SIAM J. MATRIX ANAL. APPL.
Vol. 44, No. 2, pp. 781–821

A SPECTRAL METHOD FOR JOINT COMMUNITY DETECTION AND ORTHOGONAL GROUP SYNCHRONIZATION*

YIFENG FAN†, YUEHAW KHOO‡, AND ZHIZHEN ZHAO†

Abstract. Community detection and orthogonal group synchronization are both fundamental problems with a variety of important applications in science and engineering. In this work, we consider the joint problem of community detection and orthogonal group synchronization which aims to recover the communities and perform synchronization simultaneously. To this end, we propose a simple algorithm that consists of a spectral decomposition step followed by a blockwise column pivoted QR factorization. The proposed algorithm is efficient and scales linearly with the number of edges in the graph. We also leverage the recently developed "leave-one-out" technique to establish a near-optimal guarantee for exact recovery of the cluster memberships and stable recovery of the orthogonal transforms. Numerical experiments demonstrate the efficiency and efficacy of our algorithm and confirm our theoretical characterization of it.

Key words. community detection, group synchronization, spectral method, QR factorization

MSC codes. 90C22, 60B20, 62R07

DOI. 10.1137/21M1467845

1. Introduction.

Community detection and Sync. are both fundamental problems in signal processing, machine learning, and computer vision. Recently, there is an increasing interest in their joint problem [28, 8, 50]. That is, in the presence of heterogeneous data where data points associated with random group elements (e.g., the orthogonal group $\mathrm{O}(d)$ of dimension d) fall into multiple underlying clusters, the joint problem is to simultaneously recover the cluster structures, as well as the group elements. A motivating example is the 2D class averaging process in *cryo-electron microscopy single-particle reconstruction* [32, 64, 77], whose goal is to align (with SO(2) group Sync.) and average projection images of a single particle with similar viewing angles to improve their signal-to-noise ratio. Another application in computer vision is *simultaneous permutation group Sync. and clustering* on heterogeneous object collections consisting of 2D images or 3D shapes [8].

In this work, we study the joint problem based on the probabilistic model introduced in [28] which extends the celebrated stochastic block model (SBM) [19, 21, 22, 31, 44, 47, 55, 56, 57, 58] for community detection (see Figure 1 for an illustration, which is slightly modified based on [28, Figure 1]). In particular, we focus on the orthogonal group $\mathrm{O}(d)$ that covers a wide range of applications mentioned above. Formally, given a network of n nodes (data points) with K underlying disjoint communities, each node i is additionally associated with an unknown orthogonal group

*Received by the editors December 28, 2021; accepted for publication (in revised form) by L. Lin January 9, 2023; published electronically June 5, 2023.
https://doi.org/10.1137/21M1467845
Funding: The work of the first and third authors was supported by the National Science Foundation (NSF) grant DMS-1854791 and the Alfred P. Sloan Foundation. The work of the second author was supported by the NSF grant DMS-2111563.

†Department of Electrical and Computer Engineering and Coordinate Science Laboratory, University of Illinois at Urbana-Champaign, Urbana, IL 61801 USA (yifengf2@illinois.edu, zhizhenz@illinois.edu).
‡Department of Statistics, The University of Chicago, Chicago, IL 60637 USA (ykhoo@uchicago.edu).

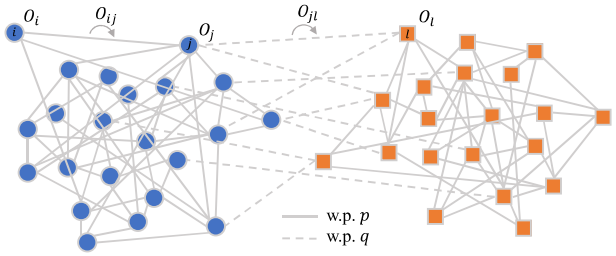


FIG. 1. *We present a network with two communities shown in circles and squares, respectively. Each node is associated with an orthogonal group element. Each pair of nodes within the same cluster (resp., across clusters) is independently connected with probability p (resp., q) as shown in solid (resp., dash) lines. Also, a pairwise alignment $\boldsymbol{O}_{ij}$ is observed on each edge (i,j).*

element $\boldsymbol{O}_i \in \mathrm{O}(d)$. For each pair of nodes (i,j), their orthogonal group transformation $\boldsymbol{O}_{ij}$ is independently observed with probability p (resp., q) when node i and node j belong to the same cluster (resp., different clusters). In particular, the clean measurement $\boldsymbol{O}_{ij} = \boldsymbol{O}_i \boldsymbol{O}_j^\top$ is obtained if i and j are in the same cluster. Otherwise, $\boldsymbol{O}_{ij}$ is uniformly drawn from $\mathrm{O}(d)$, implying the measurement is completely noisy. Notably, such a model of corruption is widely considered in Sync. (e.g., [64, 27, 29, 26, 52]).

Under this probabilistic setting, we want to simultaneously recover the clusters and group elements by combining the ideas for community detection and Sync. A naive two-stage approach is to first apply classical graph-clustering algorithms (e.g., [41, 33, 30, 2]), then perform Sync. within each identified cluster. However, as shown in [28], a better optimization program that solves the two problems jointly can be formulated, which maximizes not only the edge connections but also the consistency of the observed group transformations within each cluster. Directly solving such optimization programs is usually NP-hard and computationally intractable, which gives rise to the convex relaxation methods such as semidefinite relaxation studied in [28] or the spectral method in [8] that yield approximate solutions with polynomial time complexity. The algorithm proposed in this work is also based on a spectral method that first computes the top eigenvectors of an $n \times n$ block matrix of observed data. Then, different from [8], a blockwise column-pivoted QR factorization is performed on the top eigenvectors to identify the cluster structure and orthogonal group elements, which scales linearly with the number of data points. As a result, our method is able to achieve competitive performance with lower computational cost compared to [8].

1.1. Related work and our contributions. Given the practical importance to a variety of applications, both community detection and group synchronization have been extensively studied over the past decades. Due to the vast volume of literature, we are not able to present a complete review of all previous works but only highlight the ones most related to this work. Community detection aims to find the underlying communities within a network by using the network topology information. It is commonly studied under the SBM [22, 31, 44], where obtaining the maximum likelihood estimator for clustering is often NP-hard. Therefore, different approaches such as semidefinite programming (SDP) [1, 41, 42, 61, 39, 4, 9], spectral method [2, 72, 74, 55, 48, 59], and belief propagation [19, 3] are considered for a practical solution. In particular, semidefinite relaxation generally yields state-of-the-art performance as

it is able to achieve the information-theoretic limits of SBM [1, 3, 61, 41, 42]. However, solving large-scale SDPs is still computationally expensive. In contrast, algorithms based on the spectral method are more efficient and also give competitive results (e.g., achieving the information-theoretic limits in the case of two equal-sized clusters [2, 74, 34]). This motivates the proposed method, which extends the spectral method in [16] for clustering, to solve our joint problem.

On the other hand, group synchronization wants to recover the underlying group elements $\{g_i\}_{i=1}^n$ from a set of noisy pairwise measurements $\{g_i^{-1}g_j\}$. A common approach is using a least square estimator which is usually NP-hard. Instead, similar to the development of community detection algorithms, convex relaxations such as semidefinite relaxation [63, 46] and spectral methods [63, 6, 12, 60, 62, 36] have shown to be powerful, along with many investigations of their theoretical properties [63, 78, 52, 46, 53, 35, 76]. Again, spectral method–based algorithms are generally more favorable and appealing than SDPs due to their computational efficiency.

Joint community detection and synchronization is a relatively new topic, motivated by recent scientific applications such as cryo-electron microscopy as mentioned before. In [8], the authors addressed simultaneous permutation group synchronization and clustering via a spectral method for simultaneously mapping and clustering 3D object volumes. In [50], as motivated by the cryo-electron microscopy single-particle reconstruction problem, the authors proposed a harmonic analysis and SDP-based approach for solving the rotational alignment and classification of 2D projection images simultaneously. The recent work [28] by the authors of this paper proposed several SDPs and gave theoretical conditions for exact recovery for the probabilistic model considered. In this work, we propose an alternative spectral method–based algorithm, which greatly reduces the computational complexity of SDP and obtains competitive performance compared to the existing methods.

Besides the algorithm itself, a significant contribution of this work is to provide a near-optimal performance guarantee for exact recovery under the probabilistic model. This requires analyzing the perturbation of the eigenvectors of a low-rank matrix corrupted by random noise, which falls on a classic topic in matrix perturbation theory [65, 67, 5], where a naive ℓ_2 or Frobenius norm error bound can be easily obtained by the Davis–Kahan theorem [18]. However, such a result is not sufficient for exact recovery since it measures the "average" error, and if the error is concentrated on several entries (or blocks if each node is represented by a matrix instead of a scalar), exact recovery is not guaranteed. Instead, an ℓ_∞ norm–type error bound is necessary for exact recovery since it bounds the error entrywisely (or blockwisely). Fortunately, in the past years, we have seen a surge of developments on ℓ_∞ norm bounds. Most of them (e.g., [2, 78, 20, 13, 24, 52]) are based on the *leave-one-out technique* proposed in [2, 78]. In particular, our analysis greatly benefits from [2], which provides an entrywise error bound of the leading eigenvector of low-rank matrices, and also [52], which extends [2] to consider block matrices and gives a blockwise bound on multiple eigenvectors. Results by other approaches exist such as [23, 17]. Specifically, [17] introduces deterministic rowwise perturbation bounds for orthonormal bases of invariant subspaces of symmetric matrices. Such bounds can be applied to general forms of the perturbation matrix. Compared to [17], the leave-one-out technique can exploit the independence of random variables involved and thus achieves sharper bounds in our setting. Here, our contribution lies in handling the additional cluster structures and analyzing the QR factorization.

We summarize our contributions in the following: (1) We introduce a novel algorithm for joint community detection and orthogonal group synchronization, which

consists of three simple steps: a spectral decomposition, followed by a *blockwise* column-pivoted QR factorization (CPQR), and a step for cluster assignment and group element recovery. (2) A variant of CPQR, called *blockwise CPQR*, is designed to deal with the block matrix structure induced by the $\mathrm{O}(d)$, group transformation. (3) Under the probabilistic model, a near-optimal performance guarantee is established for the exact recovery of the cluster memberships and stable recovery of the orthogonal transforms. (4) We demonstrate the efficacy of our method and verify the theoretical characterization of the sharp phase transition for recovery via a series of numerical experiments.

1.2. Organization. The rest of this paper is organized as follows: In section 3 we introduce the probabilistic model and formulate the optimization program for the joint problem. Then in section 4 we present our algorithm. Section 5 is devoted to theoretical analysis. Numerical experiments are given in section 6 for evaluating the performance and verifying our theory. We conclude with discussions and future directions in section 7. For clarity, most of the technical proofs are deferred to the appendix.

1.3. Notations. Throughout this paper we use the following notations: The transpose of a matrix $\boldsymbol{X}$ is denoted by $\boldsymbol{X}^\top$. An $m \times n$ matrix of all zeros is denoted by $\mathbf{0}_{m\times n}$ (or $\mathbf{0}$ for brevity). An identity matrix of size $n \times n$ is denoted by $\boldsymbol{I}_n$. $\sigma_{\max}(\boldsymbol{X})$, $\sigma_{\min}(\boldsymbol{X})$, and $\sigma_l(\boldsymbol{X})$ stand for the maximum, the minimum, and the lth largest singular value of $\boldsymbol{X}$, respectively. Similarly, $\lambda_l(\boldsymbol{X})$ denotes the lth largest eigenvalue of $\boldsymbol{X}$. $\|\boldsymbol{X}\| = \max_{\|v\|=1} \|\boldsymbol{X}\boldsymbol{v}\|$ and $\|\boldsymbol{X}\|_{\mathrm{F}} = \sqrt{\mathrm{Tr}(\boldsymbol{X}^\top \boldsymbol{X})}$ denote the operator norm and the Frobenius norm of $\boldsymbol{X}$, respectively.

For a block matrix $\boldsymbol{X} \in \mathbb{R}^{md\times nd}$, without further specification, the (i,j)th block is denoted by $\boldsymbol{X}_{ij} \in \mathbb{R}^{d\times d}$ for $i = 1,\dots,m$ and $j = 1,\dots,n$. In addition, the ith *block row* (resp., jth *block column*) of $\boldsymbol{X}$ is referred to as the submatrix that contains $\boldsymbol{X}_{ij}$ for all $j = 1,\dots,n$ (resp., $i = 1,\dots,m$) and is denoted as $\boldsymbol{X}_{i\cdot} \in \mathbb{R}^{d\times nd}$ (resp., $\boldsymbol{X}_{\cdot j} \in \mathbb{R}^{md\times d}$).

For two nonnegative functions $f(n)$ and $g(n)$, $f(n) = O(g(n))$ or $f(n) \lesssim g(n)$ means there exists an absolute positive constant C such that $f(n) \le Cg(n)$ for all sufficiently large n; $f(n) = \Omega(g(n))$ or $f(n) \gtrsim g(n)$ means there exists an absolute positive constant C such that $f(n) \ge Cg(n)$ for all sufficiently large n; and $f(n) = o(g(n))$ indicates that, for every positive constant C, the inequality $f(n) \le Cg(n)$ holds for all sufficiently large n.

2. Preliminaries. We start with some important definitions for matrix factorization and decomposition that will be used for algorithm development and analysis.

Definition 2.1 (polar decomposition). *Given a squared matrix $\boldsymbol{X} \in \mathbb{R}^{d\times d}$, the* polar decomposition *of $\boldsymbol{X}$ is given as*

$$\boldsymbol{X} = \mathcal{P}(\boldsymbol{X})\boldsymbol{W}, \tag{2.1}$$

where $\mathcal{P}(\boldsymbol{X}) \in \mathbb{R}^{d\times d}$ is orthogonal and $\boldsymbol{W} \in \mathbb{R}^{d\times d}$ is positive semidefinite.

Notably, such a decomposition always exists. Also, when $\boldsymbol{X}$ has full rank, $\mathcal{P}(\boldsymbol{X})$ is the closest orthogonal matrix to $\boldsymbol{X}$ such that $\mathcal{P}(\boldsymbol{X}) = \mathrm{argmin}_{\boldsymbol{Y}\in\mathrm{O}(d)} \|\boldsymbol{X} - \boldsymbol{Y}\|_{\mathrm{F}}$ (see [25]), and $\boldsymbol{W}$ is guaranteed to be positive definite. In addition, by denoting $\boldsymbol{X} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^\top$ as its singular value decomposition, one can obtain $\mathcal{P}(\boldsymbol{X}) = \boldsymbol{U}\boldsymbol{V}^\top$ and

$\boldsymbol{W} = \boldsymbol{V}\boldsymbol{\Sigma}\boldsymbol{V}^\top$, with computational cost $O(k^3)$. As a result, we also denote $\mathcal{P}(\boldsymbol{X})$ as the *polar factor* of any matrix $\boldsymbol{X}$.

DEFINITION 2.2 (QR factorization). *Given any $\boldsymbol{X} \in \mathbb{R}^{m\times n}$, the* QR factorization *of $\boldsymbol{X}$ is given as*

$$\boldsymbol{X} = \boldsymbol{Q}\boldsymbol{R},$$

where $\boldsymbol{Q} \in \mathbb{R}^{m\times m}$ is orthogonal and $\boldsymbol{R} \in \mathbb{R}^{m\times n}$ is an upper-triangular matrix.

Again, such a factorization always exists. In terms of computing it, the G*ram–Schmidt process* and *Householder transformation* are commonly used (see, e.g., [68]), where the latter approach enjoys better numerical stability.

DEFINITION 2.3 (CPQR). *Given any $\boldsymbol{X} \in \mathbb{R}^{m\times n}$ with $m \le n$ and rank m, the* CPQR *of $\boldsymbol{X}$ is given as*

$$\boldsymbol{X}\boldsymbol{\Pi}_n = \boldsymbol{Q}\,[\boldsymbol{R}_1, \boldsymbol{R}_2] = \boldsymbol{Q}\boldsymbol{R},$$

where $\boldsymbol{\Pi}_n \in \mathbb{R}^{n\times n}$ is a permutation matrix, $\boldsymbol{Q} \in \mathbb{R}^{m\times m}$ is orthogonal, $\boldsymbol{R}_1 \in \mathbb{R}^{m\times m}$ is upper-triangular, and $\boldsymbol{R}_2 \in \mathbb{R}^{m\times(n-m)}$.

