

Algorithm 4: oPCA

Input : The number K of communities and the adjacency matrix A.

1 Obtain the first K leading left and right singular vector matrices V and U of A.

2 Put V and U together to form a matrix $R = [V, U]$, and apply the K-means method to R.

Output: The community labels of the nodes in the adjacency matrix A.

Algorithm 5: Regularized graph Laplacian

Input : The adjacency matrix A.

1 Calculate the diagonal matrix O^τ, $P^\tau \in R^{n\times n}$, where $O^\tau(i,i) = \tau + \sum_{j=1}^{n} A(i,j)$ and $P^\tau(i,i) = \tau + \sum_{j=1}^{n} A(j,i)$. The regularization parameter τ is usually set as the average degree $\tau = \sum_{i,j=1}^{n} A(i,j)/n$.

2 Let $L = (O^\tau)^{-1/2} A (P^\tau)^{-1/2}$.

Output: The regularized graph Laplacian matrix L.

election. Such political blogs data can be represented by a directed graph, in which each node in the graph corresponds to a web blog labelled either as liberal or conservative. An directed edge from node i to node j indicates that there is a hypelink from blog i to blog j. Clearly, such a political blog graph is *directed*. The fact that there is a hyperlink from blog i to j does not imply there is also a hypelink from blog j to i. Hence, the adjacency matrix of the political blogs data is an asymmetric matrix.

In our experiment, we first extract the largest component of the graph, which contains 1222 nodes, and denote it by an asymmetric directed adjacency matrix $\mathbf{A}$. Then, we extract the largest components of $\mathbf{A}^T\mathbf{A}$ and $\mathbf{A}\mathbf{A}^T$, and use $\mathcal{S}_r$ and $\mathcal{S}_l$ to denote the node sets of these two largest connected components, respectively. We define the intersection set $\mathcal{S} \equiv \mathcal{S}_r \cap \mathcal{S}_l$, which contains 823 nodes.

We run all of the six spectral algorithms in the following two different approaches. In the first approach, we run these six algorithms on the entire graph that contains 1222 nodes. In the second approach, we first run the six algorithms on the intersection set $\mathcal{S}$, and then we use the attachment technique to attach nodes outside $\mathcal{S}$ to clusters (as described in Algorithm 3). We repeat each algorithm on each setting 500 times and take the mean of the total number of misclustered nodes. Since APL and BCPL are designed for undirected network, we first build a symmetric adjacency matrix based on the asymmetric one, and then apply APL and BCPL to the symmetric one. Since symmetric network does not have intersection approach, we count the misclustered node of APL and BCPL in intersection approach directly form their result in entire graph approach while limited the node only in the intersection set calculated in DSCORE algorithms.

	Entire Graph (1222)	Int. with Attach. (1222)	Intersection (823)
oPCA	434	300	217
rPCA	414	246	190
BCPL	379	379	236
APL	61	61	28
DSCORE	142	60	22
rDSCORE	139	60	22
DSCORE2	141	61	23
rDSCORE2	142	60	26

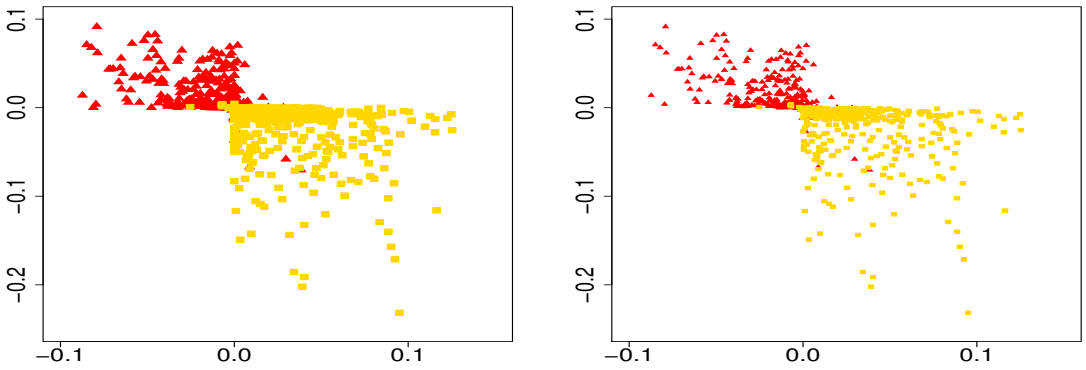
Table 1: Misclustered nodes in political blog data.

The experiment results are shown in Table 1. It can be observed from the table that D-SCORE and D-SCORE$_q$ almost have the same performance, and rD-SCORE and rD-SCORE$_q$ almost have the same performances. Furthermore, D-SCORE and D-SCORE$_2$ perform better than oPCA, which implies that the ratio step to remove the heterogeneous parameters helps greatly to improve the clustering accuracy. The same occurs in the comparison of the algorithms with the regularized graph Laplacian. Moreover, APL almost performs the same as DSCORE type algorithms while BCPL doesn't. We will show in the stimulation section that the performance of APL is easily affected by the community structure, and cause its performance unstable.

Next, by comparing the first and second columns in Table 1, we observe that for all algorithms, it is much better to run the algorithms on the intersection set and then attach the outside nodes than directly running the algorithm on the entire graph. Especially, the intersection-with-attachment technique introduced in Algorithm 3 has improved all the original D-SCORE algorithms for the entire graph.

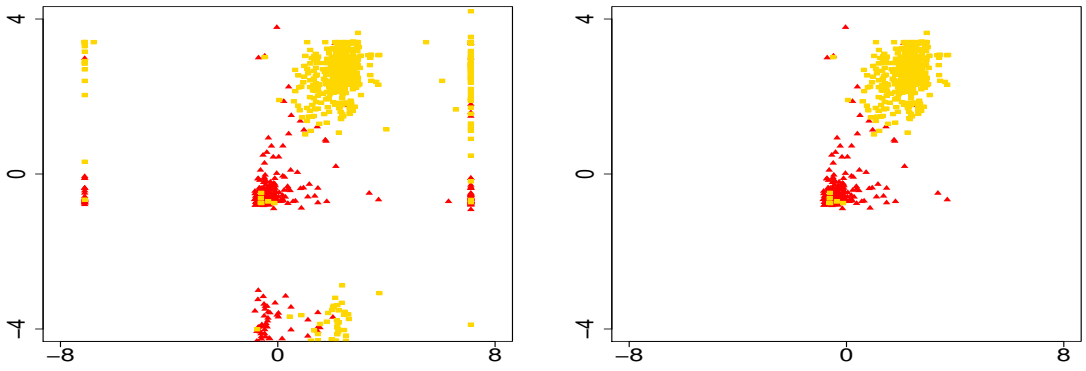
To explain this improvement further, we plot the vectors that the algorithms (i.e., oPCA, D-SCORE, D-SCORE$_q$) use in the clustering in Figs. 1 to 3. Note that in these algorithms, before the $k-$means step, each node corresponds to one row of a matrix. We thus use these row vectors as the coordinate of the nodes and plot them in the figures. The Figs. 1a, 2a and 3a include the nodes in the entire graph. The Figs. 1b, 2b and 3b include the nodes in the intersection set. We use red triangles and yellow squares to represent nodes in the liberal and conservative communities, respectively. Note that extreme coordinates in Figs. 2a and 3a are already thresholded and form the imaginary borders for better presentation; these extreme coordinates are the effect of having extremely small numbers (computational errors for 0) as the denominators when D-SCORE and D-SCORE$_q$ are used directly on the entire graph, as explained in Section 3.

First, we compare Fig. 1 (which applies the original spectral clustering) with Figs. 2 and 3 (which apply the D-SCORE and D-SCORE$_2$, respectively). It is clear that nodes in Figs. 2 and 3 are much more separable than nodes in Fig. 1 due to the ratio step in D-SCORE and D-SCORE$_2$. Furthermore, We observe that the intersection graph (Figs. 2b and 3b) extracts the center of the entire graph and deletes nodes near the border in Figs. 2a and 3a, which act as noise and mislead the clustering result. The intersection-with-attachment technique works by taking the clustering results for the intersection (shown here) and attaching



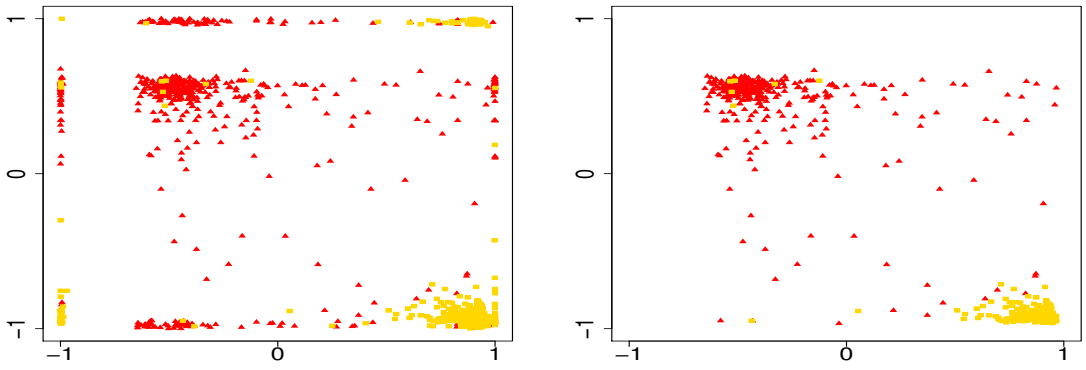
a Entire Graph b Intersection

Figure 1: Comparison of the clustering vectors in the entire graph and in the intersection graph of original spectral clustering. The x-axis is the second leading left singular vector, and the y-axis is the second leading right singular vector.



a Entire Graph b Intersection

Figure 2: Comparison of the clustering vectors of entire graph and the intersection in D-SCORE. The x-axis is the left ratio vector, and the y-axis is the right ratio vector.



a Entire Graph b Intersection

Figure 3: Comparison of the singular vectors of entire graph and the intersection in D-SCORE$_q$. The x-axis is the second left ratio vector, and the y-axis is the second right ratio vector.

the noise nodes to these clusters using the links in the original network A (not shown here), and hence yields better performance.

5.2.2. Applications to Email-Eu-Core Network

In this subsection, we apply the above mentioned eight algorithms to the email-Eu-core network introduced in Leskovec and Krevl (2014). The email data was collected from a large European research institution, and a directed edge from node i to node j indicates that person i has sent at least one email to person j. Clearly, the email-Eu-core network is also a directed network. There are many communities in this network, but we extract the top 4 largest communities which contains 297 nodes as the entire graph and 252 nodes in intersection graph. We repeat the experiment 500 times and show the mean error in table 2. The experimental observation is similar with that of the political blog data, and thus we omit it for brevity.

	Entire Graph (297)	Int. with Attach. (297)	Intersection (252)
oPCA	107	78	72
rPCA	89	57	53
BCPL	23	23	18
APL	17	17	12
DSCORE	23	7	6
rDSCORE	25	7	6
DSCORE2	15	4	3
rDSCORE2	16	4	3

Table 2: Misclustered nodes in email-Eu-core network.

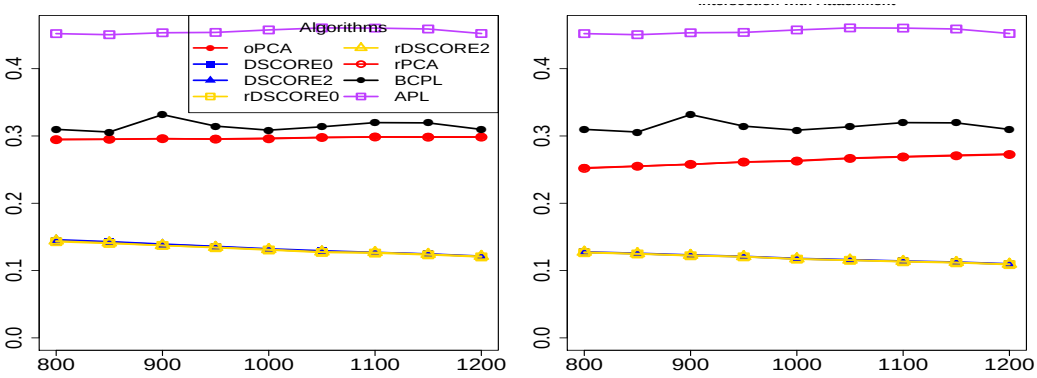
5.3. Simulations

In this section, we compare the eight algorithms described in Section 5.1 through a series of simulations. In the experiments, we first generate an adjacency matrix $\mathbf{A}_0$ by Directed-DCBM, and then extract the largest connected component $\mathbf{A}$ of $\mathbf{A}_0$ with the node set of $\mathbf{A}$ denoted by $\mathcal{S}_0$. We also extract the largest connected components of $\mathbf{A}^T\mathbf{A}$ and $\mathbf{A}\mathbf{A}^T$, and denote the node sets as $\mathcal{S}_1$ and $\mathcal{S}_2$, respectively. Let $\mathcal{S} = \mathcal{S}_1 \cap \mathcal{S}_2$. We also apply the six spectral algorithms in two approaches: (i) the entire graph approach, where we run the six algorithms over the set $\mathcal{S}_0$; and (ii) intersection-with-attachment approach, where we run the six algorithms over the intersection set $\mathcal{S}$, and then use the attachment technique to cluster nodes outside the intersection set. The usage of APL and BCPL in simulation is the same as that in real data experiment. Since the symmetric adjacency matrix does not have intersection set issue, we directly plot the result of APL and BCPL in entire graph approach in the intersection with attachment approach for comparison.

5.3.1. Block Model with Symmetric Structure

In this experiment, we generate the data by DCBM by setting the heterogeneous parameters $\boldsymbol{\theta}$ such that $P(\boldsymbol{\theta}(i) = 0.5) = 0.01$, $P(\boldsymbol{\theta}(i) = 0.1) = 0.05$ and $P(\boldsymbol{\theta}(i) = 0.6) = 0.4$. We set

$\boldsymbol{\delta}(i) = \boldsymbol{\theta}(i)$ for all $i \in \{1, \cdots n\}$. Also, we set the block matrix $\mathbf{B} = \begin{bmatrix} 1, & 0.4 \\ 0.4, & 1 \end{bmatrix}$, which is symmetric. Let $K = 2$. Then, we uniformly randomly assign community labels to nodes and let the total number n of nodes go from 800 to 1200 with the step size 50. For each n, we repeat the experiment 500 times and Fig. 4 plots the average of the misclustered rate.



a Entire graph b Intersection with attachment

Figure 4: Comparison of the misclustering rate under SBM with symmetric structure. The horizontal axis is the number of nodes in the entire graph, and the vertical axis is the minsclutering rate.

It can be observed that although the model is symmetric, D-SCORE and D-SCORE$_q$ still perform better than oPCA, APL and BCPL, and the performance is similar with its corresponding pre-precessing version. Also, by comparing Figs. 4a and 4b, we observe that the intersection-with-attachment technique improves all variants of the D-SCORE algorithms.

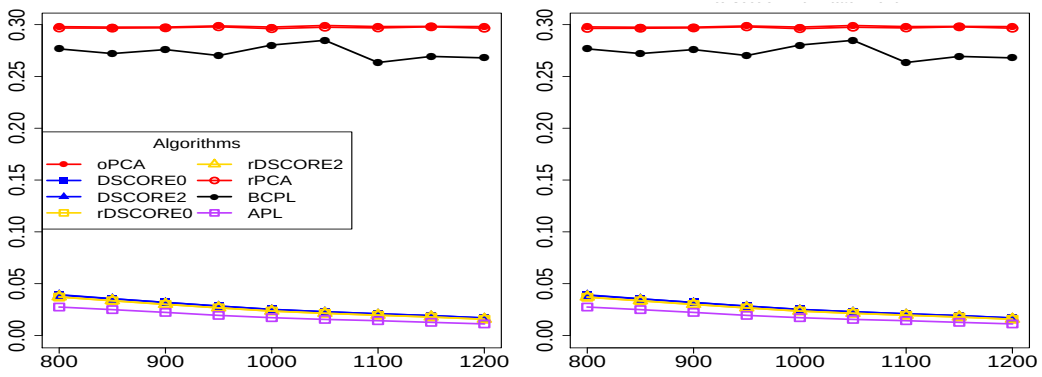
5.3.2. DCBM with Symmetric and Dense Structure

In this experiment, we set the block matrix $\mathbf{B} = \begin{bmatrix} 1, & 0.4 \\ 0.4, & 1 \end{bmatrix}$ with two communities. We randomly choose the heterogeneous parameter $\boldsymbol{\theta}$ for nodes with $P(\boldsymbol{\theta}(i) = 0.5) = 0.05$, $P(\boldsymbol{\theta}(i) = 0.1) = 0.05$ and $P(\boldsymbol{\theta}(i) = 0.6) = 0.4$. We set $\boldsymbol{\delta}(i) = \boldsymbol{\theta}(i)$ for all $i \in \{1, \cdots n\}$. Other parameters are chosen to the same as the previous experiment.