The column-pivoted QR differs from the vanilla QR by introducing a permutation $\boldsymbol{\Pi}_n$ that ideally makes $\boldsymbol{R}_1$ as well-conditioned as possible given $\boldsymbol{X}$. In practice, the *Golub–Businger algorithm* [11], which is based on the Householder transformation, chooses $\boldsymbol{\Pi}_n$ using a greedy heuristic: at each step, the column with the largest remaining norm in $\boldsymbol{X}$ is picked as the pivot for computing the new orthogonal basis in $\boldsymbol{Q}$. As a result, CPQR avoids selecting columns that are highly linearly dependent for determining $\boldsymbol{Q}$, which improves the numerical stability especially when $\boldsymbol{X}$ is rank deficient. Because of this, CPQR serves as the backbone of the so-called *rank-revealing QR factorization* [38] which is used to determine the rank of a matrix.

In the following, we introduce an extension of CPQR which considers the case of a block matrix and perform the decomposition blockwisely. This serves as the core of our proposed algorithm in this paper.

DEFINITION 2.4 (blockwise CPQR). *Given any $m \times n$ block matrix $\boldsymbol{X} \in \mathbb{R}^{md\times nd}$ with block size $d \times d$, $m < n$ and rank md (full row rank), the* blockwise CPQR *of $\boldsymbol{X}$ is given as $\boldsymbol{X}\boldsymbol{\Pi}_{nd} = \boldsymbol{Q}[\boldsymbol{R}_1, \boldsymbol{R}_2]$, where $\boldsymbol{\Pi}_{nd} \in \mathbb{R}^{nd\times nd}$ is a permutation matrix with a Kronecker product (denoted by $\otimes$) structure such that*

$$\boldsymbol{\Pi}_{nd} = \boldsymbol{\Pi}_n \otimes \boldsymbol{I}_d$$

for some permutation matrix $\boldsymbol{\Pi}_n \in \mathbb{R}^{n\times n}$, where $\boldsymbol{Q} \in \mathbb{R}^{md\times md}$ is orthogonal, $\boldsymbol{R}_1 \in \mathbb{R}^{md\times md}$ is upper-triangular, and $\boldsymbol{R}_2 \in \mathbb{R}^{md\times(n-m)d}$.

Here, one can view the blockwise CPQR as a special form of the vanilla CPQR, where the matrix $\boldsymbol{X}$ is decomposed blockwisely such that its block structure and the relative orders of columns within each block are always preserved, and the pivot becomes a block column instead. The details for computing it are deferred to section 4.

3. Problem setup. Given a network with n nodes and K underlying communities, we assume each node i has a clustering label $\kappa(i) \in \{1, \dots, K\}$ and is associated with an orthogonal transform $\boldsymbol{O}_i \in \mathrm{O}(d)$. We use m_k and C_k to denote the size of the kth cluster and the set of nodes belonging to it, respectively, such that $m_k = |C_k|$.

Formally, our probabilistic model generates a random graph $G = (V, E)$ with node set V and edge set E. Each pair of nodes (i, j) is independently connected with probability p if $\kappa(j) = \kappa(i)$ that belong to the same community; otherwise they are connected with probability q if $\kappa(j) \neq \kappa(i)$. Also, an orthogonal transformation $\boldsymbol{O}_{ij} \in \mathrm{O}(d)$ is observed on each edge connection $(i, j) \in E$, and when $\kappa(j) = \kappa(i)$ we obtain $\boldsymbol{O}_{ij} = \boldsymbol{O}_i \boldsymbol{O}_j^\top$, which is equal to the true alignment from j to i; otherwise we assume $\boldsymbol{O}_{ij} \sim \mathrm{Unif}(\mathrm{O}(d))$, which is a random orthogonal transformation uniformly drawn from $\mathrm{O}(d)$ that carries no information but only noise.

Given the above, our observation from the model can be represented by an *observation matrix* $\boldsymbol{A} \in \mathbb{R}^{nd \times nd}$, which is an $n \times n$ symmetric block matrix whose (i, j)th block $\boldsymbol{A}_{ij} \in \mathbb{R}^{d \times d}$ for any $i < j$ satisfies

$$\boldsymbol{A}_{ij} = \begin{cases} \boldsymbol{O}_i \boldsymbol{O}_j^\top & \text{with probability } p \text{ and when } \kappa(j) = \kappa(i), \\ \boldsymbol{O}_{ij} \sim \mathrm{Unif}(\mathrm{O}(d)) & \text{with probability } q \text{ and when } \kappa(j) \neq \kappa(i), \\ \boldsymbol{0} & \text{otherwise.} \end{cases} \tag{3.1}$$

Then $\boldsymbol{A}_{ij} = \boldsymbol{A}_{ji}^\top$, and we set the diagonal blocks $\boldsymbol{A}_{ii} = \boldsymbol{0}$ for $i = 1, \ldots, n$. In this way, unlike an adjacency matrix which only has $\{0, 1\}$-valued entries that indicate the connectivity, $\boldsymbol{A}$ defined in (3.1) extends to include orthogonal transformation connected nodes as well. In addition, we define the *clean observation matrix* $\boldsymbol{A}^{\mathrm{clean}} \in \mathbb{R}^{nd \times nd}$, whose (i, j)th block $\boldsymbol{A}_{ij}^{\mathrm{clean}}$ satisfies

$$\boldsymbol{A}_{ij}^{\mathrm{clean}} = \begin{cases} \boldsymbol{O}_i \boldsymbol{O}_j^\top, & \kappa(j) = \kappa(i), \\ \boldsymbol{0} & \text{otherwise.} \end{cases} \tag{3.2}$$

As a result, $\boldsymbol{A}^{\mathrm{clean}}$ is equivalent to $\boldsymbol{A}$ in the clean case when $p = 1$ and $q = 0$.

Given the observation $\boldsymbol{A}$, we formulate the following optimization program for recovery:

$$\max_{\boldsymbol{\Theta}_i, \ldots, \boldsymbol{\Theta}_n \in \mathrm{O}(d), \mathcal{C}_1, \ldots, \mathcal{C}_K} \sum_{k=1}^{K} \sum_{i,j \in \mathcal{C}_k} \left\langle \boldsymbol{A}_{ij}, \frac{1}{|\mathcal{C}_k|} \boldsymbol{\Theta}_i \boldsymbol{\Theta}_j^\top \right\rangle, \tag{3.3}$$

where $\mathcal{C}_k$ denotes the set of nodes assigned to the kth cluster and $\boldsymbol{\Theta}_i$ is the identified orthogonal transform for node i. As a result, (3.3) simultaneously recovers the cluster memberships and the orthogonal group elements by maximizing not only the edge connectivity but also the consistency of transformations among nodes within each cluster. Notably, compared to the formulation in a previous work [28, equation (5)], the additional factor $1/|\mathcal{C}_k|$ in (3.3) is introduced to make the contribution of each cluster to the cost more balanced. Such an idea is in the same spirit as the "RatioCut" [40, 71] studied in the graph partition problem.

To proceed, we perform a change of optimization variables in (3.3): for each cluster $\mathcal{C}_k$, let us define a block column vector $\boldsymbol{V}^{(k)} = [\boldsymbol{V}_i^{(k)}]_{i=1}^n \in \mathbb{R}^{nd \times d}$ of length n whose ith block $\boldsymbol{V}_i^{(k)} \in \mathbb{R}^{d \times d}$ satisfies

$$\boldsymbol{V}_i^{(k)} := \begin{cases} \frac{1}{\sqrt{|\mathcal{C}_k|}} \boldsymbol{\Theta}_i, & i \in \mathcal{C}_k, \\ \boldsymbol{0} & \text{otherwise.} \end{cases} \tag{3.4}$$

As a result, $\boldsymbol{V}_i^{(k)}$ indicates the cluster membership of $\mathcal{C}_k$ and also includes the identified orthogonal group elements of nodes in $\mathcal{C}_k$. Then, (3.3) can be rewritten as

Algorithm 1. Joint spectral clustering and synchronization

Input: The observation matrix $\boldsymbol{A}$, the number of clusters K.

1. (Spectral decomposition) Compute the top Kd eigenvectors $\boldsymbol{\Phi} \in \mathbb{R}^{nd\times Kd}$ of $\boldsymbol{A}$ such that $\boldsymbol{\Phi}^\top\boldsymbol{\Phi} = \boldsymbol{I}_{Kd}$.
2. (Blockwise CPQR) Compute the *blockwise* CPQR (detailed in Algorithm 2) of $\boldsymbol{\Phi}^\top$, which yields

$$\boldsymbol{\Phi}^\top\boldsymbol{\Pi}_{nd} = \boldsymbol{Q}\boldsymbol{R} \quad \Rightarrow \quad \boldsymbol{\Phi}^\top = \boldsymbol{Q}\boldsymbol{R}\boldsymbol{\Pi}_{nd}^\top. \tag{4.1}$$

   Update $\boldsymbol{R} \leftarrow \boldsymbol{R}\boldsymbol{\Pi}_{nd}^\top$.
3. (Recovery of the cluster memberships and orthogonal group elements) For each node $i = 1, \dots, n$, assign its cluster as

$$\hat{\kappa}(i) \leftarrow \operatorname*{argmax}_{k=1,\dots,K} \|\boldsymbol{R}_{ki}\|_{\mathrm{F}}. \tag{4.2}$$

   Identifying the orthogonal group element from the polar factor (Definition 2.1) as

$$\hat{\boldsymbol{O}}_i \leftarrow \mathcal{P}(\boldsymbol{R}_{ki})^\top, \quad \text{where} \quad k = \hat{\kappa}(i). \tag{4.3}$$

4. (Optional) Perform the refinement step described in section 4.3.

Output: Cluster assignments $\{\hat{\kappa}(i)\}_{i=1}^n$ and orthogonal group elements $\{\hat{\boldsymbol{O}}_i\}_{i=1}^n$.

(3.5)

$$\begin{aligned} \max_{\boldsymbol{V}\in\mathbb{R}^{nd\times Kd}} \quad & \left\langle \boldsymbol{A}, \boldsymbol{V}\boldsymbol{V}^\top \right\rangle \\ \text{s.t.} \quad & \boldsymbol{V}\boldsymbol{V}^\top = \sum_{k=1}^K \boldsymbol{V}^{(k)}(\boldsymbol{V}^{(k)})^\top \text{ for } \{\boldsymbol{V}^{(k)}\}_{k=1}^K \text{ satisfies the form in (3.4).} \end{aligned}$$

It is clear to see that (3.5) is nonlinear and nonconvex and is thus computationally intractable to be exactly solved. In [28], the authors use semidefinite relaxations to obtain an approximate solution with polynomial time complexity (on the program without the factor $1/|\mathcal{C}_k|$). However, solving large-scale SDPs is still highly nontrivial in general, especially when n is large. Therefore, here we propose a spectral method detailed in section 4 to improve the efficiency while not sacrificing the accuracy.

4. Algorithms. Given the observation matrix $\boldsymbol{A}$, we propose the following algorithm for simultaneously finding the underlying clusters and synchronizing within each cluster. Our algorithm, as summarized in Algorithm 1, is strikingly simple and only consists of three steps. First, we get the matrix $\boldsymbol{\Phi}$ which contains the top Kd eigenvectors of $\boldsymbol{A}$ via a spectral decomposition. Secondly, we get the matrix $\boldsymbol{R}$ through a blockwise CPQR of $\boldsymbol{\Phi}^\top$. We end up with cluster assignment and orthogonal group element recovery based on the individual subblock of $\boldsymbol{R}$, followed by an optional step for refining the recovery result. In particular, we refer to section 4.2 for the details of the blockwise CPQR given in Definition 2.4.

We highlight that Algorithm 1 is deterministic such that it has no dependency on any sort of random initialization. In comparison, the performance of other

classical algorithms such as k-means [7] largely depends on the initial guess of the cluster centers. Also, in terms of the computational complexity, our algorithm scales linearly with the number of data points n apart from the spectral decomposition step (step 1), which is desirable especially when n is large (see section 4.4 for a detailed discussion). In addition, from an implementation perspective, our algorithm is very easy to implement since it only consists of common and efficient matrix operations.

We point out that CPQR has been widely used in different scientific fields [75, 16, 14, 15], and our algorithm is mainly inspired by [16], which proposes using CPQR for clustering after a spectral embedding of the graph. The algorithm in [16] achieves a competitive and more robust performance against the classical Lloyd algorithm [54] with k-means++ initialization [7] under the SBM model. For our problem, we introduce the blockwise CPQR which naturally extends such a QR-based algorithm to handle the extra group transformations and block structures.

4.1. Algorithm motivation. In this section, we provide motivations for Algorithm 1. We start from the original problem (3.5). By noticing that $\{\boldsymbol{V}^{(k)}\}_{k=1}^{K}$ in (3.4) forms an orthogonal basis, the spectral method relaxes (3.5) by replacing the constraint in (3.5) with $\boldsymbol{V}^\top \boldsymbol{V} = \boldsymbol{I}_{Kd}$ and yields the following relaxed program:

$$\boldsymbol{\Phi} = \underset{\boldsymbol{V}\in\mathbb{R}^{nd\times Kd}}{\operatorname{argmax}} \quad \langle \boldsymbol{A}, \boldsymbol{V}\boldsymbol{V}^\top \rangle \quad \text{s.t.} \quad \boldsymbol{V}^\top \boldsymbol{V} = \boldsymbol{I}_{Kd}, \tag{4.4}$$

whose global optimizer turns out to be the top Kd eigenvectors of $\boldsymbol{A}$ denoted by $\boldsymbol{\Phi} \in \mathbb{R}^{nd\times Kd}$. This leads to *step* 1 (spectral decomposition) in Algorithm 1.

To proceed, we split $\boldsymbol{A}$ into deterministic and random parts,

$$\boldsymbol{A} = \mathbb{E}[\boldsymbol{A}] + (\boldsymbol{A} - \mathbb{E}[\boldsymbol{A}]) = \mathbb{E}[\boldsymbol{A}] + \boldsymbol{\Delta}, \tag{4.5}$$

where $\mathbb{E}[\boldsymbol{A}] = p\boldsymbol{A}^{\text{clean}}$ with the clean observation matrix $\boldsymbol{A}^{\text{clean}}$ defined in (3.2), and the residual $\boldsymbol{\Delta}$ is a random perturbation with $\mathbb{E}[\boldsymbol{\Delta}] = \boldsymbol{0}$. Specifically, $\mathbb{E}[\boldsymbol{A}]$ is a low-rank matrix which satisfies the following decomposition:

$$\mathbb{E}[\boldsymbol{A}] = p\sum_{k=1}^{K} m_k \boldsymbol{\Psi}^{(k)} (\boldsymbol{\Psi}^{(k)})^\top, \quad \text{with} \quad \boldsymbol{\Psi}_i^{(k)} := \begin{cases} \frac{1}{\sqrt{m_k}} \boldsymbol{O}_i, & i \in C_k, \\ \boldsymbol{0}, & \text{otherwise}, \end{cases}$$

where $\boldsymbol{\Psi}^{(k)} = [\boldsymbol{\Psi}_i^{(k)}]_{i=1}^{n} \in \mathbb{R}^{nd\times d}$ is a block column vector of length n that is defined in a similar manner to $\boldsymbol{V}^{(k)}$ in (3.4). Then each $\boldsymbol{\Psi}^{(k)}$ indicates the true cluster memberships of C_k and the exact orthogonal group element $\boldsymbol{O}_i$ of node i within C_k. Therefore, the matrix $\boldsymbol{\Psi} = [\boldsymbol{\Psi}^{(1)}, \boldsymbol{\Psi}^{(2)}, \ldots, \boldsymbol{\Psi}^{(K)}]$ satisfies $\boldsymbol{\Psi}^\top \boldsymbol{\Psi} = \boldsymbol{I}_{Kd}$.

We first consider the clean case when $p = 1$ and $q = 0$; then we have $\boldsymbol{A} = \boldsymbol{A}^{\text{clean}}$, $\boldsymbol{\Delta} = \boldsymbol{0}$, and $\boldsymbol{\Phi} = \boldsymbol{\Psi}\boldsymbol{O}$, where $\boldsymbol{O} \in \mathbb{R}^{Kd\times Kd}$ is some orthogonal matrix. Then for recovering the communities and group elements, it suffices to extract $\{\boldsymbol{\Psi}^{(k)}\}_{k=1}^{K}$ from $\boldsymbol{\Phi}$. However, such extraction is nontrivial since $\boldsymbol{O}$ is unknown to us. To resolve this, without loss of generality, we assume that the first m_1 nodes form the first community C_1, the following m_2 nodes form C_2, and so on and then notice that $\boldsymbol{\Phi}^\top$ can be decomposed as

$$
\begin{aligned}
\boldsymbol{\Phi}^\top &= \boldsymbol{O}^\top \begin{bmatrix} \boldsymbol{\Psi}^{(1)} & \cdots & \boldsymbol{\Psi}^{(k)} \end{bmatrix}^\top \\
&= \boldsymbol{O}^\top \begin{bmatrix} \frac{1}{\sqrt{m_1}}\boldsymbol{O}_1^\top & \cdots & \frac{1}{\sqrt{m_1}}\boldsymbol{O}_m^\top & \cdots & \mathbf{0} & \cdots & \mathbf{0} \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ \mathbf{0} & \cdots & \mathbf{0} & \cdots & \frac{1}{\sqrt{m_K}}\boldsymbol{O}_{n-m_K+1} & \cdots & \frac{1}{\sqrt{m_K}}\boldsymbol{O}_n \end{bmatrix} \\
&= \underbrace{\boldsymbol{O}^\top \begin{bmatrix} \boldsymbol{O}_1^\top & \cdots & \mathbf{0} \\ \vdots & \ddots & \vdots \\ \mathbf{0} & \cdots & \boldsymbol{O}_{n-m_K+1}^\top \end{bmatrix}}_{=:\boldsymbol{Q}} \\
&\quad \times \underbrace{\begin{bmatrix} \frac{1}{\sqrt{m_1}}\boldsymbol{I}_d & \cdots & \frac{1}{\sqrt{m_1}}\boldsymbol{O}_1\boldsymbol{O}_m^\top & \cdots & \mathbf{0} & \cdots & \mathbf{0} \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ \mathbf{0} & \cdots & \mathbf{0} & \cdots & \frac{1}{\sqrt{m_K}}\boldsymbol{I}_d & \cdots & \frac{1}{\sqrt{m_K}}\boldsymbol{O}_{n-m_K+1}\boldsymbol{O}_n^\top \end{bmatrix}}_{=:\boldsymbol{R}} \\
&= \boldsymbol{Q}\boldsymbol{R}.
\end{aligned} \tag{4.6}
$$

The resulting decomposition $\boldsymbol{\Phi}^\top = \boldsymbol{Q}\boldsymbol{R}$ corresponds to *step* 2 (blockwise CPQR) in Algorithm 1 up to some column permutation, and here we assume $\boldsymbol{\Pi}_{nd} = \boldsymbol{I}_{nd}$ such that no column pivoting is performed. In this way, $\boldsymbol{Q} \in \mathbb{R}^{Kd\times Kd}$ is an orthogonal matrix that includes the unknown orthogonal matrix $\boldsymbol{O}$, and $\boldsymbol{R} \in \mathbb{R}^{Kd\times nd}$ is a $K \times n$ block matrix that excludes $\boldsymbol{O}$. More importantly, $\boldsymbol{R}$ incorporates all the information needed for recovery, as shown in the following.

To recover the cluster memberships from $\boldsymbol{R}$ in (4.6), one can see that, for each node i, the ith block column of $\boldsymbol{R}$ (i.e., $\boldsymbol{R}_{\cdot i}$) is sparse such that its kth block $\boldsymbol{R}_{ki}$ is nonzero only if $k = \kappa(i)$, which indicates the cluster membership of node i. Also, the orthogonal group element $\boldsymbol{O}_i$ can be determined from the nonzero block (up to some global orthogonal transformation). This leads to *step* 3 in Algorithm 1, where the polar factor $\mathcal{P}(\cdot)$ in (4.3) ensures the orthogonality of the estimation $\hat{\boldsymbol{O}}_i$.

In practice, when Algorithm 1 is applied to the noisy observation $\boldsymbol{A}$, the exact recovery of the cluster assignments is still possible as long as the perturbation to $\mathbb{E}[\boldsymbol{A}]$, which is controlled by p and q, is less than a certain threshold such that $\boldsymbol{\Phi}$ is still close to $\boldsymbol{\Psi}\boldsymbol{O}$. Indeed, this will be verified by both theoretical analysis in section 5 and numerical experiments in section 6.

4.2. Blockwise column-pivoted QR. An important step of CPQR in Definition 2.3 is selecting the pivots, which amounts to finding a subset of columns that are as linearly independent as possible and are used for determining the basis. In our setting as we present in (4.6), we further require the QR factorization to always preserve the block structure which our recovery relies on; in other words, each $d \times d$ block should be treated as a single unit during the whole process. To handle this requirement, we propose a simple variant of CPQR, called *blockwise CPQR* given in Definition 2.4, which preserves the block structure and can be computed by selecting a block column instead of a single column as the pivot. We present our blockwise CPQR in Algorithm 2, where the *Householder transformation* [45] (see Algorithm 3) is adopted to ensure its numerical stability. Basically, Algorithm 2 modifies the original CPQR algorithm [68, 37] in the following three aspects:

Algorithm 2. Blockwise column-pivoted QR

Input: A block matrix $\boldsymbol{X} \in \mathbb{R}^{Kd \times nd}$ with $K \leq n$ and the block size d.
Initialize: $\boldsymbol{\Pi}_{nd} \leftarrow \boldsymbol{I}_{nd}$, $\boldsymbol{Q} \leftarrow \boldsymbol{I}_{Kd}$, $\boldsymbol{R} \leftarrow \boldsymbol{X}$
1 **for** $t = 1, 2, \ldots, K$ **do**
  `/* Pivot selection */`
2 **for** $j = t, t+1, \ldots, n$ **do**
3 Compute the residual $\rho_j \leftarrow \|\boldsymbol{R}_{t:,\,j}\|_{\mathrm{F}}$, where $\boldsymbol{R}_{t:,\,j} \in \mathbb{R}^{(K-t+1)d \times d}$ is the segment of the jth block column from the tth block to the end
4 **end**
5 Determine the pivot $j^* \leftarrow \operatorname{argmax}_{j=t,\ldots,n} \rho_j$
6 For both $\boldsymbol{R}$ and $\boldsymbol{\Pi}_{nd}$, swap the tth block column with the pivot (j^*th) block column
  `/* Determine orthonormal basis from the pivot block column */`
7 **for** $j = 1, 2, \ldots, d$ **do**
8 Let $l \leftarrow (t-1)d + j$, apply Householder transformation in Algorithm 3 on $\boldsymbol{r}_l = \boldsymbol{R}_{l:Kd,l} \in \mathbb{R}^{Kd-l+1}$, and get the Householder matrix $\widetilde{\boldsymbol{Q}}_l$.
9 Update $\boldsymbol{Q}_l \leftarrow \begin{bmatrix} \boldsymbol{I}_{l-1} & \boldsymbol{0} \\ \boldsymbol{0} & \widetilde{\boldsymbol{Q}}_l \end{bmatrix}$
10 Update $\boldsymbol{R} \leftarrow \boldsymbol{Q}_l \boldsymbol{R}$ and $\boldsymbol{Q} \leftarrow \boldsymbol{Q}\boldsymbol{Q}_l$
11 **end**
12 **end**
Output: $\boldsymbol{Q}$, $\boldsymbol{R}$, and $\boldsymbol{\Pi}_{nd}$.

Algorithm 3. Householder transformation

Input: A column vector $\boldsymbol{x} \in \mathbb{R}^n$.
1 $\alpha \leftarrow -\operatorname{sign}(x_1)\|\boldsymbol{x}\|$
2 $\boldsymbol{u} \leftarrow \boldsymbol{x} - \alpha \boldsymbol{e}_1$, where $\boldsymbol{e}_1 = [1, 0, \ldots, 0]^\top$
3 $\boldsymbol{v} \leftarrow \boldsymbol{u}/\|\boldsymbol{u}\|$
4 $\boldsymbol{Q} \leftarrow \boldsymbol{I}_n - 2\boldsymbol{v}\boldsymbol{v}^\top$
Output: Householder matrix $\boldsymbol{Q}$.

1. At each iteration, we select a *block column* instead of a column as a pivot.
2. We determine the pivot by finding the block column with the largest Frobenius norm of its residual (see line 3 in Algorithm 2).
3. After each pivot selection, we compute d orthonormal basis (instead of only one basis) from the pivot (see lines 7–11 in Algorithm 2) by using the Householder transformation (Algorithm 3).

As a result, the relative order of columns in each block is always preserved.

4.3. Refinements.

On cluster memberships. We propose the following step that further improves the clustering result by Algorithm 1. In (4.2), $\max_k \|\boldsymbol{R}_{ki}\|_{\mathrm{F}}$ can be interpreted a "confidence score" that node i belongs to its assigned cluster. Then our idea is

to refine those cluster assignments with low confidence scores. Formally, given a threshold $\varepsilon \in (0,1)$, we define

$$S_\varepsilon = \left\{ i \,\middle|\, \max_k \frac{\|\boldsymbol{R}_{ki}\|_{\mathrm{F}}}{\|\boldsymbol{R}_{\cdot i}\|_{\mathrm{F}}} \le \varepsilon \right\} \tag{4.7}$$

as the set of ill-defined nodes, where $\boldsymbol{R}_{\cdot i}$ denotes the ith block column. Notably, ε is determined based on the distribution of $\{\max_k \|\boldsymbol{R}_{ki}\|_{\mathrm{F}}/\|\boldsymbol{R}_{\cdot i}\|_{\mathrm{F}}\}_{i=1}^n$ such that S_ε includes a small fraction of nodes and $|S_\varepsilon| \ll n$. As a result, this strategy enables us to control the "level of refinement" instead of setting ε directly. We find that, in practice, refining on a small portion of nodes (e.g., 10%) usually yields a satisfying result with a mild computational complexity. Then, for each node $i \in S_\varepsilon$, we refine its cluster to be

$$\hat{\kappa}(i) = \operatorname*{argmax}_{k=1,\dots,K} \frac{1}{|\hat{C}_k|} \sum_{j \in \hat{C}_k} \sqrt{|\hat{C}_k|}\, \big\|(\boldsymbol{R}_{\cdot i})^\top \boldsymbol{R}_{\cdot j}\big\|_{\mathrm{F}} = \operatorname*{argmax}_{k=1,\dots,K} \frac{1}{\sqrt{|\hat{C}_k|}} \sum_{j \in \hat{C}_k} \big\|(\boldsymbol{R}_{\cdot i})^\top \boldsymbol{R}_{\cdot j}\big\|_{\mathrm{F}}, \tag{4.8}$$

where $\hat{C}_k$ is the kth cluster identified by Algorithm 1. To understand (4.8), consider the clean case as shown in (4.6): for each cluster, we have $\hat{C}_k = C_k$, $|\hat{C}_k| = m_k$, and (4.8) measures the averaged similarity between node i and all nodes $j \in \hat{C}_k$ as

$$\frac{1}{\sqrt{|\hat{C}_k|}} \sum_{j \in \hat{C}_k} \big\|(\boldsymbol{R}_{\cdot i})^\top \boldsymbol{R}_{\cdot j}\big\|_{\mathrm{F}} = \begin{cases} \sqrt{d}, & i \in \hat{C}_k \\ 0, & i \notin \hat{C}_k. \end{cases}$$

Then (4.8) selects the cluster with the maximum similarity. Notably, the factor $\sqrt{|\hat{C}_k|}$ in (4.8) removes the dependency of (4.8) on cluster sizes.

On orthogonal transforms. We also have a step for refining orthogonal transforms as follows: for each cluster $\hat{C}_k$ identified by (4.2), we collect all the available pairwise transforms $\boldsymbol{O}_{ij}$ for nodes in $\hat{C}_k$ and form an observation matrix $\boldsymbol{A}^{(k)} \in \mathbb{R}^{\hat{m}_k d \times \hat{m}_k d}$, where $\hat{m}_k = |\hat{C}_k|$. In other words, $\boldsymbol{A}^{(k)}$ is a restricting of $\boldsymbol{A}$ on $\hat{C}_k$. Then we compute the top-d eigenvectors of $\boldsymbol{A}^{(k)}$ denoted by $\boldsymbol{\Phi}^{(k)} \in \mathbb{R}^{m_k d \times d}$. For each node i, we have

$$\hat{\boldsymbol{O}}_i^{\mathrm{refine}} = \mathcal{P}(\boldsymbol{\Phi}_i^{(k)}) \tag{4.9}$$

as the refined orthogonal transform. Notably, under the probabilistic model in section 3, where the pairwise transforms are noiseless within clusters, one can perfectly identify all the orthogonal transforms $\{\boldsymbol{O}_i\}_{i=1}^n$ by (4.9) as long as the cluster memberships are exactly recovered.[1]

4.4. Complexity. We summarize the complexity of Algorithm 1 in Table 1. Overall, the cost of Algorithm 1 is largely dominated by the spectral decomposition step whose complexity depends on the sparsity of the network and is linear with the number of edges observed in the graph. The remaining steps of Algorithm 1 all together are relatively efficient and scale linearly in n and quadratically in K. As a result,[2] when the data network G is densely connected with $m = O(n^2)$ edges

[1] Here, we also assume the subgraph that corresponds to each cluster is connected (a spanning tree exists). Otherwise, one can add an arbitrary global offset to a connected component without violating the observations.

[2] Note that we only need to compute the top Kd eigenvectors, which is assumed to be done by the Lanczos method [66].

TABLE 1
The complexity of Algorithm 1 in each step. We consider two cases where the network (graph) is densely or sparsely connected. For the case of a sparse network, m denotes the number of edges observed in the graph.

Steps		A dense network	A sparse network
1	Spectral decomposition[2]	$O(Kd^3n^2)$	$O(Kd^3m)$
2	Blockwise CPQR	$O(K^2d^2n + K^2d^3n)$	
3	Clustering by (4.2)	$O(Kd^2n)$	
	Synchronization by (4.3)	$O(d^3n)$	
Total complexity		$O(Kd^3n^2 + K^2d^3n)$	$O(Kd^3m + K^2d^3n)$

observed, Algorithm 1 ends up with $O(n^2)$ complexity. While in practice, it is more common (see, e.g., [51]) to obtain a sparse network G such as $m = O(n \log n)$ or $m = O(n)$, then Algorithm 1 runs efficiently as the complexity reduces to $O(n \log n)$ or $O(n)$, respectively.

We also remark that, in practice, the complexity of spectral decomposition can be further reduced from being linear in K, i.e., from $O(K)$ to $O(\log K)$, by using the randomized algorithm described in, e.g., [43, 73]. But notice that the resulting decomposition is an approximation and could lead to additional errors in the recovery of the cluster memberships and the orthogonal transforms.

5. Analysis. In this section, we investigate the performance of Algorithm 1 under the probabilistic model described in section 3 by finding the condition that guarantees that the clusters $\{C_k\}_{k=1}^K$ are exactly recovered and the orthogonal transforms $\{\boldsymbol{O}_i\}_{i=1}^n$ are estimated with bounded error. For simplicity, we focus on the case of two underlying clusters with equal cluster sizes, namely, $K = 2$ and $m_1 = m_2 = m = n/2$.