The mean of misclustering rate is plotted in Fig. 5. It can be observed that DSCORE, DSCORE$_q$, rDSCORE and rDSCORE$_q$ have almost the same performance and perform much better than oPCA and rPCA. This implies that the ratio technique in these algorithms greatly helps to improve the clustering accuracy. The performance of BCPL is better than oPCA and rPCA while worse than the proposed algorithms. What surprises us is that APL performs pretty well in this setting.

5.3.3. DCBM with Asymmetric and Sparse Structure

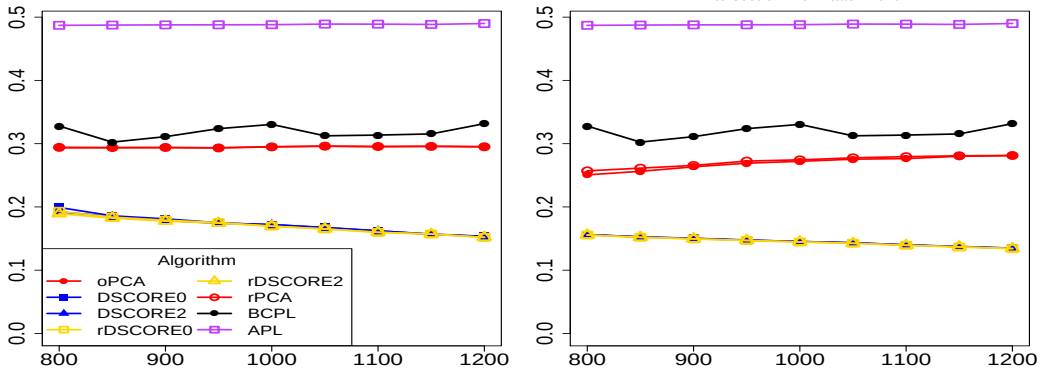
In this experiment, we set the block matrix $\mathbf{B} = \begin{bmatrix} 1, & 0.4 \\ 0.5, & 1 \end{bmatrix}$, the number of communities $K = 2$, and the heterogeneous parameter $\boldsymbol{\theta}$ such that $P(\boldsymbol{\theta}(i) = 0.5) = 0.01$, $P(\boldsymbol{\theta}(i) = 0.1) =$



a Entire graph b Intersection with attachment

Figure 5: Comparison of the misclustering rate under DCBM with symmetric structure. The vertical axis is the number of nodes in the entire graph, and the horizontal axis is the minsclutering rate.

0.01 and $P(\boldsymbol{\theta}(i) = 0.6) = 0.4$. In this experiment, we randomly pick $\boldsymbol{\delta}$ in the same way as $\boldsymbol{\theta}$ instead of setting $\boldsymbol{\theta}(i) = \boldsymbol{\delta}(i)$, which increases the asymmetric structure of the model. Other parameters are chosen to the same as the previous experiment. The mean of the misclustering rate is plotted in Fig. 6.



a Entire graph b Intersection with attachment

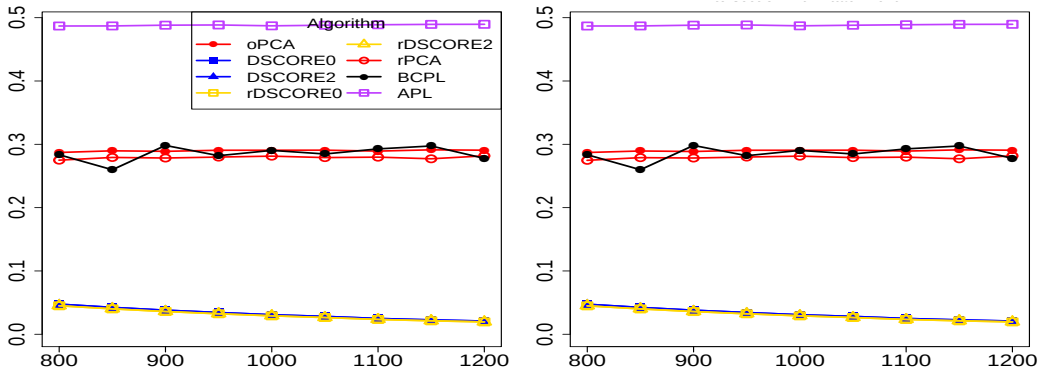
Figure 6: Comparison of the misclustering rate under DCBM with asymmetric and sparse structure. The horizontal axis is the number of nodes in the entire graph, and the vertical axis is the minsclutering rate.

We observe form Fig. 6b that DSCORE, DSCORE$_q$, rDSCORE and rDSCORE$_q$ perform the same and are better than oPCA, rPCA, APL and BCPL, which implies that the ratio technique greatly helps. Also, by comparing Figs. 6a and 6b, we observe that the intersection-with-attachment approach performs better than the entire graph approach.

5.3.4. DCBM with Asymmetric and Dense Structure

In this experiment, we set the block matrix $\mathbf{B} = \begin{bmatrix} 1, & 0.4 \\ 0.5, & 1 \end{bmatrix}$, the number of communities $K = 2$, and the heterogeneous parameter $\boldsymbol{\theta}$ such that $P(\boldsymbol{\theta}(i) = 0.5) = 0.05$, $P(\boldsymbol{\theta}(i) = 0.1) =$

0.01 and $P(\boldsymbol{\theta}(i) = 0.6) = 0.4$. The parameter $\boldsymbol{\delta}$ is randomly picked in the same way as $\boldsymbol{\theta}$. Other parameters are chosen to the same as the previous experiment. The mean of the misclustering rate is plotted in Fig. 7.



a Entire graph b Intersection with attachment

Figure 7: Comparison of the misclustering rate under DCBM with asymmetric and dense structure. The horizontal axis is the number of nodes in the entire graph, and the vertical axis is the minsclutering rate.

Here, our setting of parameters makes the graph denser than that in the previous experiment (Section 5.3.3). It can be seen that the performance of the entire graph is almost the same as that of the intersection with attachment. This suggests that the intersection-with-attachment technique is more efficient for sparse networks. This should not be surprising because, for dense networks, the nodes are more connected and noise nodes that have low degrees and need the attachment step are reduced.

6. Conclusion

In this paper, we provided theoretical guarantee and experimental results for two spectral clustering algorithms for networks with directed edges. In theory, we established the performance guarantee for D-SCORE and D-SCORE$_q$ under Direct-DCBM. We also conducted extensive experiments to demonstrate the advantage of the improved D-SCORE algorithms over the original version and the competitive algorithms. As an extension, since the translation of network structures into Euclidean coordinates using D-SCORE and SCORE can be easily extended to multi-layer networks and node-attributed networks, the theory presented in this paper can be potentially extended to those more general scenarios.

Acknowledgments

Z. Wang and Y. Liang would like to thank the partial support of the U.S. National Science Foundation under the grants ECCS-1818904 and CCF-1801855. The authors appreciate the valuable discussion with Jiashun Jin at Carnegie Mellon University.

Appendices

Appendix A. Proof of Theorem 1 (Convergence of D-SCORE)

We first provide the proofs for Propositions 1-5, and then combine all these properties together to prove Theorem 1. Note that all the propositions and lemmas that we show below need Assumption 1 and Assumption 2 to hold.

A.1. Proof of Proposition 1

Proof We first let $\mathbf{\Theta}_\theta$ and $\mathbf{\Theta}_\delta$ denote the $n \times K$ matrices such that for $1 \leqslant i \leqslant n$ and $1 \leqslant k \leqslant K$,

$$\mathbf{\Theta}_\theta(i,k) = \begin{cases} \frac{\boldsymbol{\theta}(i)}{\|\boldsymbol{\theta}^{(k)}\|} & \text{if} \quad c_i = k \\ 0 & \text{if} \quad c_i \neq k \end{cases} \quad \text{and} \quad \mathbf{\Theta}_\delta(i,k) = \begin{cases} \frac{\boldsymbol{\delta}(i)}{\|\boldsymbol{\theta}^{(k)}\|} & \text{if} \quad c_i = k \\ 0 & \text{if} \quad c_i \neq k \end{cases}.$$

The matrix $\mathbf{\Theta}_\theta$ serves as a membership matrix with each row, say, the ith row, containing only one nonzero entry, whose column index corresponds to the community that node i belongs to.

Then by the above definitions of $\mathbf{\Theta_\theta}, \mathbf{\Theta_\delta}$ and the definitions of $\mathbf{\Psi_\theta}, \mathbf{\Psi}_\delta$ (see eq. (4.1)), we can express the expectation matrix $\mathbf{\Omega} = (\mathbf{\Theta_\theta}\|\boldsymbol{\theta}\|\mathbf{\Psi_\theta})\mathbf{B}(\mathbf{\Theta_\delta}\|\boldsymbol{\delta}\|\mathbf{\Psi_\delta})^T$. Denoting $\mathbf{S} \equiv \mathbf{\Psi_\theta}\mathbf{B}\mathbf{\Psi}_{\boldsymbol{\delta}}^T$, we obtain

$$\mathbf{\Omega} = \|\boldsymbol{\theta}\|\|\boldsymbol{\delta}\|\mathbf{\Theta_\theta}\mathbf{S}\mathbf{\Theta}_{\boldsymbol{\delta}}^T. \tag{A.1}$$

Since the diagonal matrices $\mathbf{\Psi_\theta}$ and $\mathbf{\Psi_\delta}$ are of full rank, $\text{rank}(\mathbf{S}) = \text{rank}(\mathbf{\Psi_\theta}\mathbf{B}\mathbf{\Psi}_{\boldsymbol{\delta}}^T) = \text{rank}(\mathbf{B}) = K$. Thus, the $K \times K$ matrix $\mathbf{S}$ is also of full rank and has only non-zero singular values. Then, we denote the SVD of the matrix $\mathbf{S}$ as

$$\mathbf{S} = \mathbf{Y}\mathbf{\Lambda_S}\mathbf{H}^T, \tag{A.2}$$

where $\mathbf{\Lambda_S}$ is a $K \times K$ non-zero diagonal matrix with the singular values arranged in a decreasing order, and $\mathbf{H}$ and $\mathbf{Y}$ are $K \times K$ orthogonal matrices.

We substitute eq. (A.2) into eq. (A.1) and obtain

$$\mathbf{\Omega} = \|\boldsymbol{\theta}\|\|\boldsymbol{\delta}\|(\mathbf{\Theta_\theta}\mathbf{Y})\mathbf{\Lambda_S}(\mathbf{H}\mathbf{\Theta_\delta})^T. \tag{A.3}$$

By the definitions of $\mathbf{\Theta_\theta}$ and $\mathbf{\Theta_\delta}$, $\mathbf{\Theta}_{\boldsymbol{\theta}}^T\mathbf{\Theta_\theta} = \mathbf{I}$ and $\mathbf{\Theta}_{\boldsymbol{\delta}}^T\mathbf{\Theta_\delta} = \mathbf{I}$. Thus,

$$\begin{aligned} (\mathbf{\Theta_\theta}\mathbf{Y})^T\mathbf{\Theta_\theta}\mathbf{Y} &= \mathbf{Y}^T\mathbf{\Theta}_{\boldsymbol{\theta}}^T\mathbf{\Theta_\theta}\mathbf{Y} = \mathbf{I}, \\ (\mathbf{\Theta_\delta}\mathbf{H})^T\mathbf{\Theta_\delta}\mathbf{H} &= \mathbf{H}^T\mathbf{\Theta}_{\boldsymbol{\delta}}^T\mathbf{\Theta_\delta}\mathbf{H} = \mathbf{I}. \end{aligned} \tag{A.4}$$

By eq. (A.4), we observe that $\mathbf{\Theta_\theta}\mathbf{Y}$ and $\mathbf{\Theta_\delta}\mathbf{H}$ have orthogonal columns. Thus, eq. (A.3) is the compact SVD of the matrix $\mathbf{\Omega}$. Denoting the compact SVD of $\mathbf{\Omega}$ as $\mathbf{\Omega} = \mathbf{U}\mathbf{\Lambda}\mathbf{V}^T$, we have

$$\mathbf{V} = \mathbf{\Theta_\delta}\mathbf{H}, \tag{A.5}$$

$$\mathbf{U} = \mathbf{\Theta_\theta}\mathbf{Y}, \tag{A.6}$$

$$\mathbf{\Lambda} = \|\boldsymbol{\theta}\|\|\boldsymbol{\delta}\|\mathbf{\Lambda_S} \tag{A.7}$$

where $\mathbf{\Lambda}$ is a $K \times K$ non-zero diagonal matrix, and $\mathbf{V}$ and $\mathbf{U}$ are $n \times K$ matrices with orthogonal columns.

By eq. (A.7), $\mathbf{\Omega}$ has only K non-zero singular values because $\mathbf{\Lambda}_S$ is a $K \times K$ non-zero diagonal matrix, i.e., $\sigma_i(\mathbf{\Omega}) = \|\boldsymbol{\theta}\|\|\boldsymbol{\delta}\|\sigma_i(\mathbf{S})$ for $i \leqslant K$ and $\sigma_i(\mathbf{\Omega}) = 0$ for $i > K$. Therefore, eq. (4.3) follows from the forms of individual rows of eqs. (A.5) and (A.6).

Since $\mathbf{H}$ is an orthogonal matrix, eq. (4.6) follows because $\|\mathbf{V}_{\bar{\boldsymbol{i}}}\| = \left\| \frac{\boldsymbol{\delta}(i)}{\|\boldsymbol{\delta}^{(c_i)}\|} \mathbf{H}_{\bar{\boldsymbol{c}}_{\boldsymbol{i}}} \right\| = \frac{\boldsymbol{\delta}(i)}{\|\boldsymbol{\delta}^{(c_i)}\|}$. By eq. (2.9), $\|\boldsymbol{V}_{\bar{\boldsymbol{i}}}\| \asymp \frac{\boldsymbol{\delta}(i)}{\|\boldsymbol{\delta}\|}$. Following the arguments similar to the above, we have $\|\mathbf{U}_{\bar{\boldsymbol{i}}}\| \asymp \frac{\boldsymbol{\theta}(i)}{\|\boldsymbol{\theta}\|}$. ■

A.2. Proof of Proposition 2

In Proposition 2, we bound the distance between the singular vectors of $\mathbf{\Omega}$ and those of $\mathbf{A}$. In order to bound such distance, we first show a few lemmas, including Lemma A.1 that establishes the eigenvalues of $\mathbf{\Omega}$ to be at the level of $\|\boldsymbol{\theta}\|^2\|\boldsymbol{\delta}\|^2$, Lemma A.2 that bounds the distance between the random adjacency matrix $\mathbf{A}$ and its expected version $\mathbf{\Omega}$, and Lemma A.3 that lower bounds $\lambda_1(\mathbf{S}^T\mathbf{S}) - \lambda_2(\mathbf{S}^T\mathbf{S})$ away from zero. Combining all these lemmas, we apply Davis-Kahan Theorem (Lemma A.4) to establish Proposition 2.

Now, we formally state the lemmas mentioned above and relegate their proofs to Appendix A.7.