To begin with, recall (4.5) that $\boldsymbol{A} = \mathbb{E}[\boldsymbol{A}] + \boldsymbol{\Delta}$, the Davis–Kahan theorem [18] (see Theorem A.3), bounds the difference between the noisy eigenvectors $\boldsymbol{\Phi}$ of $\boldsymbol{A}$ and the clean ones $\boldsymbol{\Psi}$ of $\mathbb{E}[\boldsymbol{A}]$ in terms of the spectral norm as

$$\|\boldsymbol{\Phi} - \boldsymbol{\Psi}\boldsymbol{O}\| \lesssim \frac{\|\boldsymbol{\Delta}\boldsymbol{\Psi}\|}{\delta} \leq \frac{\|\boldsymbol{\Delta}\|\|\boldsymbol{\Psi}\|}{\delta}, \tag{5.1}$$

with a global orthogonal transformation $\boldsymbol{O} = \mathcal{P}(\boldsymbol{\Psi}^\top \boldsymbol{\Phi})$ and a certain spectral gap δ (see Theorem A.3 in Appendix A.1 for the details). Then one may expect that exact recovery can be achieved as long as the error in (5.1) is shown to be sufficiently small. However, even a tight bound of $\|\boldsymbol{\Phi} - \boldsymbol{\Psi}\boldsymbol{O}\|$ cannot guarantee exact recovery since if a few blocks of $\boldsymbol{\Phi}$ have much larger error than others, then exact recovery of every cluster membership could still be failed. Therefore, a more appealing way to show exact recovery is to obtain a blockwise error bound between $\boldsymbol{\Phi}$ and $\boldsymbol{\Psi}$ as

$$\max_{1 \leq i \leq n} \|\boldsymbol{\Phi}_{i\cdot} - \boldsymbol{\Psi}_{i\cdot}\boldsymbol{O}\| \leq \epsilon \tag{5.2}$$

for some uniform ϵ, where $\boldsymbol{\Phi}_{i\cdot}, \boldsymbol{\Psi}_{i\cdot} \in \mathbb{R}^{d \times Kd}$ denote the ith block row of $\boldsymbol{\Phi}$ and $\boldsymbol{\Psi}$, respectively. As a result, the error on each node (block) is uniformly bounded, and one can show that exact recovery is guaranteed as long as ϵ is sufficiently small.

5.1. Main theorems. Our first theoretical result provides a condition that guarantees (5.2) is satisfied for a sufficiently small ϵ, which further leads to the performance guarantee that Algorithm 1 achieves exact recovery under such condition.

THEOREM 5.1 (blockwise error bound). *Under the setting of two equal-sized clusters with model parameters (n,p,q,d), for a sufficiently large n, suppose*

$$\eta := \frac{\sqrt{(p(1-p)+q)\log(nd)}}{p\sqrt{n}} \le c_0 \tag{5.3}$$

for some small constant $c_0 \ge 0$. Then with probability $1-O(n^{-1})$,

$$\max_{1\le i\le n} \|\boldsymbol{\Phi}_{i\cdot} - \boldsymbol{\Psi}_{i\cdot}\boldsymbol{O}\| \lesssim \frac{\eta}{\sqrt{n}}, \tag{5.4}$$

where $\boldsymbol{O} = \mathcal{P}(\boldsymbol{\Psi}^\top \boldsymbol{\Phi})$.

THEOREM 5.2 (performance guarantee). *Under the assumption of Theorem 5.1, for $i = 1,\dots,n$, with probability $1-O(n^{-1})$, Algorithm 1 exactly recovers the cluster memberships $\kappa(i)$ defined in section 3, and $\hat{\boldsymbol{O}}_i$ satisfies*

$$\|\hat{\boldsymbol{O}}_i - \boldsymbol{O}_i \bar{\boldsymbol{O}}_{\kappa(i)}\| \lesssim \eta, \tag{5.5}$$

where $\bar{\boldsymbol{O}}_{\kappa(i)}$ is orthogonal and only depends on the cluster that i belongs to.

As we can see, (5.3) is the condition that leads to exact recovery of the cluster memberships, and (5.5) guarantees the estimation error of $\hat{\boldsymbol{O}}_i$ for each node i is uniformly bounded. Also, by letting $q=0$ and $d=O(1)$, (5.3) reduces to

$$\frac{p}{1-p} \gtrsim \frac{\log n}{n}. \tag{5.6}$$

As a result, (5.6) implies that (5.3) holds and exact recovery is possible if $p \gtrsim \log n/n$. Notably, such a threshold lies in the same regime by using the SDPs studied in [28], which indicates that the spectral method yields competitive results against SDP.

In the following, we outline the key steps for proving Theorems 5.1 and 5.2. The complete proof is deferred to Appendix A. Our proof follows from two important ingredients. One is the *leave-one-out* technique presented in [2, 78] which enables us to have a tight, entrywise analysis on the eigenvectors of low-rank matrices; the second one is [52] which extends the entrywise analysis to a blockwise error bound for studying group synchronization. Here, our main contribution lies in handling the difficulty introduced by the combination of community structures and orthogonal group elements, as well as the QR factorization for clustering.

5.2. The sketch of proof. Throughout the analysis, we assume the first $m = n/2$ nodes belong to C_1 and the remaining m nodes belong to C_2. Without loss of generality, we assume $\boldsymbol{O}_i = \boldsymbol{I}_d$, $i = 1,\dots,n$. For brevity, we use "w.h.p." (with high probability) to represent "with probability $1-O(n^{-1})$".

• *Step 1: Bound $\|\boldsymbol{\Phi}_{i\cdot} - \boldsymbol{\Phi}_{j\cdot}\|$.* We first bound the distance $\|\boldsymbol{\Phi}_{i\cdot} - \boldsymbol{\Phi}_{j\cdot}\|$ for any pair of nodes within the same cluster. The key point lies in finding a suitable surrogate of $\boldsymbol{\Phi}$ such that the difference between $\boldsymbol{\Phi}$ and its surrogate can be tightly bounded. To this end, let $\boldsymbol{\Lambda} \in \mathbb{R}^{2d\times 2d}$ be a diagonal matrix that contains the top $2d$ eigenvalues of

$\boldsymbol{A}$ as diagonal entries; then we have $\boldsymbol{\Phi} = \boldsymbol{A}\boldsymbol{\Phi}\boldsymbol{\Lambda}^{-1}$. To proceed, inspired by [52, 2], let us define the following function of $\boldsymbol{\Phi}$:

$$f(\boldsymbol{\Phi}) := \boldsymbol{A}\boldsymbol{\Phi}\boldsymbol{\Lambda}^{-1}; \tag{5.7}$$

then $\boldsymbol{\Phi}$ is a fixed point of $f(\cdot)$ such that $f(\boldsymbol{\Phi}) = \boldsymbol{\Phi}$. Since (5.1) indicates that $\boldsymbol{\Phi}$ is close to $\boldsymbol{\Psi}\boldsymbol{O}$ with $\boldsymbol{O} = \mathcal{P}(\boldsymbol{\Psi}^\top\boldsymbol{\Phi})$, then we hope the following choice of surrogate

$$f(\boldsymbol{\Psi}\boldsymbol{O}) = \boldsymbol{A}\boldsymbol{\Psi}\boldsymbol{O}\boldsymbol{\Lambda}^{-1} \tag{5.8}$$

serves as a good estimation of $\boldsymbol{\Phi}$ *blockwisely* such that the block error $\|\boldsymbol{\Phi}_{i\cdot} - f(\boldsymbol{\Psi}\boldsymbol{O})_{i\cdot}\|$ is uniformly bounded for any i. If this holds, we can further bound $\|\boldsymbol{\Phi}_{i\cdot} - \boldsymbol{\Phi}_{j\cdot}\|$ as

$$\begin{aligned}\|\boldsymbol{\Phi}_{i\cdot} - \boldsymbol{\Phi}_{j\cdot}\| &= \|\boldsymbol{\Phi}_{i\cdot} - f(\boldsymbol{\Psi}\boldsymbol{O})_{i\cdot} - (\boldsymbol{\Phi}_{j\cdot} - f(\boldsymbol{\Psi}\boldsymbol{O})_{j\cdot}) + f(\boldsymbol{\Psi}\boldsymbol{O})_{i\cdot} - f(\boldsymbol{\Psi}\boldsymbol{O})_{j\cdot}\| \\ &\le \|\boldsymbol{\Phi}_{i\cdot} - f(\boldsymbol{\Psi}\boldsymbol{O})_{i\cdot}\| + \|\boldsymbol{\Phi}_{j\cdot} - f(\boldsymbol{\Psi}\boldsymbol{O})_{j\cdot}\| + \|f(\boldsymbol{\Psi}\boldsymbol{O})_{i\cdot} - f(\boldsymbol{\Psi}\boldsymbol{O})_{j\cdot}\|.\end{aligned} \tag{5.9}$$

To bound $\|\boldsymbol{\Phi}_{i\cdot} - f(\boldsymbol{\Psi}\boldsymbol{O})_{i\cdot}\|$, by definition it satisfies

$$\begin{aligned}\|\boldsymbol{\Phi}_{i\cdot} - f(\boldsymbol{\Psi}\boldsymbol{O})_{i\cdot}\| &= \|[\boldsymbol{\Phi} - \boldsymbol{A}\boldsymbol{\Psi}\boldsymbol{O}\boldsymbol{\Lambda}^{-1}]_{i\cdot}\| = \|[\boldsymbol{A}(\boldsymbol{\Phi} - \boldsymbol{\Psi}\boldsymbol{O})\boldsymbol{\Lambda}^{-1}]_{i\cdot}\| \\ &\le \|\boldsymbol{\Lambda}^{-1}\|\|[\boldsymbol{A}(\boldsymbol{\Phi} - \boldsymbol{\Psi}\boldsymbol{O})]_{i\cdot}\| = \|\boldsymbol{\Lambda}^{-1}\|\|[(\mathbb{E}[\boldsymbol{A}] + \boldsymbol{\Delta})(\boldsymbol{\Phi} - \boldsymbol{\Psi}\boldsymbol{O})]_{i\cdot}\| \\ &\le \|\boldsymbol{\Lambda}^{-1}\|\left(\|[\mathbb{E}[\boldsymbol{A}]]_{i\cdot}(\boldsymbol{\Phi} - \boldsymbol{\Psi}\boldsymbol{O})\| + \|\boldsymbol{\Delta}_{i\cdot}(\boldsymbol{\Phi} - \boldsymbol{\Psi}\boldsymbol{O})\|\right).\end{aligned} \tag{5.10}$$

Here, $\|\boldsymbol{\Lambda}^{-1}\|$ and $\|[\mathbb{E}[\boldsymbol{A}]]_{i\cdot}(\boldsymbol{\Phi} - \boldsymbol{\Psi}\boldsymbol{O})\|$ are tightly bounded by Weyl's inequality (see Theorem A.2) and the Davis–Kahan theorem (see Theorem A.3), respectively. For $\|\boldsymbol{\Delta}_{i\cdot}(\boldsymbol{\Phi} - \boldsymbol{\Psi}\boldsymbol{O})\|$, it is natural to consider applying concentration inequalities for getting a tight bound since $\boldsymbol{\Delta}_i$ consists of independent noisy blocks. However, this is impossible since $\boldsymbol{\Delta}_{i\cdot}$ and $\boldsymbol{\Phi} - \boldsymbol{\Psi}\boldsymbol{O}$ are not statistically independent, which only allows the Cauchy–Schwarz inequality such that $\|\boldsymbol{\Delta}_{i\cdot}(\boldsymbol{\Phi} - \boldsymbol{\Psi}\boldsymbol{O})\| \le \|\boldsymbol{\Delta}_{i\cdot}\|\|\boldsymbol{\Phi} - \boldsymbol{\Psi}\boldsymbol{O}\|$ that is bounded loosely. To resolve this, we resort to the *leave-one-out* technique introduced in [78, 2, 52]; the idea is to define an auxiliary matrix $\boldsymbol{A}^{(i)}$ that "leaves out" the ith block row and column as

$$\boldsymbol{A}^{(i)} := \mathbb{E}[\boldsymbol{A}] + \boldsymbol{\Delta}^{(i)}, \quad \boldsymbol{\Delta}^{(i)}_{kl} := \begin{cases} \boldsymbol{\Delta}_{kl} & \text{if } k \ne i \text{ and } l \ne i, \\ \boldsymbol{0} & \text{if } k = i \text{ or } l = i. \end{cases} \tag{5.11}$$

In other words, $\boldsymbol{A}^{(i)}$ only differs from $\boldsymbol{A}$ on its ith block row and ith block column. Because of this tiny difference, the noise perturbation $\boldsymbol{\Delta}_{i\cdot}$ is not included in $\boldsymbol{A}^{(i)}$, and thus $\boldsymbol{\Delta}_{i\cdot}$ is independent of $\boldsymbol{\Phi}^{(i)}$, which denotes the top Kd eigenvectors of $\boldsymbol{A}^{(i)}$. This enables us to tightly bound $\|\boldsymbol{\Delta}_{i\cdot}(\boldsymbol{\Phi} - \boldsymbol{\Psi}\boldsymbol{O})\|$ by replacing $\boldsymbol{\Phi}$ with $\boldsymbol{\Phi}^{(i)}$ and applying concentration inequalities. To this end, by defining

$$\boldsymbol{O}^{(i)} := \mathcal{P}((\boldsymbol{\Phi}^{(i)})^\top\boldsymbol{\Phi}), \quad \boldsymbol{S}^{(i)} := \mathcal{P}(\boldsymbol{\Psi}^\top\boldsymbol{\Phi}^{(i)}),$$

then $\|\boldsymbol{\Delta}_{i\cdot}(\boldsymbol{\Phi} - \boldsymbol{\Psi}\boldsymbol{O})\|$ can be decomposed as

$$\begin{aligned}\|\boldsymbol{\Delta}_{i\cdot}(\boldsymbol{\Phi} - \boldsymbol{\Psi}\boldsymbol{O})\| &= \|\boldsymbol{\Delta}_{i\cdot}(\boldsymbol{\Phi} - \boldsymbol{\Phi}^{(i)}\boldsymbol{O}^{(i)} + \boldsymbol{\Phi}^{(i)}\boldsymbol{O}^{(i)} - \boldsymbol{\Psi}\boldsymbol{S}^{(i)}\boldsymbol{O}^{(i)} + \boldsymbol{\Psi}\boldsymbol{S}^{(i)}\boldsymbol{O}^{(i)} - \boldsymbol{\Psi}\boldsymbol{O})\| \\ &\le \underbrace{\|\boldsymbol{\Delta}_{i\cdot}(\boldsymbol{\Phi} - \boldsymbol{\Phi}^{(i)}\boldsymbol{O}^{(i)})\|}_{=:T_1} + \underbrace{\|\boldsymbol{\Delta}_{i\cdot}(\boldsymbol{\Phi}^{(i)} - \boldsymbol{\Psi}\boldsymbol{S}^{(i)})\|}_{=:T_2} + \underbrace{\|\boldsymbol{\Delta}_{i\cdot}\boldsymbol{\Psi}(\boldsymbol{S}^{(i)}\boldsymbol{O}^{(i)} - \boldsymbol{O})\|}_{=:T_3} \\ &= T_1 + T_2 + T_3.\end{aligned} \tag{5.12}$$

Then, under the condition (5.3), we bound T_1, T_2, and T_3 separately with details given in Appendix A.2, where particularly T_2 is bounded by matrix Bernstein inequality [69,

Theorem 1.6.2]. Here, we remark that the condition (5.3) is necessary for bounding T_1, T_2, and T_3, as η should be sufficiently small (i.e., $\eta \le c_0$ for some small c_0) for obtaining the intermediate result (see (B.4) in the proof of Lemma A.9 for details)

$$\max_j \|\boldsymbol{\Phi}_{j\cdot}^{(i)}\| \lesssim \max_j \|\boldsymbol{\Phi}_{j\cdot}\|,$$

which implies that after leave-one-out, the norm of the block row of $\boldsymbol{\Phi}^{(i)}$ is universally (at most) in the same order as the original $\boldsymbol{\Phi}$. Then, (5.2) gives the bounds on $\{T_i\}_{i=1}^3$ and further leads to the following bound on $\|\boldsymbol{\Phi}_{i\cdot} - f(\boldsymbol{\Psi O})_{i\cdot}\|$.

LEMMA 5.3. *Under the condition* (5.3)*, for all* $i = 1, \dots, n$*,*

$$\|\boldsymbol{\Phi}_{i\cdot} - f(\boldsymbol{\Psi O})_{i\cdot}\| \lesssim \eta \max_j \|\boldsymbol{\Phi}_{j\cdot}\|$$

with probability $1 - O(n^{-1})$*.*

Lemma 5.3 confirms our previous hypothesis that $\boldsymbol{\Phi}_{i\cdot}$ is well-approximated by its surrogate $f(\boldsymbol{\Psi O})_{i\cdot}$ uniformly as long as η defined in (5.3) is sufficiently small. Back to (5.9), $\|f(\boldsymbol{\Psi O})_{i\cdot} - f(\boldsymbol{\Psi O})_{j\cdot}\|$ can be bounded as

$$\|f(\boldsymbol{\Psi O})_{i\cdot} - f(\boldsymbol{\Psi O})_{j\cdot}\| = \|(\boldsymbol{\Delta}_{i\cdot} - \boldsymbol{\Delta}_{j\cdot})\boldsymbol{\Psi O}\boldsymbol{\Lambda}^{-1}\| \le \|(\boldsymbol{\Delta}_{i\cdot} - \boldsymbol{\Delta}_{j\cdot})\boldsymbol{\Psi}\|\|\boldsymbol{\Lambda}^{-1}\| \lesssim \eta \max_j \|\boldsymbol{\Phi}_{j\cdot}\|$$

with the details deferred to Appendix A.2. Combining the results above yields the following.

LEMMA 5.4. *Under the condition* (5.3)*, for any pair of nodes* (i, j) *that belong to the same cluster, it satisfies*

$$\|\boldsymbol{\Phi}_{i\cdot} - \boldsymbol{\Phi}_{j\cdot}\| \lesssim \eta \max_j \|\boldsymbol{\Phi}_{j\cdot}\| \lesssim \frac{\eta}{\sqrt{n}}$$

with probability $1 - O(n^{-1})$*.*

• *Step 2: Bound* $\|\boldsymbol{\Phi}_{i\cdot} - \boldsymbol{\Psi}_{i\cdot}\boldsymbol{O}\|$. Now we use the results from Step 1 to prove Theorem 5.1, which gives the blockwise error bound $\|\boldsymbol{\Phi}_{i\cdot} - \boldsymbol{\Psi}_{i\cdot}\boldsymbol{O}\|$. The main idea is to combine Lemma 5.4 with the bound on $\|\boldsymbol{\Phi} - \boldsymbol{\Psi O}\|$ given by the Davis–Kahan theorem. To this end, let us define $\boldsymbol{\Phi} - \boldsymbol{\Psi O} = [\boldsymbol{N}_{C_1}^\top, \boldsymbol{N}_{C_2}^\top]^\top$, where $\boldsymbol{N}_{C_1}$ is defined as

$$\boldsymbol{N}_{C_1} := \begin{bmatrix} \boldsymbol{\Phi}_{1\cdot} - \boldsymbol{\Psi}_{1\cdot}\boldsymbol{O} \\ \vdots \\ \boldsymbol{\Phi}_{m\cdot} - \boldsymbol{\Psi}_{m\cdot}\boldsymbol{O} \end{bmatrix} = \underbrace{\begin{bmatrix} \boldsymbol{\Phi}_{i\cdot} - \boldsymbol{\Psi}_{1\cdot}\boldsymbol{O} \\ \vdots \\ \boldsymbol{\Phi}_{i\cdot} - \boldsymbol{\Psi}_{m\cdot}\boldsymbol{O} \end{bmatrix}}_{=:\boldsymbol{N}_i} + \underbrace{\begin{bmatrix} \boldsymbol{\Phi}_{1\cdot} - \boldsymbol{\Phi}_{i\cdot} \\ \vdots \\ \boldsymbol{\Phi}_{m\cdot} - \boldsymbol{\Phi}_{i\cdot} \end{bmatrix}}_{=:\boldsymbol{N}_{\boldsymbol{\Delta},i}} = \boldsymbol{N}_i + \boldsymbol{N}_{\boldsymbol{\Delta},i} \tag{5.13}$$

for all $i \in C_1$, and $\boldsymbol{N}_{C_2}$ is defined similarly for block indices that belong to C_2. Then we can derive a lower bound of $\|\boldsymbol{\Phi} - \boldsymbol{\Psi O}\|$ as

$$\begin{aligned} \|\boldsymbol{\Phi} - \boldsymbol{\Psi O}\| &= \max_{\|\boldsymbol{x}\|=1} \|[\boldsymbol{N}_{C_1}^\top, \boldsymbol{N}_{C_2}^\top]\boldsymbol{x}\| \ge \max_{\|\boldsymbol{y}\|=1} \|\boldsymbol{N}_{C_1}^\top \boldsymbol{y}\| = \|\boldsymbol{N}_{C_1}\| \\ &\ge \|\boldsymbol{N}_i\| - \|\boldsymbol{N}_{\boldsymbol{\Delta},i}\| = \sqrt{m}\|\boldsymbol{\Phi}_{i\cdot} - \boldsymbol{\Psi}_{i\cdot}\boldsymbol{O}\| - O(\eta) \end{aligned}$$

w.h.p. On the other hand, by the Davis–Kahan theorem we can obtain an upper bound as $\|\boldsymbol{\Phi} - \boldsymbol{\Psi O}\| \lesssim \eta/\sqrt{\log(n)}$ w.h.p. (see Lemma A.8 for the detail). Combining the lower bound and the upper bound yields

$$\|\boldsymbol{\Phi}_{i\cdot} - \boldsymbol{\Psi}_{i\cdot}\boldsymbol{O}\| \lesssim \frac{\eta}{\sqrt{n}} \quad \text{for } i \in C_1$$

w.h.p. Similarly, we get the same bound for $i \in C_2$. Then applying the union bound for $i = 1, \ldots, n$ completes the proof of Theorem 5.1.

• *Step 3: Find the performance guarantee of Algorithm 1.* We start by showing the exact recovery of the cluster memberships by Algorithm 1. Recall the matrix $\boldsymbol{R}$ output from the blockwise CPQR in (4.1), and without loss of generality, we assume $p_1 \in C_1$ is the first pivot block column selected. Then for any node j, the two blocks $\boldsymbol{R}_{1j}$ and $\boldsymbol{R}_{2j}$ satisfy

$$\|\boldsymbol{R}_{1j}\|_{\mathrm{F}}^2 = \|\mathcal{P}(\boldsymbol{\Phi}_{p_1\cdot})(\boldsymbol{\Phi}_{j\cdot})^\top\|_{\mathrm{F}}^2, \quad \|\boldsymbol{R}_{2j}\|_{\mathrm{F}}^2 = \|(\boldsymbol{\Phi}_{j\cdot})^\top\|_{\mathrm{F}}^2 - \|\mathcal{P}(\boldsymbol{\Phi}_{p_1\cdot})(\boldsymbol{\Phi}_{j\cdot})^\top\|_{\mathrm{F}}^2,$$

which correspond to the projection of $(\boldsymbol{\Phi}_{j\cdot})^\top$ onto the column space $\mathcal{R}((\boldsymbol{\Phi}_{p_1\cdot})^\top)$ and the complement of $\mathcal{R}((\boldsymbol{\Phi}_{p_1\cdot})^\top)$, respectively. According to Algorithm 1, node j would be assigned to C_1 if $\|\boldsymbol{R}_{1j}\|_{\mathrm{F}} > \|\boldsymbol{R}_{2j}\|_{\mathrm{F}}$, which is equivalent to

$$\|\mathcal{P}(\boldsymbol{\Phi}_{p_1\cdot})(\boldsymbol{\Phi}_{j\cdot})^\top\|_{\mathrm{F}} > \frac{\sqrt{2}}{2}\|\boldsymbol{\Phi}_{j\cdot}\|_{\mathrm{F}}. \tag{5.14}$$

We first show that, for any $j \in C_1$, (5.14) is satisfied w.h.p. To this end, we estimate the LHS and the RHS of (5.14) by using $\boldsymbol{\Psi}_{i\cdot}\boldsymbol{O}$ as a surrogate of $\boldsymbol{\Phi}_{i\cdot}$ for $i = 1, \ldots n$, as they have shown to be similar in Theorem 5.1. Then we obtain that w.h.p.

$$\|\mathcal{P}(\boldsymbol{\Phi}_{p_1\cdot})(\boldsymbol{\Phi}_{j\cdot})^\top\|_{\mathrm{F}} = \left(\sqrt{2} - O(\eta)\right)\sqrt{\frac{d}{n}}, \quad \frac{\sqrt{2}}{2}\|\boldsymbol{\Phi}_{j\cdot}\|_{\mathrm{F}} = (1 + O(\eta))\sqrt{\frac{d}{n}}. \tag{5.15}$$

This implies when the condition (5.3) is satisfied such that η is sufficiently small, (5.14) holds, and j is correctly classified. A similar analysis applies to any $j \in C_2$, and we leave the details in Appendix A.

For recovering the orthogonal transforms $\{\boldsymbol{O}_i\}_{i=1}^n$, we follow the assumption that $p_1 \in C_1$ is the first pivot. In the case of two clusters, the orthogonal matrix $\boldsymbol{Q}$ in the block CPQR (4.1) satisfies $\boldsymbol{Q} = [\boldsymbol{Q}_{\cdot 1}, \boldsymbol{Q}_{\cdot 2}]$, where $\boldsymbol{Q}_{\cdot 1} \in \mathbb{R}^{2d\times d}$ is the polar factor of $(\boldsymbol{\Phi}_{p_1\cdot})^\top$ up to some orthogonal transform, and $\boldsymbol{Q}_{\cdot 2} \in \mathbb{R}^{2d\times d}$ is orthogonal to $\boldsymbol{Q}_{\cdot 1}$, i.e.,

$$\boldsymbol{Q}_{\cdot 1} = \mathcal{P}((\boldsymbol{\Phi}_{p_1\cdot})^\top)\bar{\boldsymbol{O}}_1, \quad \text{and} \quad (\boldsymbol{Q}_{\cdot 1})^\top\boldsymbol{Q}_{\cdot 2} = \boldsymbol{0}, \tag{5.16}$$

where $\bar{\boldsymbol{O}}_1 \in \mathbb{R}^{d\times d}$ is some orthogonal matrix. As a result, for any node $j \in C_1$, our estimation $\hat{\boldsymbol{O}}_j$ by (4.3) can be written as

$$\hat{\boldsymbol{O}}_j = \mathcal{P}(\boldsymbol{R}_{1j}) = \mathcal{P}((\boldsymbol{Q}_{\cdot 1})^\top(\boldsymbol{\Phi}_{j\cdot})^\top).$$

To proceed, by applying Theorem 5.1 we can show the estimation error $\|\hat{\boldsymbol{O}}_j - \boldsymbol{O}_j\bar{\boldsymbol{O}}_1\| \lesssim \eta$. A similar analysis applies to any $j \in C_2$, and we leave the details in Appendix A.

6. Numerical experiments. This section is devoted to numerically investigating the performance of our algorithm. All experiments[3] are performed in MATLAB on a machine with 60 Intel Xeon CPU cores, running at 2.3 GHz with 512 GB RAM in total, and only one core is used for each experiment. In each experiment, we generate the observation matrix $\boldsymbol{A}$ based on the probabilistic model in section 3 and estimate the cluster memberships and the group elements by Algorithm 1. To evaluate the result, for clustering, let $\hat{C}_k = \{i | \hat{\kappa}(i) = k\}$ be the set of nodes identified to the kth cluster by Algorithm 1; then we compute

$$\textit{success rate of exact recovery} = \text{ the rate } \{\hat{C}_k\}_{k=1}^K \text{ is identical to } \{C_k\}_{k=1}^K; \tag{6.1}$$

[3]The code is available at https://github.com/frankfyf/joint_cluster_sync_spectral.

that is, the rate that Algorithm 1 exactly recovers all the clusters memberships. After that, in order to evaluate the quality of identified orthogonal transformations, we define $\boldsymbol{O}^{(k)} = [\boldsymbol{O}_i]_{i\in C_k} \in \mathbb{R}^{m_k d\times d}$ for each cluster C_k as the matrix that concatenates the ground truth $\boldsymbol{O}_i$ for all $i \in C_k$ and similarly define $\hat{\boldsymbol{O}}^{(k)} = [\hat{\boldsymbol{O}}_i]_{i\in C_k}$, the estimated orthogonal transformations. Then we remove the orthogonal ambiguity by aligning $\hat{\boldsymbol{O}}^{(k)}$ with $\boldsymbol{O}^{(k)}$ within each cluster C_k as the following:

$$\boldsymbol{G}^{(k)} = \underset{\boldsymbol{G}\in \mathrm{O}(d)}{\operatorname{argmin}} \|\hat{\boldsymbol{O}}^{(k)} - \boldsymbol{O}^{(k)}\boldsymbol{G}\|_{\mathrm{F}}, \quad k = 1, \dots, K,$$

whose analytical solution is $\boldsymbol{G}^{(k)} = \mathcal{P}((\boldsymbol{O}^{(k)})^\top \hat{\boldsymbol{O}}^{(k)})$. In this way, the error of synchronization is defined as

$$\textit{Error of synchronization} = \log\left(\frac{1}{\sqrt{d}} \max_{k=1,\dots,K} \max_{i\in C_k} \|\hat{\boldsymbol{O}}_i - \boldsymbol{O}_i \boldsymbol{G}^{(k)}\|_{\mathrm{F}}\right), \tag{6.2}$$

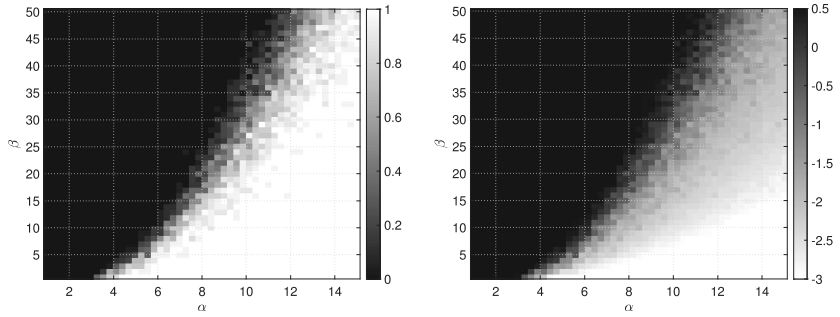
which is the maximum error of our estimation $\hat{\boldsymbol{O}}_i$ over all nodes. As a result, (6.2) is small only if the estimation error of each $\boldsymbol{O}_i$ is bounded. Both (6.1) and (6.2) are averaged over 20 different realizations for each experiment.

We first test the performance of Algorithm 1. We consider the case of two clusters with equal cluster sizes, where we fix $n = 1000$ and test under two settings with $d = 2$ or 3. In particular, since Theorem 5.2 implies that exact recovery is possible at the regime $p, q = O(\log n/n)$, we measure the recovery performance on different $p = \alpha \log n/n$ and $q = \beta \log n/n$ with varying α and β. In Figure 2 we plot the success rate of exact recovery (6.1) and the error of synchronization (6.2). As a result, in both Figure 2(a) and Figure 2(c) we observe sharp phase transitions on the success rate of exact recovery. In Figure 2(b) and Figure 2(d) the error of synchronization follows a similar pattern such that when exact recovery fails we observe a large error, and the error dramatically decreases as exact recovery is achieved. Such observations agree with our theory in Theorem 5.2.

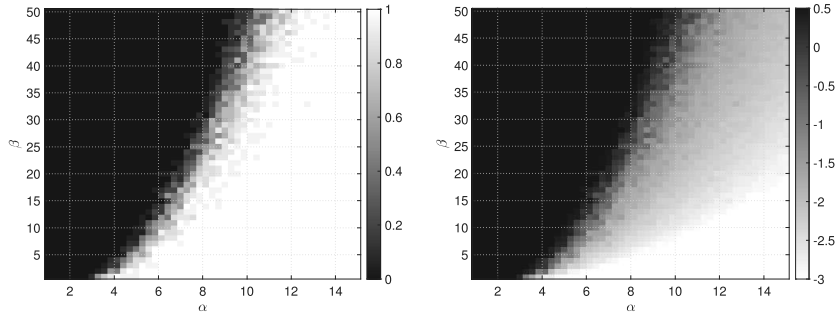
To better visualize the scaling of the phase transition curve, in Figure 3 we plot the success rate of exact recovery under different η defined in (5.3) with varying $p = \alpha \log n/n$ or $q = \beta \log n/n$. Specifically, we set $m_1 = m_2 = 200$, and for a fixed α (resp., β), we adjust η from 0 to 2 by changing β (resp., α) accordingly. As we can see, Figure 3 implies that exact recovery can be achieved with high possibility as $\eta \le 0.5$, which agrees with the theoretical condition (5.3) in Theorem 5.2 that exact recovery is possible as long as $\eta \le c_0$ for some constant c_0. Such observation indicates the sharpness of our condition (5.3).