Lemma A.1 *Under Directed-DCBM, for $1 \leqslant i \leqslant K$, we obtain*

$$\lambda_i(\mathbf{\Omega}^T\mathbf{\Omega}) \asymp \|\boldsymbol{\theta}\|^2\|\boldsymbol{\delta}\|^2. \tag{A.8}$$

Proof The proof can be found in Appendix A.7.1. ■

Lemma A.2 *For sufficiently large n, with probability at least $1 - o(n^{-4})$,*

$$\|\mathbf{A} - \mathbf{\Omega}\| \leqslant 6\sqrt{\log(n)Z}. \tag{A.9}$$

Proof The proof can be found in Appendix A.7.2. ■

Lemma A.3 *With $\mathbf{S} = \mathbf{Y}\mathbf{\Lambda}_{\boldsymbol{S}}\mathbf{H}^T$, for $i = 1, \cdots, K$, we have*

$$0 < C \leqslant \mathbf{H}_1(i) \leqslant 1 \quad \textit{and} \quad 0 < C \leqslant \mathbf{Y}_1(i) \leqslant 1, \tag{A.10}$$

$$\lambda_1(\mathbf{S}^T\mathbf{S}) - \lambda_2(\mathbf{S}^T\mathbf{S}) \geqslant C, \tag{A.11}$$

$$\mathbf{V}_1(i) > 0, \mathbf{U}_1(i) > 0 \quad \textit{for} \quad 1 \leqslant i \leqslant n. \tag{A.12}$$

Proof The proof can be found in Appendix A.7.3. ■

From eq. (A.12), we observe that the singular vector corresponding to the largest singular value of $\mathbf{\Omega}$ has all positive entries. Thus, we can use it as the denominator to generate ratio matrix.

The following lemma is a variant of Davis-Kahan theorem.

Lemma A.4 (Yu et al. (2015), Theorem 2) *Let $\mathbf{A}, \hat{\mathbf{A}} \in R^{n\times n}$ be symmetric, with eigenvalues $\lambda_1 \geqslant \cdots \geqslant \lambda_n$ and $\hat{\lambda}_1 \geqslant \cdots \geqslant \hat{\lambda}_n$ and corresponding eigenvectors $\mathbf{v}_1, \cdots, \mathbf{v}_n$ and $\hat{\mathbf{v}}_1, \cdots, \hat{\mathbf{v}}_n$, respectively. Fix $1 \leqslant r \leqslant s \leqslant n$ and assume that $\min(\lambda_{r-1} - \lambda_r, \lambda_s - \lambda_{s+1}) \geqslant 0$, where we define $\lambda_0 = \infty$ and $\lambda_{n+1} = -\infty$. Let $k = s-r+1$, $\mathbf{V} = (\mathbf{v}_r, \mathbf{v}_{(r+1)}, \cdots, \mathbf{v}_s) \in R^{n\times k}$ and $\hat{\mathbf{V}} = (\hat{\mathbf{v}}_r, \hat{\mathbf{v}}_{(r+1)}, \cdots, \hat{\mathbf{v}}_s) \in R^{n\times k}$. Then there exists an orthogonal matrix $O \in R^{k\times k}$ such that*

$$\|\mathbf{V}\boldsymbol{O} - \hat{\mathbf{V}}\| \leqslant \frac{2^{\frac{3}{2}} k^{\frac{1}{2}} \|\mathbf{A} - \hat{\mathbf{A}}\|}{\min(\lambda_{r-1} - \lambda_r, \lambda_s - \lambda_{s+1})}. \tag{A.13}$$

Now, we are ready to prove Proposition 2.

Proof [Proof of Proposition 2] First, we derive

$$\begin{aligned}
\|\mathbf{X}^T\mathbf{X} - \boldsymbol{\Omega}^T\boldsymbol{\Omega}\| &\leqslant \|\mathbf{X}^T\mathbf{X} - \mathbf{X}^T\boldsymbol{\Omega}\| + \|\mathbf{X}^T\boldsymbol{\Omega} - \boldsymbol{\Omega}^T\boldsymbol{\Omega}\| \\
&\leqslant \|\mathbf{X}\|\|\mathbf{X} - \boldsymbol{\Omega}\| + \|\mathbf{X} - \boldsymbol{\Omega}\|\|\boldsymbol{\Omega}\| \\
&\leqslant \|\mathbf{X} - \boldsymbol{\Omega}\|(\|\mathbf{X}\| + \|\boldsymbol{\Omega}\|) \\
&\overset{\text{(i)}}{\leqslant} C\sqrt{\log(n)Z}(2\|\boldsymbol{\Omega}\| + 6\sqrt{\log(n)Z}) \\
&\overset{\text{(ii)}}{\leqslant} C_1\sqrt{\log(n)Z}\|\boldsymbol{\theta}\|\|\boldsymbol{\delta}\| + C_2\log(n)Z,
\end{aligned} \tag{A.14}$$

where (i) follows from Lemma A.2, which shows that $\|\mathbf{X} - \boldsymbol{\Omega}\| \leqslant 6\sqrt{\log(n)Z}$, and hence we have $\|\mathbf{X}\| \leqslant \|\boldsymbol{\Omega}\| + 6\sqrt{\log(n)Z}$, and (ii) follows from Lemma A.1, which implies $\|\boldsymbol{\Omega}\| = \sqrt{\lambda_1(\boldsymbol{\Omega}^T\boldsymbol{\Omega})} \asymp \|\boldsymbol{\theta}\|\|\boldsymbol{\delta}\|$.

Applying Lemma A.4 (Davis-Kahan theorem), we obtain

$$\begin{aligned}
\|\hat{\mathbf{V}}_1 - \mathbf{V}_1 C_V\|_F &\leqslant \frac{C\|\mathbf{X}^T\mathbf{X} - \boldsymbol{\Omega}^T\boldsymbol{\Omega}\|}{\lambda_1(\boldsymbol{\Omega}^T\boldsymbol{\Omega}) - \lambda_2(\boldsymbol{\Omega}^T\boldsymbol{\Omega})} \\
&\overset{\text{(i)}}{\leqslant} \frac{C_1\sqrt{\log(n)Z}\|\boldsymbol{\theta}\|\|\boldsymbol{\delta}\| + C_2\log(n)Z}{\lambda_1(\boldsymbol{\Omega}^T\boldsymbol{\Omega}) - \lambda_2(\boldsymbol{\Omega}^T\boldsymbol{\Omega})} \\
&\overset{\text{(ii)}}{\leqslant} \frac{C_1\sqrt{\log(n)Z}\|\boldsymbol{\theta}\|\|\boldsymbol{\delta}\| + C_2\log(n)Z}{C\|\boldsymbol{\theta}\|^2\|\boldsymbol{\delta}\|^2} \\
&\leqslant C_1\frac{\sqrt{\log(n)Z}}{\|\boldsymbol{\theta}\|\|\boldsymbol{\delta}\|} + C_2\left(\frac{\sqrt{\log(n)Z}}{\|\boldsymbol{\theta}\|\|\boldsymbol{\delta}\|}\right)^2 \\
&\overset{\text{(iii)}}{\leqslant} C_1\frac{\sqrt{\log(n)Z}}{\|\boldsymbol{\theta}\|\|\boldsymbol{\delta}\|},
\end{aligned}$$

where (i) follows from eq. (A.14), (ii) follows from eq. (4.2) and eq. (A.11), which implies that

$$\lambda_1(\boldsymbol{\Omega}^T\boldsymbol{\Omega}) - \lambda_2(\boldsymbol{\Omega}^T\boldsymbol{\Omega}) = \|\boldsymbol{\theta}\|^2\|\boldsymbol{\delta}\|^2(\lambda_1(\mathbf{S}^T\mathbf{S}) - \lambda_2(\mathbf{S}^T\mathbf{S})) \geqslant C\|\boldsymbol{\theta}\|^2\|\boldsymbol{\delta}\|^2, \tag{A.15}$$

and (iii) follows from eq. (2.11), which gives $\lim_{n\to\infty} \frac{\sqrt{\log(n)Z}}{\|\boldsymbol{\theta}\|\|\boldsymbol{\delta}\|} = 0$, and thus on the right hand side of the inequality, the first term dominates the second term for large n.

Similarly, we apply Lemma A.4 to bound the singular vectors corresponding to the 2nd to Kth largest singular values, and have

$$\begin{aligned}
\|\hat{\mathbf{V}}_{2\sim K} - \mathbf{V}_{2\sim K}\mathbf{O}_{\mathbf{V}}\|_F &\leqslant \frac{C\|\mathbf{X}^T\mathbf{X} - \boldsymbol{\Omega}^T\boldsymbol{\Omega}\|}{\min(\lambda_1(\boldsymbol{\Omega}^T\boldsymbol{\Omega}) - \lambda_2(\boldsymbol{\Omega}^T\boldsymbol{\Omega}), \lambda_K(\boldsymbol{\Omega}^T\boldsymbol{\Omega}) - \lambda_{(K+1)}(\boldsymbol{\Omega}^T\boldsymbol{\Omega}))} \\
&\overset{\text{(i)}}{\leqslant} \frac{C\|\mathbf{X}^T\mathbf{X} - \boldsymbol{\Omega}^T\boldsymbol{\Omega}\|}{\min(\lambda_1(\boldsymbol{\Omega}^T\boldsymbol{\Omega}) - \lambda_2(\boldsymbol{\Omega}^T\boldsymbol{\Omega}), \lambda_K(\boldsymbol{\Omega}^T\boldsymbol{\Omega}))} \\
&\overset{\text{(ii)}}{\leqslant} \frac{C_1\sqrt{\log(n)Z}\|\boldsymbol{\theta}\|\|\boldsymbol{\delta}\| + C_2\log(n)Z}{C\|\boldsymbol{\theta}\|^2\|\boldsymbol{\delta}\|^2} \\
&\leqslant C_1\frac{\sqrt{\log(n)Z}}{\|\boldsymbol{\theta}\|\|\boldsymbol{\delta}\|} + C_2\left(\frac{\sqrt{\log(n)Z}}{\|\boldsymbol{\theta}\|\|\boldsymbol{\delta}\|}\right)^2 \\
&\overset{\text{(iii)}}{\leqslant} C\frac{\sqrt{\log(n)Z}}{\|\boldsymbol{\theta}\|\|\boldsymbol{\delta}\|},
\end{aligned}$$

where (i) follows from Proposition 1, which implies $\lambda_{(K+1)}(\boldsymbol{\Omega}^T\boldsymbol{\Omega}) = 0$, (ii) follows from Lemma A.1, eqs. (A.14) and (A.15), and (iii) follows from eq. (2.11), and as we argued above, the first term dominates the second term for large n.

Following the proof procedure similar to the above arguments, we can obtain that $\|\hat{\mathbf{U}}_1 - \mathbf{U}_1 C_U\| \leqslant C\frac{\sqrt{\log(n)Z}}{\|\boldsymbol{\theta}\|\|\boldsymbol{\delta}\|}$ and $\|\hat{\mathbf{U}}_{2\sim K} - \mathbf{U}_{2\sim K}\mathbf{O}_{\mathbf{U}}\|_F \leqslant C\frac{\sqrt{\log(n)Z}}{\|\boldsymbol{\theta}\|\|\boldsymbol{\delta}\|}$. ■

A.3. Proof of Proposition 3

Proof In the following, we deal with row vectors, and the row ℓ_2-norm. Take two nodes i and j from the graph. Then, by definition, we have

$$\|\mathbf{R}_{\bar{i}} - \mathbf{R}_{\bar{j}}\|^2 = \|(\mathbf{R}_{\mathbf{V}})_{\bar{i}} - (\mathbf{R}_{\mathbf{V}})_{\bar{j}}\|^2 + \|(\mathbf{R}_{\mathbf{U}})_{\bar{i}} - (\mathbf{R}_{\mathbf{U}})_{\bar{j}}\|^2.$$

Thus, to prove Proposition 3, it is sufficient to show $\|(\mathbf{R}_{\mathbf{V}})_{\bar{i}} - (\mathbf{R}_{\mathbf{V}})_{\bar{j}}\|^2 \geqslant 2$ and $\|(\mathbf{R}_{\mathbf{U}})_{\bar{i}} - (\mathbf{R}_{\mathbf{U}})_{\bar{j}}\|^2 \geqslant 2$ for $c_i \neq c_j$, , and $\|(\mathbf{R}_{\mathbf{V}})_{\bar{i}} - (\mathbf{R}_{\mathbf{V}})_{\bar{j}}\|^2 = 0$ and $\|(\mathbf{R}_{\mathbf{U}})_{\bar{i}} - (\mathbf{R}_{\mathbf{U}})_{\bar{j}}\|^2 = 0$ for $c_i = c_j$.

We first show that these hold for $\|(\mathbf{R}_{\mathbf{V}})_{\bar{i}} - (\mathbf{R}_{\mathbf{V}})_{\bar{j}}\|^2$. We derive the follow equations.

$$\begin{aligned}
\|(\mathbf{R}_{\mathbf{V}})_{\bar{i}} - (\mathbf{R}_{\mathbf{V}})_{\bar{j}}\|^2 &\overset{\text{(i)}}{=} \left\|\frac{(\mathbf{V}_{2\sim K}\mathbf{O}_{\mathbf{V}})_{\bar{i}}}{C_V\mathbf{V}_1(i)} - \frac{(\mathbf{V}_{2\sim K}\mathbf{O}_{\mathbf{V}})_{\bar{j}}}{C_V\mathbf{V}_1(j)}\right\|^2 \\
&\overset{\text{(ii)}}{=} \left\|\frac{(\mathbf{V}_{2\sim K})_{\bar{i}}}{\mathbf{V}_1(i)} - \frac{(\mathbf{V}_{2\sim K})_{\bar{j}}}{\mathbf{V}_1(j)}\right\|^2 \\
&\overset{\text{(iii)}}{=} \left\|\frac{(\mathbf{V}_{2\sim K})_{\bar{i}}}{\mathbf{V}_1(i)} - \frac{(\mathbf{V}_{2\sim K})_{\bar{j}}}{\mathbf{V}_1(j)}\right\|^2 + \left\|\frac{\mathbf{V}_1(i)}{\mathbf{V}_1(i)} - \frac{\mathbf{V}_1(j)}{\mathbf{V}_1(j)}\right\|^2 \\
&= \left\|\frac{\mathbf{V}_{\bar{i}}}{\mathbf{V}_1(i)} - \frac{\mathbf{V}_{\bar{j}}}{\mathbf{V}_1(j)}\right\|^2
\end{aligned}$$

$$\overset{\text{(iv)}}{=} \left\| \frac{\frac{\boldsymbol{\delta}(i)}{\|\boldsymbol{\delta}^{(c_i)}\|}\mathbf{H}_{\bar{c}_i}}{\frac{\boldsymbol{\delta}(i)}{\|\boldsymbol{\delta}^{(c_i)}\|}\mathbf{H}_{\bar{c}_i}(1)} - \frac{\frac{\boldsymbol{\delta}(j)}{\|\boldsymbol{\delta}^{(c_j)}\|}\mathbf{H}_{\bar{c}_i}}{\frac{\boldsymbol{\delta}(j)}{\|\boldsymbol{\delta}^{(c_j)}\|}\mathbf{H}_{\bar{c}_j}(1)} \right\|^2$$

$$= \left\| \frac{\mathbf{H}_{\bar{c}_i}}{\mathbf{H}_{\bar{c}_i}(1)} - \frac{\mathbf{H}_{\bar{c}_j}}{\mathbf{H}_{\bar{c}_j}(1)} \right\|^2, \tag{A.16}$$

where (i) follows from the definition of $(R_V)_{\bar{i}}$ (see eq. (4.7)), (ii) follows from Proposition 2, which gives $|C_V| = |C_U| = 1$, and $\mathbf{O_V}$ and O_U are orthogonal matrices, (iii) follows because the second term equals 0, and (iv) follows from eq. (4.3), which shows $\mathbf{V}_{\bar{i}} = \frac{\boldsymbol{\delta}(i)}{\|\boldsymbol{\delta}^{(c_i)}\|}\mathbf{H}_{\bar{c}_i}$.