We further test our algorithm on a more general scenario with five clusters such that $(m_1, m_2, m_3, m_4, m_5) = (100, 200, 200, 200, 300)$ and $d = 2$. We report the result of our algorithm in terms of the metrics (6.1) and (6.2) in Figure 4(a) and Figure 4(b). As a result, we still observe a clear phase transition boundary, which verifies our algorithm can handle arbitrary underlying cluster structures.

We also test the optional refinement step for cluster memberships described in section 4.3, where the threshold ϵ in (4.7) is specified in a way that 10% nodes are included in the set S_ϵ. The result is then displayed in Figure 4(c) and Figure 4(d). We also show another one in Figure 5 under the setting $m_1 = m_2 = 200$ and $d = 2$. As a result, in both examples, we observe a clear improvement in the phase transition boundary of exact recovery of the cluster memberships after the refinement step is applied, which demonstrates the efficacy of refinement.

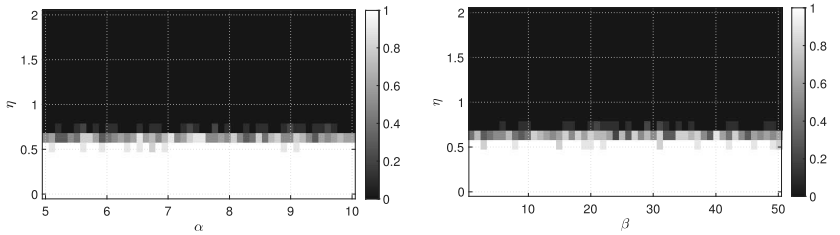


(a) Success rate of exact recovery, $d = 2$ (b) Error of synchronization, $d = 2$

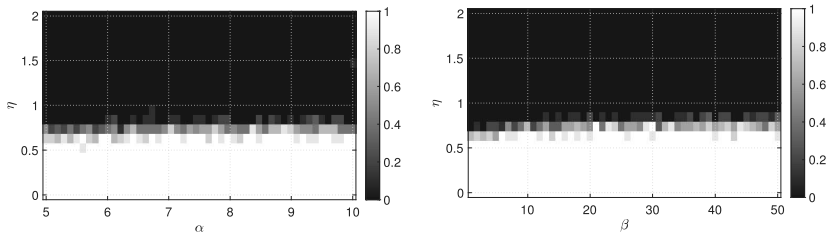


(c) Success rate of exact recovery, $d = 3$ (d) Error of synchronization, $d = 3$

FIG. 2. *Results on two clusters by Algorithm* 1. *We test under the setting* $m_1 = m_2 = 500$, $d = 2$, *or* $d = 3$. (a) *and* (c)*: the success rate of exact recovery by* (6.1), *under varying* α *in* $p = \alpha \log n/n$ *and* β *in* $q = \beta \log n/n$; (b) *and* (d)*: the synchronization error by* (6.2).

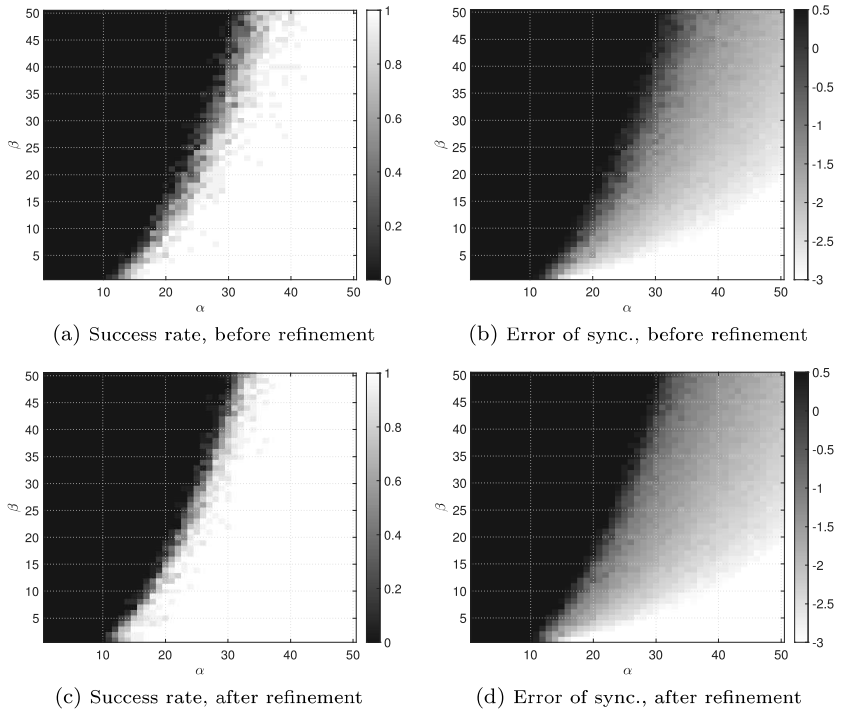


(a) Success rate with varying η and α, $d = 2$ (b) Success rate with varying η and α, $d = 2$



(c) Success rate with varying η and β, $d = 3$ (d) Success rate with varying η and β, $d = 3$

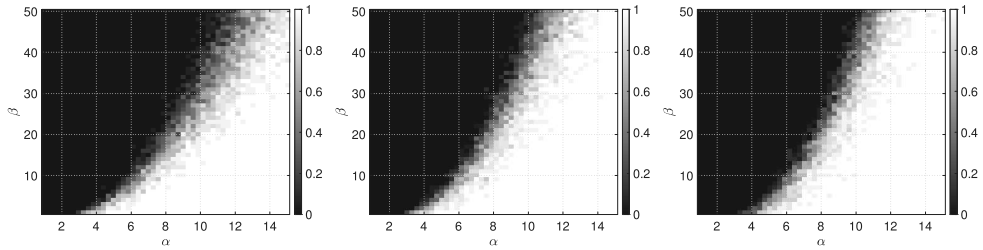
FIG. 3. *Results on two clusters by Algorithm* 1. *We test under the setting* $m_1 = m_2 = 200$, $d = 2$, *or* $d = 3$. (a) *and* (c)*: the success rate of exact recovery under varying* η *defined in* (5.3) *and* α *in* $p = \alpha \log n/n$; (b) *and* (d)*: the success rate of exact recovery under varying* η *and* β *in* $q = \beta \log n/n$. *For a fixed* α *(resp.,* β*), we adjust* η *by changing* β *(resp.,* α*).*



(a) Success rate, before refinement (b) Error of sync., before refinement

(c) Success rate, after refinement (d) Error of sync., after refinement

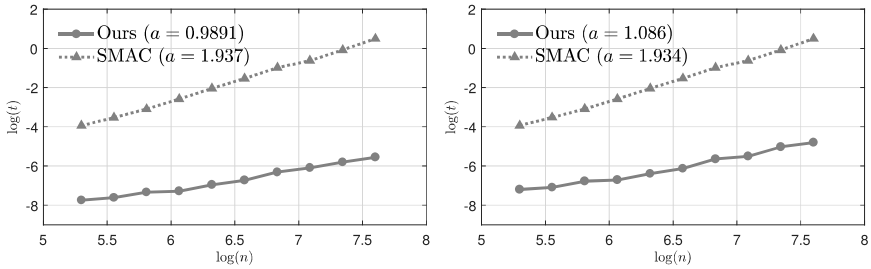
FIG. 4. *Results on five clusters by Algorithm* 1. *We test under the setting* $m_1 = 100$, $m_2 = m_3 = m_4 = 200$, $m_5 = 300$, *and* $d = 2$. *We plot the success rate of exact recovery by* (6.1) *and the synchronization error by* (6.2). (a) *and* (b)*: result by Algorithm* 1*;* (c) *and* (d)*: result after the refinement step* (4.8) *in section* 4.3.



(a) (Ours) Before refinement (b) (Ours) After refinement (c) Result by SMAC [8]

FIG. 5. *We test under the setting* $m_1 = m_2 = 200$ *and* $d = 2$. *We show the success rate of exact recovery of the result by* (a)*: Algorithm* 1*;* (b)*: the refinement step in* (4.8)*;* (c)*: SMAC proposed in* [8].

In addition, we compare our algorithm with an existing method simultaneous mapping and clustering (SMAC) proposed in [8], which is also based on spectral decomposition. Although the original version of SMAC is designed for permutation group synchronization, it can be easily extended to handle the orthogonal group. We test under the setting of $m_1 = m_2 = 200$ and $d = 2$. In Figure 5 we present the success rate of recovery by our algorithm and SMAC. As we can see, the result by our algorithm after the refinement has a similar phase transition boundary as SMAC. However, our method has much less computational complexity than SMAC, as we shall see in the following.



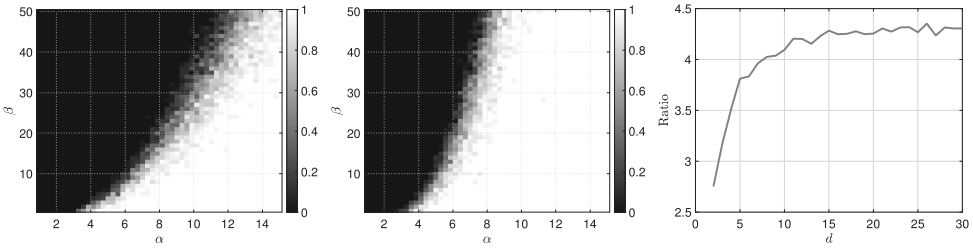
(a) Without spectral decomposition (b) With spectral decomposition

FIG. 6. *Runtime test. We plot the runtime (denoted by t) of different algorithms versus different sizes of the adjacency matrix* $\boldsymbol{A}$ *(denoted by n) in* log*-scale. The slope of each runtime versus problem size curve is noted by a.* (a) *and* (b) *represent the runtime that includes and excludes step* 1 *(spectral decomposition) in Algorithm* 1*, respectively. We set* $p = q = 10 \log n/n$*,* $K = 2$*, and* $d = 2$*. The refinement step in Algorithm* 1 *is excluded.*

To investigate the computational complexity, we further test the runtime of the proposed Algorithm 1 (excluding the refinement step) and SMAC [8] in Figure 6 under the setting $p = q = 10 \log n/n$, $K = 2$, and $d = 2$, and we let n vary from 200 to 2000. All the experiments are performed using the same computational resource mentioned at the beginning of section 6, and we obtain the average timing over 50 trials. Figure 6(a) and Figure 6(b) display the runtime results including and excluding step 1 (spectral decomposition) in Algorithm 1, respectively. From the slopes of the curves, we observe that our algorithm without refinement scales almost linearly with the data size n, and the slope slightly increases as the spectral decomposition is included (recall that the order is of $O(n \log n)$ in theory due to the sparsity). In contrast, SMAC scales quadratically in n since it performs pairwise synchronization. Such observation agrees with our complexity analysis in Table 1 and demonstrates the efficiency of Algorithm 1.

Besides, empirically we observe the phase transition boundary of exact recovery changes as the dimension of the orthogonal group d increases, as shown in Figure 7(a) and Figure 7(b), where we compare the result when $d = 2$ with $d = 10$. One can see that under the same setting of (p, q, n), exact recovery gets easier as d increases. To investigate the reason behind this, we randomly pick a node $i \in C_1$ and compute the "signal-to-noise ratio" $\|\boldsymbol{R}_{1i}\|_{\mathrm{F}}/\|\boldsymbol{R}_{2i}\|_{\mathrm{F}}$, where $\|\boldsymbol{R}_{1i}\|_{\mathrm{F}}$ and $\|\boldsymbol{R}_{2i}\|_{\mathrm{F}}$ can be seen as the signal and the noise level, respectively. Clearly, a larger ratio indicates a lower error probability of clustering. Then, we fix the parameters $p = q = 0.5, m_1 = m_2 = 500$, and ratio $\min_{i \in C_1} \|\boldsymbol{R}_{1i}\|_{\mathrm{F}}/\|\boldsymbol{R}_{2i}\|_{\mathrm{F}}$ among all nodes $i \in C_1$ under different d varying from 2 to 30 in Figure 7(c). As we can see, the ratio increases and converges as d increases, which indicates that clustering becomes easier on a larger dimension d. Unfortunately, such a phenomenon is not characterized by our theory in section 5, and we leave the theoretical investigation as a future study.

7. Discussion and conclusion. In this work, we study joint community detection and orthogonal synchronization by proposing a spectral method-based algorithm. The proposed method is extremely convenient to use and only consists of a spectral decomposition step followed by a blockwise CPQR. As a simple variant of CPQR, blockwise CPQR is designed to ensure that the blockwise nature of the matrices involved is captured and the blockwise structure is always preserved. Such a QR variant is flexible and can be applied to other applications that require QR factorization on a



(a) $d = 2$ (b) $d = 10$ (c) $\min_{i \in C_1} \frac{\|\boldsymbol{R}_{1i}\|_{\mathrm{F}}}{\|\boldsymbol{R}_{2i}\|_{\mathrm{F}}}$

FIG. 7. (a) *and* (b)*: We compare the success rate of exact recovery between* $d = 2$ *and* $d = 10$*, under* $m_1 = m_2 = 500$*.* (c)*: We fix* $p = q = 0.5$ *and plot the minimum ratio* $\min_{i \in C_1} \|\boldsymbol{R}_{i1}\|_{\mathrm{F}} / \|\boldsymbol{R}_{i2}\|_{\mathrm{F}}$ *under different* d *from* 2 *to* 30*.*

block matrix. In terms of time complexity, our algorithm scales linearly with the number of data points, which exhibits a great advantage over other existing methods that at least requires $O(n^2)$ complexity. In addition, under the scenario of two equal-sized clusters, we provide a near-optimal condition that guarantees the underlying cluster memberships are exactly recovered, and the orthogonal transforms are stably recovered. In particular, such condition is obtained by deriving a blockwise error bound on each block of eigenvectors, using the leave-one-out technique [78, 2] rather than the Davis–Kahan theorem [18]. Also, we point out that our theory can be extended to the more general case when having an arbitrary number of clusters of different sizes. To evaluate our algorithm, we perform a series of numerical experiments that demonstrate the efficacy of our algorithm and confirm our theoretical characterization of the sharp phase transition of recovery.

In addition, we have considered an extension of the current result to cover a "noisy" version of the problem by considering an additive Gaussian noise model on the pairwise group transformation. We perform some initial study on this noise model which is given in Appendix C, where we (both theoretically and empirically) show that our proposed Algorithm 1 is still able to robustly recover the cluster memberships and the orthogonal transforms under mild additive noise levels.

There are several directions that can be further explored. First, it is natural to expect that the proposed algorithm can be applied to other groups such as the permutation group. Second, one should be able to extend the theoretical results to more general scenarios with unbalanced cluster sizes and a larger number of clusters,[4] but several major changes are necessary: for example, in general, the spectral norm of the residual $\|\boldsymbol{\Delta}\|$ given in Lemma A.7 should depend on K, but the current proof (see Appendix B.1) only considers two clusters; also, the analysis of exact recovery by CPQR certainly becomes more challenging as more than one pivot is needed to be considered when $K > 2$.

Appendix A. Proof of the main theorems. This section is devoted to the proof of the main theorems given in section 5.

A.1. Important technical ingredients. We first introduce several important technical ingredients that will be frequently used in our analysis.

[4]Here, we still assume the number of clusters scales as a constant, i.e., $K = O(1)$, and each cluster size scales in the same order as the total number of nodes, i.e., $m_i = \Theta(n)$, $n = 1, \ldots, K$.

Polar decomposition. Recall the polar decomposition and the polar factor $\mathcal{P}(\cdot)$ in Definition 2.1; then the following bound will be frequently used in our analysis.

LEMMA A.1 ([52, Lemma 5.2]). *Let $\boldsymbol{X}, \boldsymbol{Y} \in \mathbb{R}^{d\times d}$ be two invertible matrices; then*

$$\|\mathcal{P}(\boldsymbol{X}) - \mathcal{P}(\boldsymbol{Y})\| \le 2\sqrt{2}\min\{\sigma_{\min}^{-1}(\boldsymbol{X}), \sigma_{\min}^{-1}(\boldsymbol{Y})\}\|\boldsymbol{X} - \boldsymbol{Y}\|,$$

where $\sigma_{\min}(\cdot)$ denotes the smallest singular value of a matrix.

Lemma A.1 implies that $\|\mathcal{P}(\boldsymbol{X}) - \mathcal{P}(\boldsymbol{Y})\|$ is bounded as long as $\|\boldsymbol{X} - \boldsymbol{Y}\|$ is bounded and all the singular values of $\boldsymbol{X}$ and $\boldsymbol{Y}$ are bounded away from zero.

Matrix perturbation theory. We include the following classic results in matrix perturbation theory, where Theorem A.2 and Theorem A.3 study the perturbation on singular values and eigenvectors, respectively.

THEOREM A.2 (Weyl's inequality [65]). *Let $\boldsymbol{X}$ and $\boldsymbol{Y}$ be two matrices of the same size; then*

$$|\sigma_i(\boldsymbol{X}) - \sigma_i(\boldsymbol{Y})| \le \|\boldsymbol{X} - \boldsymbol{Y}\| \quad \text{for all } i,$$

where $\sigma_i(\cdot)$ denotes the ith largest singular value of a matrix.

THEOREM A.3 (Davis–Kahan theorem [18]). *Let $\boldsymbol{X}$ and $\hat{\boldsymbol{X}} = \boldsymbol{X} + \boldsymbol{\Delta}$ be two $n \times n$ symmetric matrices with eigenvalues $\{\lambda_i\}_{i=1}^n$ and $\{\hat{\lambda}_i\}_{i=1}^n$, respectively, with the following eigen-decompositions:*

$$\boldsymbol{X} = \begin{bmatrix}\boldsymbol{\Phi}_0 & \boldsymbol{\Phi}_1\end{bmatrix}\begin{bmatrix}\boldsymbol{\Lambda}_0 & \boldsymbol{0}\\ \boldsymbol{0} & \boldsymbol{\Lambda}_1\end{bmatrix}\begin{bmatrix}\boldsymbol{\Phi}_0 & \boldsymbol{\Phi}_1\end{bmatrix}^\top, \quad \hat{\boldsymbol{X}} = \begin{bmatrix}\hat{\boldsymbol{\Phi}}_0 & \hat{\boldsymbol{\Phi}}_1\end{bmatrix}\begin{bmatrix}\hat{\boldsymbol{\Lambda}}_0 & \boldsymbol{0}\\ \boldsymbol{0} & \hat{\boldsymbol{\Lambda}}_1\end{bmatrix}\begin{bmatrix}\hat{\boldsymbol{\Phi}}_0 & \hat{\boldsymbol{\Phi}}_1\end{bmatrix}^\top,$$

where $\boldsymbol{\Lambda}_0$ and $\boldsymbol{\Lambda}_1$ (resp., $\hat{\boldsymbol{\Lambda}}_0$ and $\hat{\boldsymbol{\Lambda}}_1$) are diagonal matrices that contain the top r eigenvalues $\{\lambda_i\}_{i=1}^r$ (resp., $\{\hat{\lambda}_i\}_{i=1}^r$) and the remaining eigenvalues, respectively. $\boldsymbol{\Phi}_k, \hat{\boldsymbol{\Phi}}_k$ for $k = 0, 1$ are the normalized eigenvectors. Then,

$$\|\hat{\boldsymbol{\Phi}}_1^\top \boldsymbol{\Phi}_0\| \le \frac{\|\boldsymbol{\Delta}\boldsymbol{\Phi}_0\|}{\delta},$$

where $\delta := |\lambda_{\min}(\hat{\boldsymbol{\Lambda}}_0) - \lambda_{\max}(\boldsymbol{\Lambda}_1)|$ denotes the spectral gap between $\hat{\boldsymbol{\Lambda}}_0$ and $\boldsymbol{\Lambda}_1$.

Remark A.4. A common technique for analyzing δ above is by obtaining the following lower bound:

$$\begin{aligned}\delta &= |\lambda_{\min}(\hat{\boldsymbol{\Lambda}}_0) - \lambda_{\max}(\boldsymbol{\Lambda}_1)| = |\lambda_r(\hat{\boldsymbol{X}}) - \lambda_{r+1}(\boldsymbol{X})|\\ &= |\lambda_r(\hat{\boldsymbol{X}}) - \lambda_r(\boldsymbol{X}) + \lambda_r(\boldsymbol{X}) - \lambda_{r+1}(\boldsymbol{X})|\\ &\overset{(a)}{\ge} |\lambda_r(\boldsymbol{X}) - \lambda_{r+1}(\boldsymbol{X})| - |\lambda_r(\hat{\boldsymbol{X}}) - \lambda_r(\boldsymbol{X})|\\ &\overset{(b)}{\ge} |\lambda_r(\boldsymbol{X}) - \lambda_{r+1}(\boldsymbol{X})| - \|\boldsymbol{\Delta}\|,\end{aligned}$$

where (a) comes from the triangular inequality and (b) applies Theorem A.2. Here, we assume that $|\lambda_r(\boldsymbol{X}) - \lambda_{r+1}(\boldsymbol{X})| > \|\boldsymbol{\Delta}\|$ such that the spectral gap is greater than the norm of perturbation (otherwise the lower bound becomes trivial). Indeed, this is the case in our analysis because $\|\boldsymbol{\Delta}\| = O(\sqrt{\log n})$ under the condition that $p = \alpha\frac{\log n}{n}$ and $q = \beta\frac{\log n}{n}$, and it is always of lower order compared to $|\lambda_r(\boldsymbol{X}) - \lambda_{r+1}(\boldsymbol{X})| = \Omega(\log n)$. Therefore we have $\delta \approx |\lambda_r(\boldsymbol{X}) - \lambda_{r+1}(\boldsymbol{X})|$. For details of the analysis on $\|\boldsymbol{\Delta}\|$, see Lemma A.6 and its proof.

LEMMA A.5 ([52, Lemma 5.6]). *Suppose* $\boldsymbol{X}, \boldsymbol{Y} \in \mathbb{R}^{n\times d}$ *are two tall orthogonal matrices, i.e.,* $\boldsymbol{X}^\top \boldsymbol{X} = \boldsymbol{Y}^\top \boldsymbol{Y} = \boldsymbol{I}_d$. *Then,*

$$\|\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{O}\| \le 2\left\|(\boldsymbol{I}_n - \boldsymbol{X}\boldsymbol{X}^\top)\boldsymbol{Y}\right\|,$$

where $\boldsymbol{O} = \mathcal{P}(\boldsymbol{X}^\top \boldsymbol{Y})$.

As a result, plugging Lemma A.5 into Theorem A.3 yields

$$\|\boldsymbol{\Phi}_0 - \hat{\boldsymbol{\Phi}}_0 \boldsymbol{O}\| \le 2\|\hat{\boldsymbol{\Phi}}_1^\top \boldsymbol{\Phi}_0\| \le \frac{2\|\boldsymbol{\Delta}\boldsymbol{\Phi}_0\|}{\delta}, \tag{A.1}$$

where $\boldsymbol{O} = \mathcal{P}(\boldsymbol{\Phi}_0^\top \hat{\boldsymbol{\Phi}}_0)$. (A.1) will be frequently used in our analysis.

Analysis of $\boldsymbol{\Delta}$. Recall the residual $\boldsymbol{\Delta} = \boldsymbol{A} - \mathbb{E}[\boldsymbol{A}]$ defined in (4.5). Then under the setting of two equal-sized clusters and the assumption $\boldsymbol{O}_i = \boldsymbol{I}_d, i = 1, \dots, n$, by definition each block $\boldsymbol{\Delta}_{ij} \in \mathbb{R}^{d\times d}$ for $j \ne i$ satisfies

$$\begin{aligned} &\text{when } \kappa(j) = \kappa(i): \ \boldsymbol{\Delta}_{ij} = \begin{cases} (1-p)\boldsymbol{I}_d & \text{with probability } p, \\ -p\boldsymbol{I}_d & \text{with probability } 1-p, \end{cases} \\ &\text{when } \kappa(j) \ne \kappa(i): \ \boldsymbol{\Delta}_{ij} = \begin{cases} \boldsymbol{O}_{ij} & \text{with probability } q, \\ \boldsymbol{0} & \text{with probability } 1-q, \end{cases} \end{aligned} \tag{A.2}$$

where $\boldsymbol{O}_{ij} \sim \mathrm{Unif}(\mathrm{O}(d))$. Also, from the setting $\boldsymbol{A}_{ii} = \boldsymbol{0}$ we have $\boldsymbol{\Delta}_{ii} = \boldsymbol{0}, i = 1, \dots, n$. Given the above, we obtain the following inequalities for $\boldsymbol{\Delta}$.

LEMMA A.6. *Given* $\boldsymbol{\Delta}$ *defined in* (A.2)*, suppose the model parameters* p, q *satisfy* $p = \Omega(\log n/n), 1-p = \Omega(\log n/n)$ *and* $q = \Omega(\log n/n)$*; then*

$$\|\boldsymbol{\Delta}\| \lesssim \sqrt{p(1-p)n} + \sqrt{qn}$$

with probability $1 - O(n^{-1})$. *Also, for the* i*th block row* $\boldsymbol{\Delta}_{i\cdot}$ *and* $\boldsymbol{\Delta}^{(i)}$ *defined in* (5.11)*, we have* $\|\boldsymbol{\Delta}_{i\cdot}\| \le \|\boldsymbol{\Delta}\|$ *and* $\|\boldsymbol{\Delta}^{(i)}\| \le \|\boldsymbol{\Delta}\|$ *for* $i = 1, \dots, n$.

LEMMA A.7. *Given any block matrix* $\boldsymbol{M} \in \mathbb{R}^{nd\times r}$ *with* n *block rows, suppose* $p = \Omega(\log n/n), 1-p = \Omega(\log n/n)$, *and* $q = \Omega(\log n/n)$*; then for each block row* $\boldsymbol{\Delta}_{i\cdot}$ *it satisfies*

$$\|\boldsymbol{\Delta}_{i\cdot}\boldsymbol{M}\| \lesssim \sqrt{(p(1-p)+q)n\log(nd)}\max_j \|\boldsymbol{M}_{j\cdot}\|$$

with probability $1 - O(n^{-1})$.

The proofs of both Lemma A.6 and Lemma A.7 are deferred to Appendix B.

A.2. Proof of Theorem 5.1. In this section, we prove Theorem 5.1 which provides the blockwise error bound in (5.4). We first introduce a series of necessary inequalities, followed by the proofs of Lemmas 5.3 and 5.4 that are presented in section 5.2. Then we end up with the proof of Theorem 5.1. All the lemmas introduced in this section are proved in Appendix B.3.

To begin with, recall that $\boldsymbol{\Phi}$, $\boldsymbol{\Psi}$, and $\boldsymbol{\Phi}^{(i)}$ are defined as the top Kd eigenvectors of $\boldsymbol{A}$, $\mathbb{E}[\boldsymbol{A}]$, and the auxiliary matrix $\boldsymbol{A}^{(i)}$ is defined in (5.11), respectively. We first bound the difference between $\boldsymbol{\Phi}$ (or $\boldsymbol{\Phi}^{(i)}$) and $\boldsymbol{\Psi}$ in the following.

LEMMA A.8. *For a sufficiently large n, let $\tau := \frac{\sqrt{p(1-p)}+\sqrt{q}}{p\sqrt{n}}$; then the following satisfies with probability $1 - O(n^{-1})$:*

$$\begin{aligned}\|\boldsymbol{\Phi} - \boldsymbol{\Psi}\boldsymbol{O}\| &\lesssim \tau,\\ \|\boldsymbol{\Phi}^{(i)} - \boldsymbol{\Psi}\boldsymbol{O}^{(i)}\| &\lesssim \tau,\\ 1 - \sigma_{\min}(\boldsymbol{\Psi}^\top\boldsymbol{\Phi}) &\lesssim \tau,\\ 1 - \sigma_{\min}(\boldsymbol{\Psi}^\top\boldsymbol{\Phi}^{(i)}) &\lesssim \tau,\end{aligned}$$

where $\boldsymbol{O} = \mathcal{P}(\boldsymbol{\Psi}^\top\boldsymbol{\Phi})$ and $\boldsymbol{O}^{(i)} = \mathcal{P}(\boldsymbol{\Psi}^\top\boldsymbol{\Phi}^{(i)})$.

The following lemma bounds the difference between $\boldsymbol{\Phi}$ and $\boldsymbol{\Phi}^{(i)}$. One can expect that such a difference is tiny since $\boldsymbol{A}$ only differs from $\boldsymbol{A}^{(i)}$ on its ith block column and ith block row.

LEMMA A.9. *Under the condition* (5.3)*, as n is sufficiently large, we have*

$$\begin{aligned}\|\boldsymbol{\Phi} - \boldsymbol{\Phi}^{(i)}\boldsymbol{O}^{(i)}\| &\lesssim \eta \max_j \|\boldsymbol{\Phi}_{j\cdot}\|,\\ \max_j \|\boldsymbol{\Phi}^{(i)}_{j\cdot}\| &\lesssim \max_j \|\boldsymbol{\Phi}_{j\cdot}\|\end{aligned}$$

with probability $1 - O(n^{-1})$, where $\boldsymbol{O}^{(i)} = \mathcal{P}((\boldsymbol{\Phi}^{(i)})^\top\boldsymbol{\Phi})$.

The following lemma further bounds the blockwise difference between $\boldsymbol{\Phi}^{(i)}$ and $\boldsymbol{\Psi}$.

LEMMA A.10. *Under the condition* (5.3)*, for a sufficiently large n, we have*

$$\begin{aligned}\max_j \|\boldsymbol{\Phi}^{(i)}_{j\cdot} - \boldsymbol{\Psi}_{j\cdot}\boldsymbol{S}^{(i)}\| &\lesssim \max_j \|\boldsymbol{\Phi}_{j\cdot}\|,\\ \max_j \|\boldsymbol{\Phi}_{j\cdot}\| &\geq \frac{1}{\sqrt{n}}\end{aligned}$$

with probability $1 - O(n^{-1})$, where $\boldsymbol{S}^{(i)} = \mathcal{P}(\boldsymbol{\Psi}^\top\boldsymbol{\Phi}^{(i)})$.

LEMMA A.11. *Under the condition* (5.3)*, for a sufficiently large n, we have*

$$\|\boldsymbol{S}^{(i)}\boldsymbol{O}^{(i)} - \boldsymbol{O}\| \lesssim \eta \max_j \|\boldsymbol{\Phi}_{j\cdot}\|,$$

where $\boldsymbol{O} = \mathcal{P}(\boldsymbol{\Psi}^\top\boldsymbol{\Phi})$, $\boldsymbol{O}^{(i)} = \mathcal{P}((\boldsymbol{\Phi}^{(i)})^\top\boldsymbol{\Phi})$, and $\boldsymbol{S}^{(i)} = \mathcal{P}(\boldsymbol{\Psi}^\top\boldsymbol{\Phi}^{(i)})$.

As we can see, most of the statistics in the previous lemmas involve $\max_j \|\boldsymbol{\Phi}_{j\cdot}\|$, and from Lemma A.9 we have $\max_j \|\boldsymbol{\Phi}_{j\cdot}\| \geq 1/\sqrt{n}$. Then in the following lemma we further show that $\max_j \|\boldsymbol{\Phi}_{j\cdot}\| = O(1/\sqrt{n})$. Therefore, this enables us to replace all $\max_j \|\boldsymbol{\Phi}_{j\cdot}\|$ in the previous bounds with $1/\sqrt{n}$.

LEMMA A.12. *Under the condition* (5.3)*, for a sufficiently large n, we have*

$$\max_j \|\boldsymbol{\Phi}_{j\cdot}\| \lesssim \frac{1}{\sqrt{n}},$$

with probability $1 - O(n^{-1})$.

Based on the results above, now we are able to prove Lemma 5.3 which bounds the difference between $\boldsymbol{\Phi}_{i\cdot}$ and its surrogate $f(\boldsymbol{\Psi}\boldsymbol{O})_{i\cdot}$.