Thus, eq. (A.16) implies that $\|(\mathbf{R_V})_{\bar{i}} - (\mathbf{R_V})_{\bar{j}}\|^2 = 0$ if $c_i = c_j$. Otherwise, if $c_i \neq c_j$, we have

$$\begin{aligned}
\|(\mathbf{R_V})_{\bar{i}} - (\mathbf{R_V})_{\bar{j}}\|^2 &\overset{\text{(i)}}{=} \left\| \frac{\mathbf{H}_{\bar{c}_i}}{\mathbf{H}_{\bar{c}_i}(1)} - \frac{\mathbf{H}_{\bar{c}_j}}{\mathbf{H}_{\bar{c}_j}(1)} \right\|^2 \\
&= \left\| \frac{\mathbf{H}_{\bar{c}_i}}{\mathbf{H}_{\bar{c}_i}(1)} \right\|^2 + \left\| \frac{\mathbf{H}_{\bar{c}_j}}{\mathbf{H}_{\bar{c}_j}(1)} \right\|^2 - 2\left\langle \frac{\mathbf{H}_{\bar{c}_i}}{\mathbf{H}_{\bar{c}_i}(1)}, \frac{\mathbf{H}_{\bar{c}_j}}{\mathbf{H}_{\bar{c}_j}(1)} \right\rangle \\
&\overset{\text{(ii)}}{=} \frac{1}{|\mathbf{H}_{\bar{c}_i}(1)|} + \frac{1}{|\mathbf{H}_{\bar{c}_j}(1)|} \\
&\overset{\text{(iii)}}{\geqslant} 1 + 1 \\
&= 2,
\end{aligned} \tag{A.17}$$

where (i) follows from eq. (A.16), (ii) follows from eq. (4.3), which shows that $\mathbf{H}$ is an orthogonal matrix, and thus $\|\mathbf{H}_{c_i}\| = 1$, and the rows of $\mathbf{H}$ are also orthogonal to each other, i.e., $\langle \mathbf{H}_{\bar{i}}, \mathbf{H}_{\bar{j}} \rangle = 0$, for $i \neq j$, (iii) follows from eq. (A.10), which shows $\mathbf{H}_{\bar{c}_i}(1) \leqslant 1$.

The inequality $\|(\mathbf{R_U})_{\bar{i}} - (\mathbf{R_U})_{\bar{j}}\|^2 \geqslant 2$ for $c_i \neq c_j$, and otherwise equals 0 can be shown in similar way, which completes the proof of Proposition 3. ■

A.4. Proof of Proposition 4

To prove Proposition 4, we first establish Lemma A.5 to bound the number of ill-behavior nodes and a technical inequality in Lemma A.6.

First, for a constant $0 < C < 1$, we define

$$\hat{S}_V \equiv \left(1 \leqslant i \leqslant n; \left| \frac{\hat{\mathbf{V}}_1(i)}{C_V \mathbf{V}_1(i)} - 1 \right| \leqslant C \right) \text{ and } \hat{S}_U \equiv \left(1 \leqslant i \leqslant n; \left| \frac{\hat{\mathbf{U}}_1(i)}{C_U \mathbf{U}_1(i)} - 1 \right| \leqslant C \right). \tag{A.18}$$

Then, we bound the number of nodes that outside $\hat{S}_V$ and $\hat{S}_U$ in Lemma A.5.

Lemma A.5 *For nodes in $\hat{S}_V$ or $\hat{S}_U$, the following equations hold*

$$\left| \hat{\mathbf{V}}_1(i) \right| \asymp \left| C_V \mathbf{V}_1(i) \right| \asymp \frac{\boldsymbol{\delta}(i)}{\|\boldsymbol{\delta}\|} \quad \textit{for} \quad i \in \hat{S}_V, \tag{A.19}$$

$$\left|\hat{\mathbf{U}}_1(i)\right| \asymp \left|C_U \mathbf{U}_1(i)\right| \asymp \frac{\boldsymbol{\theta}(i)}{\|\boldsymbol{\theta}\|} \quad \textit{for} \quad i \in \hat{S}_U. \tag{A.20}$$

Furthermore, with probability at least $1-O(n^{-4})$*, the cardinality of* $\mathcal{V}\backslash\hat{S}_V$ *and* $\mathcal{V}\backslash\hat{S}_U$ *satisfy*

$$\left|\mathcal{V}\backslash\hat{S}_V\right| \leqslant \frac{C\log(n)Z}{\|\boldsymbol{\theta}\|^2\delta_{\min}^2} \quad \textit{and} \quad \left|\mathcal{V}\backslash\hat{S}_U\right| \leqslant \frac{C\log(n)Z}{\|\boldsymbol{\delta}\|^2\theta_{\min}^2}. \tag{A.21}$$

Proof The proof can be found in Appendix A.7.4. ■

Then, we provide a technical inequality in Lemma A.6.

Lemma A.6 *For* $\mathbf{v}, \mathbf{u} \in \mathbf{R}^n$*,* $a, b \in \mathbf{R}, a > 0, b > 0$*, the following inequality holds,*

$$\left\|\frac{\mathbf{v}}{a} - \frac{\mathbf{u}}{b}\right\|^2 \leqslant 2\left(\frac{1}{a^2}\|\mathbf{v}-\mathbf{u}\|^2 + \frac{(b-a)^2}{(ab)^2}\|\mathbf{u}\|^2\right).$$

Proof The proof can be found in Appendix A.7.5. ■

Now we are ready to proof the proposition.

Proof [Proof of Proposition 4] Note that

$$\|\mathbf{R}^* - \mathbf{R}\|_F^2 = \|\mathbf{R}^*_{\hat{\mathbf{V}}} - \mathbf{R}_{\mathbf{V}}\|_F^2 + \|\mathbf{R}^*_{\hat{\mathbf{U}}} - \mathbf{R}_{\mathbf{U}}\|_F^2.$$

It is sufficient to prove $\|\mathbf{R}^*_{\hat{\mathbf{V}}} - \mathbf{R}_{\mathbf{V}}\|_F^2 \leqslant C\frac{T_n^2\log(n)Z}{\boldsymbol{\delta}^2_{min}\|\boldsymbol{\theta}\|^2}$ and $\|R^*_{\hat{U}} - R_U\|_F^2 \leqslant C\frac{T_n^2\log(n)Z}{\boldsymbol{\theta}^2_{min}\|\boldsymbol{\delta}\|^2}$, and then combining these two inequalities, we establish the proposition. We first prove $\|\mathbf{R}^*_{\hat{\mathbf{V}}} - \mathbf{R}_{\mathbf{V}}\|_F^2 \leqslant C\frac{T_n^2\log(n)Z}{\boldsymbol{\delta}^2_{min}\|\boldsymbol{\theta}\|^2}$, and the latter one can be shown similarly.

First we show $\|(\mathbf{V}_{2\sim K}\mathbf{O}_{\mathbf{V}})_{\bar{i}}\|^2 \leqslant C\frac{\boldsymbol{\delta}^2(i)}{\|\boldsymbol{\delta}\|^2}$. Note that

$$\|(\mathbf{V}_{2\sim K}\mathbf{O}_{\mathbf{V}})_{\bar{i}}\|^2 = \|(\mathbf{V}_{2\sim K})_{\bar{i}}\mathbf{O}_{\mathbf{V}}\|^2 = \|(\mathbf{V}_{2\sim K})_{\bar{i}}\|^2 \leqslant \|\mathbf{V}_{\bar{i}}\|^2 \overset{\text{(i)}}{\leqslant} C\frac{\boldsymbol{\delta}^2(i)}{\|\boldsymbol{\delta}\|^2}, \tag{A.22}$$

where (i) follows from eq. (4.6).

Next we prove $\|(\mathbf{R}_{\mathbf{V}})_{\bar{i}}\|^2 \leqslant C$. By definition of $\mathbf{R}_{\mathbf{V}}$, we have

$$\begin{aligned}\|(\mathbf{R}_{\mathbf{V}})_{\bar{i}}\|^2 = \left\|\frac{(\mathbf{V}_{2\sim K}\mathbf{O}_{\mathbf{V}})_{\bar{i}}}{C_V\mathbf{V}_1(i)}\right\|^2 &\overset{\text{(i)}}{\leqslant} \frac{C\frac{\boldsymbol{\delta}(i)^2}{\|\boldsymbol{\delta}\|^2}}{|C_V\mathbf{V}_1(i)|^2} \\ &\overset{\text{(ii)}}{\leqslant} \frac{C\frac{\boldsymbol{\delta}(i)^2}{\|\boldsymbol{\delta}\|^2}}{\frac{\boldsymbol{\delta}(i)^2}{\|\boldsymbol{\delta}^{(c_i)}\|^2}|H_1(c_i)|^2} \overset{\text{(iii)}}{\leqslant} C,\end{aligned} \tag{A.23}$$

where (i) follows from eq. (A.22), (ii) follows from eq. (4.3) which implies $V_1(i) = \frac{\boldsymbol{\delta}(i)}{\|\boldsymbol{\delta}^{(c_i)}\|}\mathbf{H}_1(c_i)$, and (iii) follows from eq. (2.9) and Lemma A.3, which implies $\mathbf{H}_1(c_i) \geqslant C > 0$.

In order to prove $\|\mathbf{R}_{\hat{\mathbf{V}}} - \mathbf{R}_{\mathbf{V}}\|_F^2 \leqslant C\frac{T_n^2\log(n)Z}{\boldsymbol{\delta}^2_{min}\|\boldsymbol{\theta}\|^2}$, we divide the sum into following two parts:

$$\|\mathbf{R}_{\hat{\mathbf{V}}} - \mathbf{R}_{\mathbf{V}}\|_F^2 = \sum_{i\in(\mathcal{V}\backslash\hat{S}_V)}\|(\mathbf{R}_{\hat{\mathbf{V}}})_{\bar{i}} - (\mathbf{R}_{\mathbf{V}})_{\bar{i}}\|^2 + \sum_{i\in\hat{S}_V}\|(\mathbf{R}_{\hat{\mathbf{V}}})_{\bar{i}} - (\mathbf{R}_{\mathbf{V}})_{\bar{i}}\|^2. \tag{A.24}$$

For the first term, we have

$$\begin{aligned}\sum_{i\in(\mathcal{V}\setminus\hat{S}_V)}\|(\mathbf{R}_{\hat{\mathbf{V}}})_{\bar{i}}-(\mathbf{R}_{\mathbf{V}})_{\bar{i}}\|^2 &\leqslant C\sum_{i\in(\mathcal{V}\setminus\hat{S}_V)}(\|(\mathbf{R}_{\hat{\mathbf{V}}})_{\bar{i}}\|^2+\|(\mathbf{R}_{\mathbf{V}})_{\bar{i}}\|^2)\\ &\overset{\text{(i)}}{\leqslant} C\sum_{i\in(\mathcal{V}\setminus\hat{S}_V)}(KT_n^2+C)\\ &\overset{\text{(ii)}}{\leqslant} C|\mathcal{V}\setminus\hat{S}_V|T_n^2\\ &\overset{\text{(iii)}}{\leqslant}\frac{CT_n^2\log(n)Z}{\|\boldsymbol{\theta}\|^2\delta_{\min}^2},\end{aligned}\tag{A.25}$$

where (i) follows from eqs. (3.1) and (A.23), (ii) follows from the fact that T_n scales with n, and thus T_n dominates C for sufficient large n, and (iii) follows from Lemma A.5.

For the second term in eq. (A.24), we have

$$\begin{aligned}&\sum_{i\in\hat{S}_V}\left\|(\mathbf{R}_{\hat{\mathbf{V}}})_{\bar{i}}-(\mathbf{R}_{\mathbf{V}})_{\bar{i}}\right\|^2\\ &\overset{\text{(i)}}{\leqslant} C\sum_{i\in\hat{S}_V}\left\|\frac{(\hat{\mathbf{V}}_{2\sim K})_{\bar{i}}}{\hat{\mathbf{V}}_1(i)}-\frac{(\mathbf{V}_{2\sim K}\mathbf{O}_{\mathbf{V}})_{\bar{i}}}{C_V\mathbf{V}_1(i)}\right\|^2\\ &\overset{\text{(ii)}}{\leqslant} C\sum_{i\in\hat{S}_V}\left(\frac{1}{(\hat{\mathbf{V}}_1(i))^2}\|(\hat{\mathbf{V}}_{2\sim K})_{\bar{i}}-(\mathbf{V}_{2\sim K}\mathbf{O}_{\mathbf{V}})_{\bar{i}}\|^2+\frac{(C_V\mathbf{V}_1(i)-\hat{\mathbf{V}}_1(i))^2}{(\hat{\mathbf{V}}_1(i)C_V\mathbf{V}_1(i))^2}\|(\mathbf{V}_{2\sim K}\mathbf{O}_{\mathbf{V}})_{\bar{i}}\|^2\right)\\ &\overset{\text{(iii)}}{\leqslant} C\sum_{i\in\hat{S}_V}\left(\frac{\|\boldsymbol{\delta}\|^2}{\boldsymbol{\delta}(i)^2}\|(\hat{\mathbf{V}}_{2\sim K})_{\bar{i}}-(\mathbf{V}_{2\sim K}\mathbf{O}_{\mathbf{V}})_{\bar{i}}\|^2+\frac{\|\boldsymbol{\delta}\|^2}{\boldsymbol{\delta}(i)^2}(C_V\mathbf{V}_1(i)-\hat{\mathbf{V}}_1(i))^2\right)\\ &\leqslant C\frac{\|\boldsymbol{\delta}\|^2}{\delta_{\min}^2}\sum_{i\in\hat{S}_V}\left(\|(\hat{\mathbf{V}}_{2\sim K})_{\bar{i}}-(\mathbf{V}_{2\sim K}\mathbf{O}_{\mathbf{V}})_{\bar{i}}\|^2+\sum_{i\in\hat{S}_V}(C_V\mathbf{V}_1(i)-\hat{\mathbf{V}}_1(i))^2\right)\\ &\leqslant C\frac{\|\boldsymbol{\delta}\|^2}{\delta_{\min}^2}\left(\|\hat{\mathbf{V}}_{2\sim K}-\mathbf{V}_{2\sim K}\mathbf{O}_{\mathbf{V}}\|_F^2+\|\hat{\mathbf{V}}_1-\mathbf{V}_1C_V\|^2\right)\\ &\overset{\text{(iv)}}{\leqslant}\frac{C\log(n)Z}{\|\boldsymbol{\theta}\|^2\delta_{\min}^2},\end{aligned}\tag{A.26}$$

where (i) follows from the fact that $|(\mathbf{R}_{\mathbf{V}})_{\bar{i}}|\leqslant C$ (see eq. (A.23)), and T_n scales with n, which implies $T_n\geqslant C\geqslant|(\mathbf{R}_{\mathbf{V}})_{\bar{i}}|$ for n large enough. Thus, although eq. (3.1) shows that $(\mathbf{R}_{\hat{\mathbf{V}}})_{\bar{i}}$ is truncated by T_n, we still have $\|(\mathbf{R}_{\hat{\mathbf{V}}})_{\bar{i}}-(\mathbf{R}_{\mathbf{V}})_{\bar{i}}\|^2\leqslant\left\|\frac{(\hat{\mathbf{V}}_{2\sim K})_{\bar{i}}}{\hat{\mathbf{V}}_1(i)}-\frac{(\mathbf{V}_{2\sim K}\mathbf{O}_{\mathbf{V}})_{\bar{i}}}{C_V\mathbf{V}_1(i)}\right\|^2$ for large n, (ii) follows from Lemma A.6, (iii) follows from Lemma A.5 and eq. (A.22), and (iv) follows from Proposition 2.

Combining eqs. (A.25) and (A.26), we obtain $\|\mathbf{R}_{\hat{\mathbf{V}}}-\mathbf{R}_{\mathbf{V}}\|_F^2\leqslant C\frac{T_n^2\log(n)Z}{\boldsymbol{\delta}_{min}^2\|\boldsymbol{\theta}\|^2}$. ■

A.5. Proof of Proposition 5

Proof Recall that $\mathbf{M}^*$ is defined as

$$\mathbf{M}^* = \underset{\mathbf{M}\in \boldsymbol{M}_{n,2K-2,K}}{\operatorname{argmin}} \left\|\mathbf{M} - \hat{\mathbf{R}}\right\|_F^2,$$

where $\boldsymbol{M}_{n,2K-2,K}$ denotes the set of $n\times(2K-2)$ matrices with only K different rows. Note that R is also in $\boldsymbol{M}_{n,2K-2,K}$. Thus,

$$\|\mathbf{M}^* - \hat{\mathbf{R}}\| \leqslant \|\mathbf{R} - \hat{\mathbf{R}}\|. \tag{A.27}$$

Then, we obtain

$$\begin{aligned}
\|\mathbf{M}^* - R\|_F^2 &\leqslant \|\mathbf{M}^* - \hat{\mathbf{R}} + \hat{\mathbf{R}} - \mathbf{R}\|_F^2 \\
&\leqslant C\|\mathbf{M}^* - \hat{\mathbf{R}}\|_F^2 + C\|\hat{\mathbf{R}} - \mathbf{R}\|_F^2 \\
&\overset{\text{(i)}}{\leqslant} C\|\mathbf{R} - \hat{\mathbf{R}}\|_F^2 + C\|\hat{\mathbf{R}} - \mathbf{R}\|_F^2 \\
&\leqslant C\|\mathbf{R} - \hat{\mathbf{R}}\|_F^2 \\
&\overset{\text{(ii)}}{\leqslant} CT_n^2 \log(n) err_n,
\end{aligned}$$

where (i) follows from eq. (A.27), and (ii) follows from Proposition 4.

■

A.6. Proof of Theorem 1

Proof First, if nodes i, j in set $\mathcal{W}$ are in different communities, then

$$\begin{aligned}
\|\mathbf{M}^*_{\bar{i}} - \mathbf{M}^*_{\bar{j}}\| &= \|\mathbf{M}^*_{\bar{i}} - \mathbf{R}_{\bar{i}} + \mathbf{R}_{\bar{i}} - \mathbf{R}_{\bar{j}} + \mathbf{R}_{\bar{j}} - \mathbf{M}^*_{\bar{j}}\| \\
&\geqslant \|\mathbf{R}_{\bar{i}} - \mathbf{R}_{\bar{j}}\| - \|\mathbf{M}^*_{\bar{i}} - \mathbf{R}_{\bar{i}} + \mathbf{R}_{\bar{j}} - \mathbf{M}^*_{\bar{j}}\| \\
&\geqslant \|\mathbf{R}_{\bar{i}} - \mathbf{R}_{\bar{j}}\| - \|\mathbf{M}^*_{\bar{i}} - \mathbf{R}_{\bar{i}}\| - \|\mathbf{M}^*_{\bar{j}} - \mathbf{R}_{\bar{j}}\|.
\end{aligned}$$

By Proposition 3, i.e., $\|\mathbf{R}_{\bar{i}} - \mathbf{R}_{\bar{j}}\| \geqslant 2$, and the definition of set $\mathcal{W}$ in Theorem 1. We obtain that for $i, j \in \mathcal{W}$,

$$\|\mathbf{M}^*_{\bar{i}} - \mathbf{M}^*_{\bar{j}}\| \geqslant (2-1) = 1.$$

Thus, if nodes i, j are in different communities, then their corresponding rows in $\mathbf{M}^*$ are sufficiently different. By the assumption $|\mathcal{V}\backslash\mathcal{W}| < \min\{n_1, n_2 \cdots n_K\}$, $\mathcal{W}$ contains at least one node in each community. Combining there two facts and the definition that $\mathbf{M}^*$ has only K different rows, we conclude that the corresponding rows in $\mathbf{M}^*$ of nodes in the same community are same. In conclusion, if two nodes in $\mathcal{W}$ are in the same community, then their corresponding rows in $\mathbf{M}^*$ are the same. Otherwise, their corresponding rows in $\mathbf{M}^*$ are sufficiently different. Thus, nodes in $\mathcal{W}$ are correctly clustered. Then, the definition of $\mathcal{W}$ and Proposition 5 directly imply

$$|\mathcal{V}\backslash\mathcal{W}| \leqslant CT_n^2 \log(n) err_n. \tag{A.28}$$

■

A.7. Proof of Lemmas for D-SCORE

A.7.1. Proof of Lemma A.1

Proof Following eq. (4.2), we obtain that $\sigma_i(\mathbf{\Omega}) = \|\boldsymbol{\theta}\|\|\boldsymbol{\delta}\|\sigma_i(\mathbf{S})$, for $1 \leqslant i \leqslant K$, it is sufficient to show $0 < C_1 \leqslant \sigma_K(\mathbf{S}) \leqslant \sigma_1(\mathbf{S}) \leqslant C_2$.

Recall $\mathbf{S} = \mathbf{\Psi}_{\boldsymbol{\theta}}\mathbf{B}\mathbf{\Psi}_{\boldsymbol{\delta}}^T$, where $\mathbf{\Psi}_{\boldsymbol{\theta}}$ and $\mathbf{\Psi}_{\boldsymbol{\delta}}$ are diagonal matrices, and $\|\mathbf{\Psi}_{\boldsymbol{\theta}}\|$ and thus $\|\mathbf{\Psi}_{\boldsymbol{\theta}}\|_{\min}$ correspond to the largest and smallest absolute value of the diagonal entries of $\mathbf{\Psi}_{\boldsymbol{\theta}}$, respectively. Following from eq. (2.9), we have

$$\|\mathbf{\Psi}_{\boldsymbol{\theta}}\| = \max_i \mathbf{\Psi}_{\boldsymbol{\theta}}(i,i) = \max_i \frac{\|\boldsymbol{\theta}^{(i)}\|}{\|\boldsymbol{\theta}\|} \leqslant C,$$
$$\|\mathbf{\Psi}_{\boldsymbol{\theta}}\|_{\min} = \min_i \mathbf{\Psi}_{\boldsymbol{\theta}}(i,i) = \min_i \frac{\|\boldsymbol{\theta}^{(i)}\|}{\|\boldsymbol{\theta}\|} \geqslant C.$$

Therefore, there exist two constants $C_m > 0$ and $C_M > 0$ such that

$$C_m \leqslant \|\mathbf{\Psi}_{\boldsymbol{\theta}}\|_{\min} \leqslant \|\mathbf{\Psi}_{\boldsymbol{\theta}}\| \leqslant C_M. \tag{A.29}$$

It can be similarly shown that

$$C_m \leqslant \|\mathbf{\Psi}_{\boldsymbol{\delta}}\|_{\min} \leqslant \|\mathbf{\Psi}_{\boldsymbol{\delta}}\| \leqslant C_M.$$

By the definition of $\mathbf{S} \equiv \mathbf{\Psi}_{\boldsymbol{\theta}}\mathbf{B}\mathbf{\Psi}_{\boldsymbol{\delta}}^T$, we have

$$\sigma_1(\mathbf{S}) = \|\mathbf{S}\| = \|\mathbf{\Psi}_{\boldsymbol{\theta}}\mathbf{B}\mathbf{\Psi}_{\boldsymbol{\delta}}^T\| \leqslant \|\mathbf{\Psi}_{\boldsymbol{\theta}}\|\|\mathbf{B}\|\|\mathbf{\Psi}_{\boldsymbol{\delta}}^T\| \overset{\text{(i)}}{\leqslant} C, \tag{A.30}$$

where (i) follows from eq. (A.29) and because $\mathbf{B}$ is a constant matrix, i.e., $\mathbf{B}$ does change with n, so that there exists a constant C, such that $\|\mathbf{B}\| \leqslant C$. On the other hand,

$$\sigma_K(\mathbf{S}) = \|\mathbf{S}\|_{\min} = \|\mathbf{\Psi}_{\boldsymbol{\theta}}\mathbf{B}\mathbf{\Psi}_{\boldsymbol{\delta}}^T\|_{\min} \geqslant \|\mathbf{\Psi}_{\boldsymbol{\theta}}\|_{\min}\|\mathbf{B}\|_{\min}\|\mathbf{\Psi}_{\boldsymbol{\delta}}^T\|_{\min} \overset{\text{(i)}}{\geqslant} C, \tag{A.31}$$

where (i) follows from eq. (A.29) and the inequality $\|\mathbf{AB}\|_{\min} \geqslant \|\mathbf{A}\|_{\min}\|\mathbf{B}\|_{\min}$. Also, since $\mathbf{B}$ is a constant matrix which does change with n, and B is non-singular (see eq. (2.4)), there exists a constant C, such that $\|\mathbf{B}\|_{\min} \geqslant C > 0$.

Combining eqs. (A.30) and (A.31), we obtain $\sigma_i(\mathbf{S}) \asymp C$. Then, by eq. (4.2), we have $\sigma_i^2(\mathbf{\Omega}) = \sigma_i^2(\mathbf{S})\|\boldsymbol{\theta}\|^2\|\boldsymbol{\delta}\|^2$. Hence, for $1 \leqslant i \leqslant K$

$$\lambda_i(\mathbf{\Omega}^T\mathbf{\Omega}) = \sigma_i^2(\mathbf{\Omega}) = \sigma_i^2(\mathbf{S})\|\boldsymbol{\theta}\|^2\|\boldsymbol{\delta}\|^2 \asymp \|\boldsymbol{\theta}\|^2\|\boldsymbol{\delta}\|^2.$$

■

A.7.2. Proof of Lemma A.2

Proof Define $\mathbf{e}_i$ as an $n \times 1$ vector, where $\mathbf{e}_i(i) = 1$ and 0 elsewhere. Thus, we can write $\mathbf{W}$ as $\mathbf{W} = \sum_{i,j=1}^n \mathbf{W}(i,j)\mathbf{e}_i\mathbf{e}_j^T$. By the definition that $\mathbf{W} \equiv \mathbf{A} - \mathbf{\Omega} = \mathbf{A} - E[\mathbf{A}]$, the entry $\mathbf{W}(i,j)$ is an independent centered Bernoulli random variable. Thus $\mathbf{W}(i,j)\mathbf{e}_i\mathbf{e}_j^T$

is an independent centered Bernoulli random matrix with the dimension $n \times n$. In fact, $\mathbf{W}(i,j)\mathbf{e}_i\mathbf{e}_j^T$ is a matrix with only one nonzero entry $\mathbf{W}(i,j) = \mathbf{A}(i,j) - \boldsymbol{\theta}(i)\mathbf{B}(c_i,c_j)\boldsymbol{\delta}(j)$ at the location (i,j).

In order to apply matrix Bernstein inequality, we need to bound the spectral norm of each summation matrix, and the variance of the entire summation. By the definition of the matrix spectral norm, for $1 \leqslant i,j \leqslant n$, we have

$$\begin{aligned}
\|\mathbf{W}(i,j)\mathbf{e}_i\mathbf{e}_j^T\| &= \big|\mathbf{W}(i,j)\big|\,\|\mathbf{e}_i\mathbf{e}_j^T\| = \big|\mathbf{W}(i,j)\big|\sqrt{\|\mathbf{e}_i\mathbf{e}_j^T(\mathbf{e}_i\mathbf{e}_j^T)^T\|} \\
&= \big|\mathbf{W}(i,j)\big|\sqrt{\|\mathbf{e}_i\mathbf{e}_i^T\|} \overset{\text{(i)}}{=} \big|\mathbf{W}(i,j)\big| \\
&\overset{\text{(ii)}}{=} \big|\mathbf{A}(i,j) - \boldsymbol{\Omega}(i,j)\big| \\
&\overset{\text{(iii)}}{\leqslant} \max\big(\,\big|0 - \boldsymbol{\theta}(i)\mathbf{B}(c_i,c_j)\boldsymbol{\delta}(j)\big|\,,\big|1 - \boldsymbol{\theta}(i)\mathbf{B}(c_i,c_j)\boldsymbol{\delta}(j)\big|\big) \\
&\overset{\text{(iv)}}{\leqslant} 1, \qquad \text{(A.32)}
\end{aligned}$$

where (i) follows because $\mathbf{e}_i\mathbf{e}_i^T$ is a diagonal matrix with only one non-zero entry 1 at location (i,i), thus $\|\mathbf{e}_i\mathbf{e}_i^T\| = 1$, (ii) follows because that $\mathbf{W} = \mathbf{A} - \boldsymbol{\Omega}$, (iii) follows because $\mathbf{A}(i,j)$ is a Bernoulli random variable that it takes the values 0 or 1, and (iv) follows because $0 < \boldsymbol{\Omega}(i,j) = \boldsymbol{\theta}(i)\mathbf{B}(c_i,c_j)\boldsymbol{\delta}(j) \leqslant 1$.

Next we consider the variance of the random matrix $V(\mathbf{W}) \equiv \max\Big(\|E(\mathbf{W}\mathbf{W}^T)\|, \|E(\mathbf{W}^T\mathbf{W})\|\Big)$. We first bound $\|E(\mathbf{W}\mathbf{W}^T)\|$, and then bound $\|E(\mathbf{W}^T\mathbf{W})\|$. Note that

$$\begin{aligned}
E(\mathbf{W}\mathbf{W}^T) &= E[(\sum_{i,j=1}^{n}\mathbf{W}(i,j)\mathbf{e}_i\mathbf{e}_j^T)(\sum_{k,l=1}^{n}\mathbf{W}(k,l)\mathbf{e}_k\mathbf{e}_l^T)^T] \\
&= E[\sum_{i,j,k,l=1}^{n}\mathbf{W}(i,j)\mathbf{W}(k,l)\mathbf{e}_i\mathbf{e}_j^T\mathbf{e}_l\mathbf{e}_k^T] \\
&\overset{\text{(i)}}{=} \sum_{i,j,k=1}^{n}E[\mathbf{W}(i,j)\mathbf{W}(k,j)\mathbf{e}_i\mathbf{e}_k^T] \\
&\overset{\text{(ii)}}{=} \sum_{i,j=1}^{n}E[\mathbf{W}^2(i,j)]\mathbf{e}_i\mathbf{e}_i^T, \qquad \text{(A.33)}
\end{aligned}$$

where (i) follows from the fact that $\mathbf{e}_j^T\mathbf{e}_l = 1$ if $j = l$ and 0 otherwise, and (ii) follows from the fact that if $i \neq k$, $\mathbf{W}(i,j)$ and $\mathbf{W}(k,j)$ are independent random Bernoulli random variables with the expected value 0, i.e., $E[\mathbf{W}(i,j)\mathbf{W}(k,j)\mathbf{e}_i\mathbf{e}_k^T)] = E[\mathbf{W}(i,j)]E[\mathbf{W}(k,j)]\mathbf{e}_i\mathbf{e}_k^T = 0 \times 0 = 0$. Thus, we only need to consider the case with $i = k$. Observing that $\boldsymbol{\Omega}(i,j) = E[\mathbf{X}(i,j)]$ and let $Var(\mathbf{X}(i,j))$ denote the variance of Bernoulli random variable $\mathbf{X}(i,j)$. Then, we obtain

$$\begin{aligned}
E[\mathbf{W}^2(i,j)] &= E[(\mathbf{X}(i,j) - \boldsymbol{\Omega}(i,j))^2] \\
&= E[(\mathbf{X}(i,j) - E[\mathbf{X}(i,j)])^2] \\
&= \mathrm{Var}(\mathbf{X}(i,j))
\end{aligned}$$

$$
\begin{aligned}
&= \boldsymbol{\theta}(i)\mathbf{B}(c_i, c_j)\boldsymbol{\delta}(j)[1 - \boldsymbol{\theta}(i)\mathbf{B}(c_i, c_j)\boldsymbol{\delta}(j)] \\
&\leqslant \boldsymbol{\theta}(i)\mathbf{B}(c_i, c_j)\boldsymbol{\delta}(j) \\
&\leqslant \boldsymbol{\theta}(i)\boldsymbol{\delta}(j). \qquad \text{(A.34)}
\end{aligned}
$$

By eq. (A.33), we have

$$
\begin{aligned}
\|E[\mathbf{W}\mathbf{W}^T]\| = \left\| \sum_{i,j=1}^{n} E[\mathbf{W}^2(i,j)]\mathbf{e}_i\mathbf{e}_i^T \right\| &\overset{\text{(i)}}{=} \max_{1\leqslant i\leqslant n} \left| \sum_{j=1}^{n} E[\mathbf{W}^2(i,j)] \right| \\
&\overset{\text{(ii)}}{\leqslant} \max_{1\leqslant i\leqslant n} \sum_{j=1}^{n} |\boldsymbol{\theta}(i)\boldsymbol{\delta}(j)| \leqslant \max_{1\leqslant i\leqslant n} \boldsymbol{\theta}(i)\|\boldsymbol{\delta}\|_1 \\
&\leqslant \boldsymbol{\theta}_{\max}\|\boldsymbol{\delta}\|_1, \qquad \text{(A.35)}
\end{aligned}
$$

where (i) follows because $E[\mathbf{W}\mathbf{W}^T] = \sum_{i=1}^{m} \left(\sum_{j=1}^{n} E[\mathbf{W}^2(i,j)] \right) \mathbf{e}_i\mathbf{e}_i^T$ is a diagonal matrix (the spectral norm of a diagonal matrix is the maximum absolute value of its diagonal entries), and (ii) follows from eq. (A.34). Following the similar proof procedure, we obtain $\|E[\mathbf{W}^T\mathbf{W}]\| \leqslant \boldsymbol{\delta}_{\max}\|\boldsymbol{\theta}\|_1$. Thus, we have

$$
\begin{aligned}
V(\mathbf{W}) &= \max\left(\|E[\mathbf{W}\mathbf{W}^T]\|, \|E[\mathbf{W}^T\mathbf{W}]\| \right) \\
&\leqslant \max\left(\boldsymbol{\theta}_{\max}\|\boldsymbol{\delta}\|_1, \boldsymbol{\delta}_{\max}\|\boldsymbol{\theta}\|_1 \right) \\
&\leqslant \max\left(\boldsymbol{\theta}_{\max}, \boldsymbol{\delta}_{\max} \right) \max\left(\|\boldsymbol{\theta}\|_1 , \|\boldsymbol{\delta}\|_1 \right). \qquad \text{(A.36)}
\end{aligned}
$$

Note that $Z \equiv \max\left(\boldsymbol{\theta}_{\max}, \boldsymbol{\delta}_{\max} \right) \max\left(\|\boldsymbol{\theta}\|_1 , \|\boldsymbol{\delta}\|_1 \right)$, by eq. (A.36), we have $V(\mathbf{W}) \leqslant Z$. Since eq. (A.32) implies that $\|\mathbf{W}(i,j)\mathbf{e}_i\mathbf{e}_j^T\|$ is bounded by 1, and these are also independent centered random matrices, we apply the asymmetric version of the matrix version of Bernstein inequality (Theorem 1.6.2 in Tropp (2015)) with $t = 6\sqrt{\log(n)Z}$ and $V(\mathbf{W}) \leqslant Z$, and obtain

$$
\begin{aligned}
P(\|\mathbf{W}\| \geqslant t) &\leqslant 2n \exp\left(\frac{-\frac{t^2}{2}}{V(\mathbf{W}) + \frac{t}{3}}\right) \\
&\leqslant 2n \exp\left(\frac{-18\log(n)Z}{Z + 2\sqrt{\log(n)Z}} \right) \\
&\leqslant 2n \exp\left(\frac{-18\log(n)}{1 + 2\sqrt{\frac{\log(n)}{Z}}} \right) \\
&\leqslant \frac{1}{n^4},
\end{aligned}
$$

where the last inequality follows from eq. (2.13), which implies that $\sqrt{\frac{\log(n)}{Z}} \leqslant 1$ for sufficiently large n. ■

A.7.3. Proof of Lemma A.3

Proof We first introduce the following useful lemma,

Lemma A.7 (Horn and Charles (1985), Theorem 8.4.4) *For every $K \times K$ irreducible, nonnegative, and positive semidefinite matrix $\mathbf{M}$, let $\mathbf{V}_1$ denote the eigenvector corresponding to the largest eigenvalue. Then, the following facts hold:*

(i) $\mathbf{V}_1$ *can be a positive vector.*

(ii) The largest eigenvalue is an algebraically simple eigenvalue.

We note that it is sufficient to prove that $\mathbf{S}^T\mathbf{S}$ and $\mathbf{S}\mathbf{S}^T$ are irreducible and nonnegative, and such properties do not change with n. Once these facts hold, (i) in Lemma A.7 implies $\mathbf{H}_1(i) \geqslant C > 0$ and $\mathbf{Y}_1(i) \geqslant C > 0$, where $\mathbf{H}_1$ and $\mathbf{Y}_1$ are the eigenvectors corresponding to the largest eigenvalues of $\mathbf{S}^T\mathbf{S}$ and $\mathbf{S}\mathbf{S}^T$, respectively. Furthermore, $\mathbf{H}_1(i) \geqslant C > 0$ implies $\mathbf{V}_1(i) > 0$ due to $\mathbf{V}_{\bar{i}} = \frac{\boldsymbol{\delta}(i)}{\|\boldsymbol{\delta}^{(c_i)}\|}\mathbf{H}_{\bar{c_i}}$ (see eq. (4.3)). Similarly, we obtain that $\mathbf{U}_1(i) > 0$ for $1 \leqslant i \leqslant n$. Moreover, (ii) in Lemma A.7 implies $\lambda_1(\mathbf{S}^T\mathbf{S}) - \lambda_2(\mathbf{S}^T\mathbf{S}) \geqslant C$. Then, we complete the proof of Lemma A.3.

Thus, we next prove that $\mathbf{S}^T\mathbf{S}$ and $\mathbf{S}\mathbf{S}^T$ are irreducible and nonnegative, and such properties do not change with n. Recall $\mathbf{S} = \boldsymbol{\Psi}_{\boldsymbol{\theta}}\mathbf{B}\boldsymbol{\Psi}_{\boldsymbol{\delta}}^T$. By eq. (A.29) and the definition of the diagonal matrices $\boldsymbol{\Psi}_{\boldsymbol{\delta}}$ and $\boldsymbol{\Psi}_{\boldsymbol{\theta}}$ (eq. (4.1)), it is clear that there exist C_1 and C_2 such that $0 < C_1 \leqslant \boldsymbol{\Psi}_{\boldsymbol{\theta}}(i,i) \leqslant C_2$ and $0 < C_1 \leqslant \boldsymbol{\Psi}_{\boldsymbol{\delta}}(i,i) \leqslant C_2$. Thus, we have

$$C_1^2\mathbf{B}(i,j) \leqslant \mathbf{S}(i,j) \leqslant C_2^2\mathbf{B}(i,j), \quad \text{for} \quad 1 \leqslant i,j \leqslant K. \tag{A.37}$$

Then, for $1 \leqslant i,j \leqslant K$, we obtain,

$$\begin{aligned}(\mathbf{S}^T\mathbf{S})(i,j) &= \sum_{k=1}^{K}\mathbf{S}^T(i,k)\mathbf{S}(k,j) \leqslant \sum_{k=1}^{K}\mathbf{S}^T(i,k)\mathbf{S}(k,j) \\ &\leqslant C_2^4\sum_{k=1}^{K}\mathbf{B}^T(i,k)\mathbf{B}(k,j) = C(\mathbf{B}^T\mathbf{B})(i,j).\end{aligned} \tag{A.38}$$

Similarly, for $1 \leqslant i,j \leqslant K$, we obtain

$$C(\mathbf{B}^T\mathbf{B})(i,j) \leqslant (\mathbf{S}^T\mathbf{S})(i,j). \tag{A.39}$$

Combining eqs. (A.38) and (A.39), for $1 \leqslant i,j \leqslant K$, we obtain $C(\mathbf{B}^T\mathbf{B})(i,j) \leqslant (\mathbf{S}^T\mathbf{S})(i,j) \leqslant C(\mathbf{B}^T\mathbf{B})(i,j)$. We further note that B is a constant matrix with positive entries, and $\mathbf{B}^T\mathbf{B}$ is irreducible and nonnegative by Assumption 1. Thus, we conclude that $\mathbf{S}^T\mathbf{S}$ is nonnegative and irreducible, and these properties do not change with n. Similarly, we obtain that $\mathbf{S}\mathbf{S}^T$ is also nonnegative and irreducible, and these properties do not change with n. This completes the proof. ■

A.7.4. Proof of Lemma A.5

We first prove $\left|C_V \mathbf{V}_1(i)\right| \asymp \left|\frac{\boldsymbol{\delta}(i)}{\|\boldsymbol{\delta}\|}\right|$ for $1 \leqslant i \leqslant n$. By eq. (4.3), $\mathbf{V}_{\bar{i}} = \frac{\boldsymbol{\delta}(i)}{\|\boldsymbol{\delta}^{(c_i)}\|}\mathbf{H}_{\bar{c}_i}$, and thus $C_V \mathbf{V}_1(i) = C_V \frac{\boldsymbol{\delta}(i)}{\|\boldsymbol{\delta}^{(c_i)}\|}\mathbf{H}_1(c_i)$, where $|C_V| = 1$ by Proposition 2. Following from eqs. (2.9) and (A.10), we have

$$\left|C_V \mathbf{V}_1(i)\right| \asymp \left|C_V \frac{\boldsymbol{\delta}(i)}{\|\boldsymbol{\delta}^{(c_i)}\|}\mathbf{H}_1(c_i)\right| \asymp \left|\frac{\boldsymbol{\delta}(i)}{\|\boldsymbol{\delta}\|}\right|, \quad \text{for } 1 \leqslant i \leqslant n. \tag{A.40}$$

Then, by eq. (A.18), nodes in the set $\hat{S}_V$ satisfies $\left|\frac{\hat{\mathbf{V}}_1(i)}{C_V \mathbf{V}_1(i)} - 1\right| \leqslant C < 1$, and hence $\left|\hat{\mathbf{V}}_1(i)\right| \asymp \left|C_V \mathbf{V}_1(i)\right|$, where $|C_V| = 1$ by Proposition 2. Then, by eq. (A.40), we have for $i \in \hat{S}_V$,

$$\left|\hat{\mathbf{V}}_1(i)\right| \asymp \left|C_V \mathbf{V}_1(i)\right| \asymp \left|\frac{\boldsymbol{\delta}(i)}{\|\boldsymbol{\delta}\|}\right|. \tag{A.41}$$

Similarly, $\left|\hat{\mathbf{U}}_1(i)\right| \asymp \left|C_U \mathbf{U}_1(i)\right| \asymp \left|\frac{\boldsymbol{\theta}(i)}{\|\boldsymbol{\theta}\|}\right|$ for $i \in \hat{S}_U$.

Next, by eq. (A.40), $\left|C_V \mathbf{V}_1(i)\right| \asymp \left|\frac{\boldsymbol{\delta}(i)}{\|\boldsymbol{\delta}\|}\right| > 0$, and thus having $C_V \mathbf{V}_1(i)$ as denominator for all $1 \leqslant i \leqslant n$ is valid. We further derive

$$\begin{aligned}
\sum_{i \in (\mathcal{V} \backslash \hat{S}_V)} \left(\frac{\hat{\mathbf{V}}_1(i)}{C_V \mathbf{V}_1(i)} - 1\right)^2 &= \sum_{i \in (\mathcal{V} \backslash \hat{S}_V)} \left(\frac{1}{C_V \mathbf{V}_1(i)}\right)^2 (\hat{\mathbf{V}}_1(i) - C_V \mathbf{V}_1(i))^2 \\
&\overset{\text{(i)}}{\leqslant} \sum_{i \in (\mathcal{V} \backslash \hat{S}_V)} \frac{\|\boldsymbol{\delta}\|^2}{\delta_{\min}^2} (\hat{\mathbf{V}}_1(i) - C_V \mathbf{V}_1(i))^2 \\
&\leqslant \sum_{i=1}^{n} \frac{\|\boldsymbol{\delta}\|^2}{\delta_{\min}^2} (\hat{\mathbf{V}}_1(i) - C_V \mathbf{V}_1(i))^2 \\
&\leqslant \frac{\|\boldsymbol{\delta}\|^2}{\delta_{\min}^2} \|\hat{V}_1 - \mathbf{V}_1 C_V\|^2 \\
&\overset{\text{(ii)}}{\leqslant} C \frac{\log(n) Z}{\|\boldsymbol{\theta}\|^2 \delta_{\min}^2},
\end{aligned} \tag{A.42}$$

where (i) follows from eq. (A.40), and (ii) follows from Proposition 2. Since nodes in the set $\mathcal{V} \backslash \hat{S}_V$ satisfy $(\frac{\hat{\mathbf{V}}_1(i)}{C_V \mathbf{V}_1(i)} - 1)^2 > C_0^2$, we have

$$|\mathcal{V} \backslash \hat{S}_V| = \sum_{i \in \mathcal{V} \backslash \hat{S}_V} 1 \leqslant \sum_{i \in \mathcal{V} \backslash \hat{S}_V} \frac{1}{C_0^2} \left(\frac{\hat{\mathbf{V}}_1(i)}{C_V \mathbf{V}_1(i)} - 1\right)^2 \overset{\text{(i)}}{\leqslant} C \frac{\log(n) Z}{\|\boldsymbol{\theta}\|^2 \delta_{\min}^2},$$

where (i) follows from eq. (A.42). Similarly, we can show that $\left|\mathcal{V} \backslash \hat{S}_U\right| \leqslant \frac{C \log(n) Z}{\|\boldsymbol{\delta}\|^2 \theta_{\min}^2}$.

A.7.5. Proof of Lemma A.6

Proof We derive the following bound:

$$\begin{aligned}\left\|\frac{\mathbf{v}}{a}-\frac{\mathbf{u}}{b}\right\|^2 &= \left\|\frac{b\mathbf{v}-a\mathbf{u}}{ab}\right\|^2 \\ &= \frac{1}{(ab)^2}\|b\mathbf{v}-b\mathbf{u}+b\mathbf{u}-a\mathbf{u}\|^2 \\ &\leqslant \frac{2}{(ab)^2}\left(\|b\mathbf{v}-b\mathbf{u}\|^2+\|b\mathbf{u}-a\mathbf{u}\|^2\right) \\ &\leqslant \frac{2}{a^2}\|\mathbf{v}-\mathbf{u}\|^2+\frac{2(b-a)^2}{(ab)^2}\|\mathbf{u}\|^2 \\ &= 2\left(\frac{1}{a^2}\|\mathbf{v}-\mathbf{u}\|^2+\frac{(b-a)^2}{(ab)^2}\|\mathbf{u}\|^2\right).\end{aligned}$$

■

Appendix B. Proof of Theorem 2 (Convergence of DSCORE$_q$)

To establish the performance guarantee for D-SCORE$_q$, the general idea is similar to that of D-SCORE, but there are technical differences. Hence, the proof here focuses only on these differences. As in Appendix A, we first prove a few propositions, which then lead to the proof of Theorem 2.

We first state the following two lemmas, which are useful in our proof.

Lemma B.1 *For $\mathbf{x},\mathbf{y}\in\mathcal{R}^d$ where d is finite, the following inequality holds,*

$$\left\|\frac{\mathbf{x}}{\|\mathbf{x}\|_q}-\frac{\mathbf{y}}{\|\mathbf{y}\|_q}\right\| \leqslant C\frac{\|\mathbf{x}-\mathbf{y}\|}{\min\left(\|\mathbf{x}\|_q,\|\mathbf{y}\|_q\right)}.$$

Proof The proof can be found in Appendix B.6.2. ■

Lemma B.2 *For any vector norm $\|\cdot\|$ in the finite dimensional space, it can be bounded by its l_2-norm, i.e., there exists two constants $0<C_1\leqslant C_2$, such that for all x in the finite dimensional space, we have*

$$C_1\|\mathbf{x}\|_2 \leqslant \|\mathbf{x}\| \leqslant C_2\|\mathbf{x}\|_2. \tag{B.1}$$

Proof The proof follows directly from Corollary 5.4.5 in Horn and Charles (1985). ■

We also note that Proposition 1 on the property of the expected adjacency matrix Ω also holds here and is very useful for the analysis of D-SCORE$_q$.

B.1. Proposition 6 and its Proof

In parallel to Proposition 2 for D-SCORE, we bound the distance between the singular vector matrices $\hat{\mathbf{U}}$ and $\hat{\mathbf{V}}$ of $\mathbf{A}$ and the singular vector matrices $\mathbf{U}$ and $\mathbf{V}$ of $\mathbf{\Omega}$. However, for D-SCORE, we need to develop the bound for the first singular vectors and the 2nd to Kth singular vectors separately, whereas for D-SCORE$_q$ we need only to develop the bound for the entire singular vector matrices. In the following proposition, we adapt the same notation for the singular vector matrices of $\mathbf{\Omega}$ and $\mathbf{A}$ as in Proposition 2.

Proposition 6 *There exist two orthogonal matrices $\mathbf{O_V}$ and $\mathbf{O_U}$, such that for n large enough, with probability at least $1 - o(n^{-4})$,*

$$\|\hat{\mathbf{V}} - \mathbf{V}\mathbf{O_V}\|_F \leqslant C\frac{\sqrt{\log(n)Z}}{\|\boldsymbol{\theta}\|\|\boldsymbol{\delta}\|}, \qquad \|\hat{\mathbf{U}} - \mathbf{V}\mathbf{O_U}\|_F \leqslant C\frac{\sqrt{\log(n)Z}}{\|\boldsymbol{\theta}\|\|\boldsymbol{\delta}\|}. \tag{B.2}$$

Proof The proof follows in the same manner as that of Proposition 2 for D-SCORE, based on the direct application of Davis-Kehan inequality. ■

B.2. Proposition 7 and its Proof

The central difference between D-SCORE$_q$ and D-SCORE lies in the way that they eliminate the heterogeneous parameters before clustering. D-SCORE divides each row of the singular vector matrices by its first entry to eliminate the heterogeneous parameters, whereas D-SCORE$_q$ divides each row by its corresponding ℓ_q norm. Then, in parallel to Proposition 3 for D-SCORE, we provide Proposition 7 as follows, which characterizes the properties of the ratio matrix $\mathbf{R} \equiv [\mathbf{R_V}, \mathbf{R_U}]$.

Proposition 7 *For the ratio matrix $\mathbf{R} = [\mathbf{R_V}, \mathbf{R_U}]$ generated by the singular vectors of the matrix $\mathbf{\Omega}$, and for $1 \leqslant i \leqslant n$ and $1 \leqslant j \leqslant n$, the following inequalities hold:*

$$\|\mathbf{R}_{\bar{i}} - \mathbf{R}_{\bar{j}}\|^2 = 0 \quad \textit{if} \quad c_i = c_j, \quad \textit{and} \quad \|\mathbf{R}_{\bar{i}} - \mathbf{R}_{\bar{j}}\|^2 \geqslant C > 0 \quad \textit{if} \quad c_i \neq c_j.$$

Proposition 7 states that if nodes i and j are in the same community, i.e., $c_i = c_j$, then their corresponding rows in the ratio matrix $\mathbf{R}$ are same; otherwise their corresponding rows in $\mathbf{R}$ are different. This property justifies why $\mathbf{R}$ is used for clustering.

Proof First, we have

$$\|\mathbf{R}_{\bar{i}} - \mathbf{R}_{\bar{j}}\|^2 = \|(\mathbf{R_V})_{\bar{i}} - (\mathbf{R_V})_{\bar{j}}\|^2 + \|(\mathbf{R_U})_{\bar{i}} - (\mathbf{R_U})_{\bar{j}}\|^2.$$

For the first term $\|(\mathbf{R_V})_{\bar{i}} - (\mathbf{R_V})_{\bar{j}}\|^2$, by eq. (4.13) which shows $(\mathbf{R_V})_{\bar{i}} = \frac{\mathbf{H}_{\bar{c}_i}\mathbf{O_V}}{\|\mathbf{H}_{\bar{c}_i}\mathbf{O_V}\|_q}$, the following equation holds,

$$\|(\mathbf{R_V})_{\bar{i}} - (\mathbf{R_V})_{\bar{j}}\|^2 = \left\| \frac{\mathbf{H}_{\bar{c}_i}\mathbf{O_V}}{\|\mathbf{H}_{\bar{c}_i}\mathbf{O_V}\|_q} - \frac{\mathbf{H}_{\bar{c}_j}\mathbf{O_V}}{\|\mathbf{H}_{\bar{c}_j}\mathbf{O_V}\|_q} \right\|^2.$$

If $c_i = c_j$, i.e., node i and j are in the same community, and then

$$\|(\mathbf{R_V})_{\bar{i}} - (\mathbf{R_V})_{\bar{j}}\|^2 = 0. \tag{B.3}$$

Otherwise, if $c_i \neq c_j$, we have

$$\begin{aligned}\|(\mathbf{R_V})_{\bar{i}} - (\mathbf{R_V})_{\bar{j}}\|^2 &= \left\| \frac{\mathbf{H}_{\bar{c}_i}\mathbf{O_V}}{\|\mathbf{H}_{\bar{c}_i}\mathbf{O_V}\|_q} - \frac{\mathbf{H}_{\bar{c}_j}\mathbf{O_V}}{\|\mathbf{H}_{\bar{c}_j}\mathbf{O_V}\|_q} \right\|^2 \\ &= \left\| \frac{\mathbf{H}_{\bar{c}_i}\mathbf{O_V}}{\|\mathbf{H}_{\bar{c}_i}\mathbf{O_V}\|_q} \right\|^2 + \left\| \frac{\mathbf{H}_{\bar{c}_j}\mathbf{O_V}}{\|\mathbf{H}_{\bar{c}_j}\mathbf{O_V}\|_q} \right\|^2 - 2\left\langle \frac{\mathbf{H}_{\bar{c}_i}\mathbf{O_V}}{\|\mathbf{H}_{\bar{c}_i}\mathbf{O_V}\|_q}, \frac{\mathbf{H}_{\bar{c}_j}\mathbf{O_V}}{\|\mathbf{H}_{\bar{c}_j}\mathbf{O_V}\|_q} \right\rangle \\ &\overset{\text{(i)}}{=} \left\| \frac{\mathbf{H}_{\bar{c}_i}\mathbf{O_V}}{\|\mathbf{H}_{\bar{c}_i}\mathbf{O_V}\|_q} \right\|^2 + \left\| \frac{\mathbf{H}_{\bar{c}_j}\mathbf{O_V}}{\|\mathbf{H}_{\bar{c}_j}\mathbf{O_V}\|_q} \right\|^2 \\ &\overset{(ii)}{\geqslant} C > 0, \end{aligned} \tag{B.4}$$

where(i) follows from Proposition 1, where $\mathbf{H}$ is an orthogonal matrix so that $\langle \mathbf{H}_{\bar{c}_i}\mathbf{O_V}, \mathbf{H}_{\bar{c}_j}\mathbf{O_V}\rangle = 0$, and (ii) follows from Lemma B.2 so that $\left\| \frac{\mathbf{H}_{\bar{c}_i}\mathbf{O_V}}{\|\mathbf{H}_{\bar{c}_i}\mathbf{O_V}\|_q} \right\|^2 = \frac{\|\mathbf{H}_{\bar{c}_i}\mathbf{O_V}\|^2}{\|\mathbf{H}_{\bar{c}_i}\mathbf{O_V}\|_q^2} \geqslant C > 0$.

Following the similar proof procedure, we obtain

$$\begin{aligned} &\|(\mathbf{R_U})_{\bar{i}} - (\mathbf{R_U})_{\bar{j}}\|^2 = 0 \quad \text{if} \quad c_i = c_j, \\ &\|(\mathbf{R_U})_{\bar{i}} - (\mathbf{R_U})_{\bar{j}}\|^2 \geqslant C > 0 \quad \text{if} \quad c_i \neq c_j. \end{aligned} \tag{B.5}$$

Combining eqs. (B.4), (B.5) and (B.3), we have

$$\|\mathbf{R}_{\bar{i}} - \mathbf{R}_{\bar{j}}\|^2 = 0 \quad \text{if} \quad c_i = c_j, \tag{B.6}$$

$$\|\mathbf{R}_{\bar{i}} - \mathbf{R}_{\bar{j}}\|^2 \geqslant C > 0 \quad \text{if} \quad c_i \neq c_j. \tag{B.7}$$

Thus, if nodes in the same community, they share the same row in $R = [R_V, R_U]$, and if they are in different communities, their corresponding rows in R are sufficiently difference. Since there are K communities, there are exactly K different rows in R. ■

B.3. Proof of Proposition 8

In this section, we develop a bound on the difference between the ratio matrix $\hat{\mathbf{R}}$ generated by the singular vectors of $\mathbf{A}$ and the ratio matrix $\mathbf{R}$ generated by the singular vectors of $\mathbf{\Omega}$, which is in parallel to Proposition 4 for D-SCORE.

Proposition 8 *For $\mathbf{R} = [\mathbf{R_V}, \mathbf{R_U}], \hat{\mathbf{R}} = [\mathbf{R}_{\hat{\mathbf{V}}}, \mathbf{R}_{\hat{\mathbf{U}}}]$, and n large enough, with probability at least $1 - O(n^{-4})$, we have*

$$\|\hat{\mathbf{R}} - \mathbf{R}\|_F^2 \leqslant CT_n^2 \log(n) err_n. \tag{B.8}$$

Proof We define the sets $\hat{S}_V$ and $\hat{S}_U$ as follows:

$$\begin{aligned} \hat{S}_V &= \left(1 \leqslant i \leqslant n; \left| \frac{\|\hat{\mathbf{V}}_{\bar{i}}\|_q}{\|(\mathbf{V}\mathbf{O_V})_{\bar{i}}\|_q} - 1 \right| \leqslant C_0, 0 < C_0 < 1 \right), \\ \hat{S}_U &= \left(1 \leqslant i \leqslant n; \left| \frac{\|\hat{\mathbf{U}}_{\bar{i}}\|_q}{\|(\mathbf{U}\mathbf{O_U})_{\bar{i}}\|_q} - 1 \right| \leqslant C_0, 0 < C_0 < 1 \right). \end{aligned} \tag{B.9}$$

Then, we have the following bounds for these sets.

Lemma B.3 *For nodes in $\hat{S}_V$ or $\hat{S}_U$, the following inequalities hold*

$$\begin{aligned}\|\hat{\mathbf{V}}_{\bar{i}}\| \asymp \|(\mathbf{VO_V})_{\bar{i}}\| \asymp \frac{\boldsymbol{\delta}(i)}{\|\boldsymbol{\delta}^{(c_i)}\|} \quad \textit{for} \quad i \in \hat{S}_V,\\ \|\hat{\mathbf{U}}_{\bar{i}}\| \asymp \|(\mathbf{UO_U})_{\bar{i}}\| \asymp \frac{\boldsymbol{\delta}(i)}{\|\boldsymbol{\delta}^{(c_i)}\|} \quad \textit{for} \quad i \in \hat{S}_U. \end{aligned} \tag{B.10}$$

For n large enough, with probability at least $1 - O(n^{-4})$, the following inequalities hold

$$\left|\mathcal{V}\backslash\hat{S}_V\right| \leqslant \frac{C\log(n)Z}{\|\boldsymbol{\theta}\|^2\delta_{\min}^2} \quad \textit{and} \quad \left|\mathcal{V}\backslash\hat{S}_U\right| \leqslant \frac{C\log(n)Z}{\|\boldsymbol{\delta}\|^2\theta_{\min}^2}. \tag{B.11}$$

Proof The proof can be found in Appendix B.6.1. ■

We are now ready to prove the proposition. By eq. (4.12) and Lemma B.2, we have

$$\|(\mathbf{R_V})_{\bar{i}}\| = \left\|\frac{(\mathbf{VO_V})_{\bar{i}}}{\|(\mathbf{VO_V})_{\bar{i}}\|_q}\right\| \asymp \|\frac{(\mathbf{VO_V})_{\bar{i}}}{\|(\mathbf{VO_V})_{\bar{i}}\|}\| = 1. \tag{B.12}$$

Note that

$$\|\mathbf{R} - \mathbf{R}\|_F^2 = \|\mathbf{R}_{\hat{\mathbf{V}}} - \mathbf{R_V}\|_F^2 + \|\mathbf{R}_{\hat{\mathbf{U}}} - \mathbf{R_U}\|_F^2.$$

We first divide $\|\mathbf{R}_{\hat{\mathbf{V}}} - \mathbf{R_V}\|_F^2$ into the following two parts:

$$\|\mathbf{R}_{\hat{\mathbf{V}}} - \mathbf{R_V}\|_F^2 = \sum_{i\in(\mathcal{V}\backslash\hat{S}_V)} \|(\mathbf{R}_{\hat{\mathbf{V}}})_{\bar{i}} - (\mathbf{R_V})_{\bar{i}}\|^2 + \sum_{i\in\hat{S}_V} \|(\mathbf{R}_{\hat{\mathbf{V}}})_{\bar{i}} - (\mathbf{R_V})_{\bar{i}}\|^2.$$

For the first term, i.e., $i \in (\mathcal{V}\backslash\hat{S}_V)$,

$$\begin{aligned}\sum_{i\in(\mathcal{V}\backslash\hat{S}_V)} \|(\mathbf{R}_{\hat{\mathbf{V}}})_{\bar{i}} - (\mathbf{R_V})_{\bar{i}}\|^2 &\leqslant C \sum_{i\in(\mathcal{V}\backslash\hat{S}_V)} (\|(\mathbf{R}_{\hat{\mathbf{V}}})_{\bar{i}}\|^2 + \|(\mathbf{R_V})_{\bar{i}}\|^2)\\ &\overset{\text{(i)}}{\leqslant} C \sum_{i\in(\mathcal{V}\backslash\hat{S}_V)} (KT_n^2 + C)\\ &\leqslant C|\mathcal{V}\backslash\hat{S}_V|T_n^2\\ &\overset{\text{(ii)}}{\leqslant} \frac{CT_n^2\log(n)Z}{\|\boldsymbol{\theta}\|^2\delta_{\min}^2}, \end{aligned} \tag{B.13}$$

where (i) follows from eq. (3.2), which shows us that the term is truncated by T_n, and eq. (B.12), and (ii) follows from Lemma B.3.

For the second term, i.e., $i \in \hat{S}_V$, we have

$$\sum_{i\in\hat{S}_V} \|(\mathbf{R}_{\hat{\mathbf{V}}})_{\bar{i}} - (\mathbf{R_V})_{\bar{i}}\|^2 \overset{\text{(i)}}{\leqslant} \sum_{i\in\hat{S}_V} \left\|\frac{\hat{\mathbf{V}}_{\bar{i}}}{\|\hat{\mathbf{V}}_{\bar{i}}\|_q} - \frac{(\mathbf{VO_V})_{\bar{i}}}{\|(\mathbf{VO_V})_{\bar{i}}\|_q}\right\|^2$$

$$\overset{\text{(ii)}}{\leqslant} \sum_{i\in\hat{S}_V} \frac{\|\hat{\mathbf{V}}_{\bar{i}} - (\mathbf{V}\mathbf{O}_{\mathbf{V}})_{\bar{i}}\|^2}{\min\left(\|\hat{\mathbf{V}}_{\bar{i}}\|_q^2, \|(\mathbf{V}\mathbf{O}_{\mathbf{V}})_{\bar{i}}\|_q^2\right)}$$
$$\overset{\text{(iii)}}{\leqslant} C\frac{\|\boldsymbol{\delta}\|^2}{\delta_{min}^2} \sum_{i\in\hat{S}_V} \|\hat{\mathbf{V}}_{\bar{i}} - (\mathbf{V}\mathbf{O}_{\mathbf{V}})_{\bar{i}}\|^2$$
$$\overset{\text{(iv)}}{\leqslant} \frac{C\log(n)Z}{\|\boldsymbol{\theta}\|^2\delta_{\min}^2}, \tag{B.14}$$

where (i) follows from eq. (B.12), which implies $\|(\mathbf{R}_{\mathbf{V}})_{\bar{i}}\| = \left\|\frac{(\mathbf{V}\mathbf{O}_{\mathbf{V}})_{\bar{i}}}{\|(\mathbf{V}\mathbf{O}_{\mathbf{V}})_{\bar{i}}\|_q}\right\| \leqslant C$, and hence $T_n \geqslant C \geqslant \|(\mathbf{R}_{\mathbf{V}})_{\bar{i}}\|$ for large n, so that, $\|(\mathbf{R}_{\hat{\mathbf{V}}})_{\bar{i}} - (\mathbf{R}_{\mathbf{V}})_{\bar{i}}\|^2 \leqslant \left\|\frac{\hat{\mathbf{V}}_{\bar{i}}}{\|\hat{\mathbf{V}}_{\bar{i}}\|_q} - \frac{(\mathbf{V}\mathbf{O}_{\mathbf{V}})_{\bar{i}}}{\|(\mathbf{V}\mathbf{O}_{\mathbf{V}})_{\bar{i}}\|_q}\right\|^2$, (ii) follows from Lemma B.1, (iii) follows from Lemma B.3, and (iv) follows from Proposition 6.

Combining eqs. (B.13) and (B.14), we obtain $\|\mathbf{R}_{\hat{\mathbf{V}}} - \mathbf{R}_{\mathbf{V}}\|_F^2 \leqslant \frac{CT_n^2\log(n)Z}{\|\boldsymbol{\theta}\|^2\delta_{\min}^2}$. Similarly, we obtain $\|\mathbf{R}_{\hat{\mathbf{U}}} - \mathbf{R}_{\mathbf{U}}\|_F^2 \leqslant \frac{CT_n^2\log(n)Z}{\|\boldsymbol{\delta}\|^2\theta_{\min}^2}$. Therefore, Proposition 8 follows by combining these two inequalities together. ■

B.4. Proposition 9 and its Proof

The following Proposition 9 is in parallel to Proposition 5 for D-SCORE.

Proposition 9 *For n large enough, with probability at least $1 - O(n^{-4})$, we have*

$$\|\mathbf{M}^* - \mathbf{R}\|_F^2 \leqslant CT_n^2\log(n)err_n.$$

Proof The proof follows in a similar manner to that for Proposition 5 for D-SCORE. ■

B.5. Proof of Theorem 2

Proof The proof follows in a similar manner to that for Theorem 1 for D-SCORE. Note that the constant C in this theorem can be chosen based on Proposition 7, where $\|\mathbf{R}_{\bar{i}} - \mathbf{R}_{\bar{j}}\|^2 \geqslant C > 0$ if $c_i \neq c_j$. ■

B.6. Proof of Lemmas for D-SCORE$_q$

B.6.1. Proof of Lemma B.3

Proof By eq. (4.3), we obtain $\mathbf{V}_{\bar{i}} = \frac{\boldsymbol{\delta}(i)}{\|\boldsymbol{\delta}^{(c_i)}\|}\mathbf{H}_{\bar{c_i}}$, and H is an orthogonal matrix. Thus,

$$\|(\mathbf{V}\mathbf{O}_{\mathbf{V}})_{\bar{i}}\| = \|\mathbf{V}_{\bar{i}}\mathbf{O}_{\mathbf{V}}\| = \|\mathbf{V}_{\bar{i}}\| = \left\|\frac{\boldsymbol{\delta}(i)}{\|\boldsymbol{\delta}^{(c_i)}\|}\mathbf{H}_{\bar{c_i}}\right\| = \frac{\boldsymbol{\delta}(i)}{\|\boldsymbol{\delta}^{(c_i)}\|}. \tag{B.15}$$

With eq. (2.9), we have

$$\|(\mathbf{V}\mathbf{O}_{\mathbf{V}})_{\bar{i}}\| = \frac{\boldsymbol{\delta}(i)}{\|\boldsymbol{\delta}^{(c_i)}\|} \asymp \frac{\boldsymbol{\delta}(i)}{\|\boldsymbol{\delta}\|}. \tag{B.16}$$

Combining Lemma B.2 with eq. (B.16), we have

$$\|(\mathbf{V}\mathbf{O}_{\mathbf{V}})_{\bar{i}}\|_q \asymp \|(\mathbf{V}\mathbf{O}_{\mathbf{V}})_{\bar{i}}\| \asymp \frac{\boldsymbol{\delta}(i)}{\|\boldsymbol{\delta}\|}. \tag{B.17}$$

By definition of $\hat{S}_V$ (eq. (B.9)), for $i \in \hat{S}_V$, we have $1 - C_0 \leqslant \frac{\|\hat{\mathbf{V}}_{\bar{i}}\|_q}{\|(\mathbf{V}\mathbf{O}_{\mathbf{V}})_{\bar{i}}\|_q} \leqslant 1 + C_0$. Thus

$$\|\hat{\mathbf{V}}_{\bar{i}}\|_q \asymp \|(\mathbf{V}\mathbf{O}_{\mathbf{V}})_{\bar{i}}\|_q. \tag{B.18}$$

Combining eqs. (B.17) and (B.18), we conclude that for $i \in \hat{S}_V$

$$\|\hat{\mathbf{V}}_{\bar{i}}\|_q \asymp \frac{\boldsymbol{\delta}(i)}{\|\boldsymbol{\delta}\|}. \tag{B.19}$$

Similarly, we obtain $\|\hat{\mathbf{U}}_{\bar{i}}\| \asymp \|(\mathbf{U}\mathbf{O}_{\mathbf{U}})_{\bar{i}}\| \asymp \frac{\boldsymbol{\theta}(i)}{\|\boldsymbol{\theta}^{(c_i)}\|}$, for $i \in \hat{S}_U$, which completes the proof for eq. (B.10).

To prove eq. (B.11), we first drive

$$\begin{aligned}
\sum_{i\in(\mathcal{V}\backslash\hat{S}_V)} \left(\frac{\|\hat{\mathbf{V}}_{\bar{i}}\|_q}{\|(\mathbf{V}\mathbf{O}_{\mathbf{V}})_{\bar{i}}\|_q} - 1\right)^2 &= \sum_{i\in(\mathcal{V}\backslash\hat{S}_V)} \frac{1}{\|(\mathbf{V}\mathbf{O}_{\mathbf{V}})_{\bar{i}}\|_q^2} (\|\hat{\mathbf{V}}_{\bar{i}}\|_q - \|(\mathbf{V}\mathbf{O}_{\mathbf{V}})_{\bar{i}}\|_q)^2 \\
&\overset{\text{(i)}}{\leqslant} \frac{\|\boldsymbol{\delta}\|^2}{\delta_{\min}^2} \sum_{i\in(\mathcal{V}\backslash\hat{S}_V)} \|\hat{\mathbf{V}}_{\bar{i}} - (\mathbf{V}\mathbf{O}_{\mathbf{V}})_{\bar{i}}\|^2 \\
&\leqslant \frac{\|\boldsymbol{\delta}\|^2}{\delta_{\min}^2} \sum_{i\in\mathcal{V}} \|\hat{\mathbf{V}}_{\bar{i}} - (\mathbf{V}\mathbf{O}_{\mathbf{V}})_{\bar{i}}\|^2 \\
&= \frac{\|\boldsymbol{\delta}\|^2}{\delta_{\min}^2} \|\hat{V} - \mathbf{V}\mathbf{O}_{\mathbf{V}}\|_F^2 \\
&\overset{\text{(ii)}}{\leqslant} C\frac{\log(n)Z}{\|\boldsymbol{\theta}\|^2\delta_{\min}^2},
\end{aligned}$$

where (i) follows because $\big|\|v\| - \|u\|\big| \leqslant \|v-u\|$ and eq. (B.17), and (ii) follows from Proposition 6.

Thus, $\left|\mathcal{V}\backslash\hat{S}_V\right| \leqslant C\frac{\log(n)Z}{\|\boldsymbol{\theta}\|^2\delta_{\min}^2}$. Similar steps can show that $\left|\mathcal{V}\backslash\hat{S}_U\right| \leqslant C\frac{\log(n)Z}{\|\boldsymbol{\delta}\|^2\theta_{\min}^2}$. ■

B.6.2. Proof of Lemma B.1

Proof Without loss of generality, we assume $\|\mathbf{x}\|_q \leqslant \|\mathbf{y}\|_q$, and only need to show

$$\left\|\frac{\mathbf{x}}{\|\mathbf{x}\|_q} - \frac{\mathbf{y}}{\|\mathbf{y}\|_q}\right\|^2 \leqslant C\frac{\|\mathbf{x}-\mathbf{y}\|^2}{\|\mathbf{x}\|_q^2}.$$

We derive

$$\begin{aligned}
\left\| \frac{\mathbf{x}}{\|\mathbf{x}\|_q} - \frac{\mathbf{y}}{\|\mathbf{y}\|_q} \right\|^2 &= \left\| \frac{\mathbf{x}}{\|\mathbf{x}\|_q} - \frac{\mathbf{y}}{\|\mathbf{x}\|_q} + \frac{\mathbf{y}}{\|\mathbf{x}\|_q} - \frac{\mathbf{y}}{\|\mathbf{y}\|_q} \right\|^2 \\
&\overset{\text{(i)}}{\leqslant} 2 \left(\left\| \frac{\mathbf{x}}{\|\mathbf{x}\|_q} - \frac{\mathbf{y}}{\|\mathbf{x}\|_q} \right\|^2 + \left\| \frac{\mathbf{y}}{\|\mathbf{x}\|_q} - \frac{\mathbf{y}}{\|\mathbf{y}\|_q} \right\|^2 \right) \\
&\leqslant 2 \frac{\|\mathbf{x}-\mathbf{y}\|^2}{\|\mathbf{x}\|_q^2} + 2\|\mathbf{y}\|^2 \left| \frac{1}{\|\mathbf{x}\|_q} - \frac{1}{\|\mathbf{y}\|_q} \right|^2 \\
&\leqslant 2 \frac{\|\mathbf{x}-\mathbf{y}\|^2}{\|\mathbf{x}\|_q^2} + 2\|\mathbf{y}\|^2 \left| \frac{\|\mathbf{y}\|_q - \|\mathbf{x}\|_q}{\|\mathbf{x}\|_q \|\mathbf{y}\|_q} \right|^2 \\
&\overset{\text{(ii)}}{\leqslant} 2 \frac{\|\mathbf{x}-\mathbf{y}\|^2}{\|\mathbf{x}\|_q^2} + 2 \frac{\|\mathbf{y}\|^2}{\|\mathbf{x}\|_q^2 \|\mathbf{y}\|_q^2} \|\mathbf{x}-\mathbf{y}\|_q^2 \\
&\overset{\text{(iii)}}{\leqslant} 2 \frac{\|\mathbf{x}-\mathbf{y}\|^2}{\|\mathbf{x}\|_q^2} + C \frac{\|\mathbf{x}-\mathbf{y}\|_q^2}{\|\mathbf{x}\|_q^2} \\
&\overset{\text{(iv)}}{\leqslant} C \frac{\|\mathbf{x}-\mathbf{y}\|^2}{\|\mathbf{x}\|_q^2},
\end{aligned}$$

where (i) follows because $\|\mathbf{x}+\mathbf{y}\|^2 \leqslant 2\big(\|\mathbf{x}\|^2+\|\mathbf{y}\|^2\big)$, and (ii) follows because $\big|\|\mathbf{x}\|_q - \|\mathbf{y}\|_q\big| \leqslant \|\mathbf{x}-\mathbf{y}\|_q$, (iii) follows from Lemma B.2 such that $\|\mathbf{y}\| \asymp \|\mathbf{y}\|_q$, and (iv) follows from Lemma B.2, which implies $\|\mathbf{x}-\mathbf{y}\| \asymp \|\mathbf{x}-\mathbf{y}\|_q$. ■

References

Emmanuel Abbe and Colin Sandon. Proof of the achievability conjectures for the general stochastic block model. *Communications on Pure and Applied Mathematics*, 2017.

Lada A Adamic and Natalie Glance. The political blogosphere and the 2004 U.S. Election. *In Proc. International Workshop on Link Discovery*, 2005.

Arash A. Amini and Elizaveta Levina. On semidefinite relaxations for the block model. *Annals of Statistics*, 46, 2018.

Arash A. Amini, Aiyou Chen, Peter J. Bickel, and Elizaveta Levina. Pseudo-likelihood methods for community detection in large sparse networks. *Annals of Statistics*, 41, 2013. doi: 10.1214/13-AOS1138.

Peter J. Bickel and Aiyou Chen. A nonparametric view of network models and Newman-Girvan and other modularities. *Proceedings of the National Academy of Sciences*, 106, 2009a. doi: 10.1073/pnas.0907096106.

Peter J. Bickel and Aiyou Chen. A nonparametric view of network models and newman–girvan and other modularities. *Proceedings of the National Academy of Sciences*, 106, 2009b.

Peter J. Bickel and Purnamrita Sarkar. Hypothesis testing for automated community detection in networks. *Journal of the Royal Statistical Society: Series B*, 78, 2016.

Kehui Chen and Jing Lei. Network cross-validation for determining the number of communities in network data. *Journal of the American Statistical Association*, 2017.

Yudong Chen, Sujay Sanghavi, and Huan Xu. Clustering sparse graphs. *In Proc. Neural Information Processing Systems (NIPS)*, 2012.

Yudong Chen, Xiaodong Li, and Jiaming Xu. Convexified modularity maximization for degree-corrected stochastic block models. *Annals of Statistics*, 46, 2018.

Erik D. Demaine and Nicole Immorlica. Correlation clustering with partial information. *Approximation, Randomization, and Combinatorial Optimization: Algorithms and Techniques*, 2003.

Anna Goldenberg, Alice X. Zheng, Steven Fienberg, and Edoardo Airoldi. A survey of statistical network models. *Foundations and Trends in Machine Learning*, 2, 2009.

Lennart Gulikers, Marc Lelarge, and Laurent Massoulié. A spectral method for community detection in moderately sparse degree-corrected stochastic block models. *Advances in Applied Probability*, 49, 2017.

Paul W. Holland, Kathryn Blackmond Laskey, and Samuel Leinhardt. Stochastic blockmodels: First steps. *Social Network*, 5, 1983.

A. Roger Horn and R. Johnson Charles. *Matrix Analysis*. Cambridge University Press, 1985.

Pengsheng Ji and Jiashun Jin. Coauthorship and citation networks for statisticians. *Annals of Applied Statistics*, 10, 2016.

Jiashun Jin. Fast community detection by SCORE. *Annals of Statistics*, 43, 2015.

Antony Joseph and Bin Yu. Impact of regularization on spectral clustering. *Annals of Statistics*, 44, 2016.

Brian Karrer and M. E J Newman. Stochastic blockmodels and community structure in networks. *Physical Review E*, 83, 2011.

Jing Lei and Alessandro Rinaldo. Consistency of spectral clustering in stochastic block models. *Annals of Statistics*, 43, 2015.

Jure Leskovec and Andrej Krevl. SNAP Datasets: Stanford large network dataset collection. `http://snap.stanford.edu/data`, 2014.

Fragkiskos D. Malliaros and Michalis Vazirgiannis. Clustering and community detection in directed networks: A survey. *Physics Reports*, 533, 2013.

Elchanan Mossel, Joe Neeman, and Allan Sly. Consistency thresholds for the planted bisection model. *In Proc. ACM Symposium on Theory of Computing*, 2016.

Elchanan Mossel, Joe Neeman, and Allan Sly. A proof of the block model threshold conjecture. *Combinatorica*, 2017.

Mark E. J. Newman. Community detection in networks: Modularity optimization and maximum likelihood are equivalent. *arXiv:1606.02319*, 2016.

Mark E J Newman and Michelle Girvan. Finding and evaluating community structure in networks. *Physical Review E*, 69, 2004.

Hae-Sang Park and Chi-Hyuck Jun. A simple and fast algorithm for K-medoids clustering. *Expert Systems with Applications*, 36, 2009.

Tai Qin and Karl Rohe. Regularized spectral clustering under the degree-corrected stochastic blockmodel. *In Proc. Advances in Neural Information Processing Systems (NIPS)*, 2013.

Joerg Reichardt and Douglas R. White. Role models for complex networks. *The European Physical Journal B*, 60, 2007.

Karl Rohe, Sourav Chatterjee, and Bin Yu. Spectral clustering and the high-dimensional stochastic blockmodel. *Annals of Statistics*, 39, 2011. doi: 10.1214/11-AOS887.

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, 2016.

D. Franco Saldaña, Yi Yu, and Yang Feng. How many communities are there? *Journal of Computational and Graphical Statistics*, 26, 2017.

Daniel L. Sussman, Minh Tang, Donniell E. Fishkind, and Carey E. Priebe. A consistent adjacency embedding for stochastic blockmodel graphs. *Journal of the American Statistical Association*, 107, 2012.

Joel A. Tropp. An introduction to matrix concentration inequalities. *Foundations and Trends in Machine Learning*, 8, 2015.

Yuchung J Wang and George Y Wong. Stochastic blockmodels for directed graphs. *Journal of the American Statistical Association*, 82, 1987.

Tianbao Yang, Yun Chi, Shenghuo Zhu, Yihong Gong, and Rong Jin. Directed network community detection: A popularity and productivity link model. *SIAM International Conference on Data Mining*, 2010.

Yi Yu, Tengyao Wang, and Richard J Samworth. A useful variant of the Davis–Kahan theorem for statisticians. *Biometrika*, 102, 2015.

Yunpeng Zhao, Elizaveta Levina, and Ji Zhu. Community extraction for social networks. *Proceedings of the National Academy of Sciences*, 108, 2011.

Yunpeng Zhao, Elizaveta Levina, and Ji Zhu. Consistency of community detection in networks under degree-corrected stochastic block models. *Annals of Statistics*, 40, 2012.