Proof of Lemma 5.3. We start from (5.10) in our proof sketch section 5.2, which is given as

$$\|\boldsymbol{\Phi}_{i\cdot} - f(\boldsymbol{\Psi}\boldsymbol{O})_{i\cdot}\| \le \|\boldsymbol{\Lambda}^{-1}\| \left(\|[\mathbb{E}[\boldsymbol{A}]]_{i\cdot}(\boldsymbol{\Phi} - \boldsymbol{\Psi}\boldsymbol{O})\| + \|\boldsymbol{\Delta}_{i\cdot}(\boldsymbol{\Phi} - \boldsymbol{\Psi}\boldsymbol{O})\|\right), \tag{A.3}$$

where the three terms $\|\boldsymbol{\Lambda}^{-1}\|$, $\|[\mathbb{E}[\boldsymbol{A}]]_{i\cdot}(\boldsymbol{\Phi} - \boldsymbol{\Psi}\boldsymbol{O})\|$, and $\|\boldsymbol{\Delta}_{i\cdot}(\boldsymbol{\Phi} - \boldsymbol{\Psi}\boldsymbol{O})\|$ are bounded separately. For $\|\boldsymbol{\Lambda}^{-1}\|$, by definition $\|\boldsymbol{\Lambda}^{-1}\| = \sigma_{2d}^{-1}(\boldsymbol{A})$, and by applying Weyl's inequality (Theorem A.2) we have

$$\sigma_{2d}(\boldsymbol{A}) \ge \sigma_{2d}(\mathbb{E}[\boldsymbol{A}]) - \|\boldsymbol{\Delta}\| = \frac{pn}{2} - \|\boldsymbol{\Delta}\| = \Omega(pn)$$

w.h.p., where $\|\boldsymbol{\Delta}\|$ is bounded by Lemma A.6. This leads to

$$\|\boldsymbol{\Lambda}^{-1}\| \lesssim (pn)^{-1} \tag{A.4}$$

w.h.p. For $\|[\mathbb{E}[\boldsymbol{A}]]_{i\cdot}(\boldsymbol{\Phi} - \boldsymbol{\Psi}\boldsymbol{O})\|$, it satisfies

$$\begin{aligned}\|[\mathbb{E}[\boldsymbol{A}]]_{i\cdot}(\boldsymbol{\Phi} - \boldsymbol{\Psi}\boldsymbol{O})\| &\le \|[\mathbb{E}[\boldsymbol{A}]]_{i\cdot}\|\|\boldsymbol{\Phi} - \boldsymbol{\Psi}\boldsymbol{O}\| \lesssim p\sqrt{m}\cdot\frac{\sqrt{p(1-p)}+\sqrt{q}}{p\sqrt{n}}\\ &\overset{(a)}{\lesssim} p\sqrt{m}\cdot\underbrace{\frac{\sqrt{p(1-p)+q}}{p\sqrt{n}}}_{=\eta/\log(nd)} \le \frac{p\eta\sqrt{n}}{\log(nd)}\end{aligned}$$

w.h.p., where $\|\boldsymbol{\Phi} - \boldsymbol{\Psi}\boldsymbol{O}\|$ is bounded by Lemma A.8 and (a) uses the fact $\sqrt{x}+\sqrt{y} \le \sqrt{2(x+y)}$ for any $x, y \ge 0$. From (5.12), it remains to bound $\|\boldsymbol{\Delta}_{i\cdot}(\boldsymbol{\Phi} - \boldsymbol{\Psi}\boldsymbol{O})\|$ as

$$\begin{aligned}\|\boldsymbol{\Delta}_{i\cdot}(\boldsymbol{\Phi} - \boldsymbol{\Psi}\boldsymbol{O})\| &\le \underbrace{\|\boldsymbol{\Delta}_{i\cdot}(\boldsymbol{\Phi} - \boldsymbol{\Phi}^{(i)}\boldsymbol{O}^{(i)})\|}_{=:T_1} + \underbrace{\|\boldsymbol{\Delta}_{i\cdot}(\boldsymbol{\Phi}^{(i)} - \boldsymbol{\Psi}\boldsymbol{S}^{(i)})\|}_{=:T_2} + \underbrace{\|\boldsymbol{\Delta}_{i\cdot}\boldsymbol{\Psi}(\boldsymbol{S}^{(i)}\boldsymbol{O}^{(i)} - \boldsymbol{O})\|}_{=:T_3}\\ &= T_1 + T_2 + T_3,\end{aligned} \tag{A.5}$$

w.h.p. Here, T_1, T_2, and T_3 satisfy

$$\begin{aligned}T_1 &= \|\boldsymbol{\Delta}_{i\cdot}(\boldsymbol{\Phi} - \boldsymbol{\Phi}^{(i)}\boldsymbol{O}^{(i)})\| \le \|\boldsymbol{\Delta}_{i\cdot}\|\|\boldsymbol{\Phi} - \boldsymbol{\Phi}^{(i)}\boldsymbol{O}^{(i)}\|\\ &\lesssim (\sqrt{p(1-p)n} + \sqrt{qn})\cdot\eta\max_j\|\boldsymbol{\Phi}_{j\cdot}\| \overset{(a)}{\lesssim} \sqrt{(p(1-p)+q)n}\eta\max_j\|\boldsymbol{\Phi}_{j\cdot}\|,\\ T_2 &= \|\boldsymbol{\Delta}_{i\cdot}(\boldsymbol{\Phi}^{(i)} - \boldsymbol{\Psi}\boldsymbol{S}^{(i)})\| \lesssim \sqrt{(p(1-p)+q)n\log(nd)}\max_j\|\boldsymbol{\Phi}^{(i)}_{j\cdot} - \boldsymbol{\Psi}_{j\cdot}\boldsymbol{S}^{(i)}\|\\ &\lesssim \sqrt{(p(1-p)+q)n\log(nd)}\max_j\|\boldsymbol{\Phi}_{j\cdot}\|,\\ T_3 &= \|\boldsymbol{\Delta}_{i\cdot}\boldsymbol{\Psi}(\boldsymbol{S}^{(i)}\boldsymbol{O}^{(i)} - \boldsymbol{O})\| \le \|\boldsymbol{\Delta}_{i\cdot}\boldsymbol{\Psi}\|\|\boldsymbol{S}^{(i)}\boldsymbol{O}^{(i)} - \boldsymbol{O}\|\\ &\lesssim \sqrt{(p(1-p)+q)n\log(nd)}\max_j\|\boldsymbol{\Psi}_{j\cdot}\|\cdot\eta\max_j\|\boldsymbol{\Phi}_{j\cdot}\|\\ &\lesssim \sqrt{(p(1-p)+q)\log(nd)}\eta\max_j\|\boldsymbol{\Phi}_{j\cdot}\|.\end{aligned} \tag{A.6}$$

For T_1, $\|\boldsymbol{\Delta}_{i\cdot}\|$ and $\|\boldsymbol{\Phi} - \boldsymbol{\Phi}^{(i)}\boldsymbol{O}^{(i)}\|$ are bounded by Lemma A.6 and Lemma A.9, respectively; for T_2, $\|\boldsymbol{\Delta}_{i\cdot}(\boldsymbol{\Phi}^{(i)} - \boldsymbol{\Psi}\boldsymbol{S}^{(i)})\|$ is bounded by Lemma A.7 with $\boldsymbol{M} = \boldsymbol{\Phi}^{(i)} - \boldsymbol{\Psi}\boldsymbol{S}^{(i)}$, and $\max_j\|\boldsymbol{\Phi}^{(i)}_{j\cdot} - \boldsymbol{\Psi}_{j\cdot}\boldsymbol{S}^{(i)}\|$ is bounded by Lemma A.10; for T_3, $\|\boldsymbol{\Delta}_{i\cdot}\boldsymbol{\Psi}\|$ and $\|\boldsymbol{S}^{(i)}\boldsymbol{O}^{(i)} - \boldsymbol{O}\|$ are bounded by Lemma A.7 and Lemma A.11, respectively. As a result, one can see that T_2 is the dominant term among $\{T_i\}_{i=1}^3$ and (A.5) becomes

$$\|\boldsymbol{\Delta}_{i\cdot}(\boldsymbol{\Phi} - \boldsymbol{\Psi}\boldsymbol{O})\| \lesssim \sqrt{(p(1-p)+q)n\log(nd)}\max_j\|\boldsymbol{\Phi}_{j\cdot}\| = \eta pn\max_j\|\boldsymbol{\Phi}_{j\cdot}\|$$

w.h.p. Putting the above together into (A.3) yields

$$\begin{aligned}\|\boldsymbol{\Phi}_{i\cdot} - f(\boldsymbol{\Psi O})_{i\cdot}\| &\lesssim \|\boldsymbol{\Lambda}^{-1}\| \left(\|[\mathbb{E}[\boldsymbol{A}]]_{i\cdot}(\boldsymbol{\Phi} - \boldsymbol{\Psi O})\| + \|\boldsymbol{\Delta}_{i\cdot}(\boldsymbol{\Phi} - \boldsymbol{\Psi O})\|\right) \\ &\lesssim (pn)^{-1} \cdot \left(\frac{p\eta\sqrt{n}}{\log(nd)} + \eta pn \max_j \|\boldsymbol{\Phi}_{j\cdot}\|\right) \overset{(a)}{\lesssim} \eta \max_j \|\boldsymbol{\Phi}_{j\cdot}\|\end{aligned}$$

w.h.p., where (a) uses $\max_j \|\boldsymbol{\Phi}_{j\cdot}\| \geq 1/\sqrt{n}$ shown in the proof of Lemma A.10. ∎

Next, we prove Lemma 5.4 which bounds the difference between $\boldsymbol{\Phi}_{i\cdot}$ and $\boldsymbol{\Phi}_{j\cdot}$ for $\kappa(i) = \kappa(j)$.

Proof of Lemma 5.4. We start from (5.9) such that

$$\|\boldsymbol{\Phi}_{i\cdot} - \boldsymbol{\Phi}_{j\cdot}\| \leq \|\boldsymbol{\Phi}_{i\cdot} - f(\boldsymbol{\Psi O})_{i\cdot}\| + \|\boldsymbol{\Phi}_{j\cdot} - f(\boldsymbol{\Psi O})_{j\cdot}\| + \|f(\boldsymbol{\Psi O})_{i\cdot} - f(\boldsymbol{\Psi O})_{j\cdot}\|,$$

where $\|\boldsymbol{\Phi}_{i\cdot} - f(\boldsymbol{\Psi O})_{i\cdot}\|$ and $\|\boldsymbol{\Phi}_{j\cdot} - f(\boldsymbol{\Psi O})_{j\cdot}\|$ have been bounded by Lemma 5.3. For $\|f(\boldsymbol{\Psi O})_{i\cdot} - f(\boldsymbol{\Psi O})_{j\cdot}\|$, by definition

$$\begin{aligned}\|f(\boldsymbol{\Psi O})_{i\cdot} - f(\boldsymbol{\Psi O})_{j\cdot}\| &= \|(\boldsymbol{\Delta}_{i\cdot} - \boldsymbol{\Delta}_{j\cdot})\boldsymbol{\Psi O}\boldsymbol{\Lambda}^{-1}\| \leq \|(\boldsymbol{\Delta}_{i\cdot} - \boldsymbol{\Delta}_{j\cdot})\boldsymbol{\Psi}\|\|\boldsymbol{\Lambda}^{-1}\| \\ &\leq (\|\boldsymbol{\Delta}_{i\cdot}\boldsymbol{\Psi}\| + \|\boldsymbol{\Delta}_{j\cdot}\boldsymbol{\Psi}\|)\|\boldsymbol{\Lambda}^{-1}\| \lesssim \sqrt{(p(1-p)+q)\log(nd)} \cdot (pn)^{-1} \\ &= \frac{\eta}{\sqrt{n}} \overset{(a)}{\leq} \eta \max_j \|\boldsymbol{\Phi}_{j\cdot}\|\end{aligned} \tag{A.7}$$

w.h.p., where $\|\boldsymbol{\Delta}_{i\cdot}\boldsymbol{\Psi}\|$ and $\|\boldsymbol{\Delta}_{j\cdot}\boldsymbol{\Psi}\|$ are bounded by Lemma A.7, $\|\boldsymbol{\Lambda}^{-1}\|$ is bounded in (A.4), and (a) uses $\max_j \|\boldsymbol{\Phi}_{j\cdot}\| \geq 1/\sqrt{n}$ in Lemma A.10. Combining (A.7) with Lemma 5.3 yields

$$\|\boldsymbol{\Phi}_{i\cdot} - \boldsymbol{\Phi}_{j\cdot}\| \lesssim \eta \max_j \|\boldsymbol{\Phi}_{j\cdot}\| \tag{A.8}$$

w.h.p. This further leads to $\max_j \|\boldsymbol{\Phi}_{j\cdot}\| = O(1/\sqrt{n})$ given in Lemma A.12. Plugging this back into (A.8) completes the proof. ∎

Now we are ready to prove Theorem 5.1.

Proof of Theorem 5.1. By defining $\boldsymbol{N}_{C_1}$, $\boldsymbol{N}_i$, and $\boldsymbol{N}_{\boldsymbol{\Delta},i}$ as in (5.13) such that

$$\boldsymbol{N}_{C_1} := \begin{bmatrix} \boldsymbol{\Phi}_{1\cdot} - \boldsymbol{\Psi}_{1\cdot}\boldsymbol{O} \\ \vdots \\ \boldsymbol{\Phi}_{m\cdot} - \boldsymbol{\Psi}_{m\cdot}\boldsymbol{O} \end{bmatrix} = \underbrace{\begin{bmatrix} \boldsymbol{\Phi}_{i\cdot} - \boldsymbol{\Psi}_{1\cdot}\boldsymbol{O} \\ \vdots \\ \boldsymbol{\Phi}_{i\cdot} - \boldsymbol{\Psi}_{m\cdot}\boldsymbol{O} \end{bmatrix}}_{=:\boldsymbol{N}_i} + \underbrace{\begin{bmatrix} \boldsymbol{\Phi}_{1\cdot} - \boldsymbol{\Phi}_{i\cdot} \\ \vdots \\ \boldsymbol{\Phi}_{m\cdot} - \boldsymbol{\Phi}_{i\cdot} \end{bmatrix}}_{=:\boldsymbol{N}_{\boldsymbol{\Delta},i}} = \boldsymbol{N}_i + \boldsymbol{N}_{\boldsymbol{\Delta},i},$$

we have $\|\boldsymbol{N}_i\| = \sqrt{m}\|\boldsymbol{\Phi}_{i\cdot} - \boldsymbol{\Psi}_{i\cdot}\boldsymbol{O}\|$ since $\boldsymbol{\Psi}_{1\cdot} = \cdots = \boldsymbol{\Psi}_{m\cdot}$, and

$$\|\boldsymbol{N}_{\boldsymbol{\Delta},i}\| \overset{(a)}{\leq} \sqrt{\sum_{j=1}^{m} \|\boldsymbol{\Phi}_{i\cdot} - \boldsymbol{\Phi}_{j\cdot}\|^2} \overset{(b)}{\lesssim} \sqrt{m} \cdot \frac{\eta}{\sqrt{n}} = O(\eta)$$

w.h.p., where (a) holds by definition of the operator norm and (b) comes from Lemma 5.4. Then the following satisfies:

$$\begin{aligned}\|\boldsymbol{\Phi} - \boldsymbol{\Psi O}\| &\overset{(a)}{=} \max_{\|\boldsymbol{x}\|=1} \|[\boldsymbol{N}_{C_1}^\top, \boldsymbol{N}_{C_2}^\top]\boldsymbol{x}\| \geq \max_{\|\boldsymbol{y}\|=1} \|\boldsymbol{N}_{C_1}^\top \boldsymbol{y}\| \overset{(b)}{=} \|\boldsymbol{N}_{C_1}\| \\ &\overset{(c)}{\geq} \|\boldsymbol{N}_i\| - \|\boldsymbol{N}_{\boldsymbol{\Delta},i}\| = \sqrt{m}\|\boldsymbol{\Phi}_{i\cdot} - \boldsymbol{\Psi}_{i\cdot}\boldsymbol{O}\| - O(\eta)\end{aligned} \tag{A.9}$$

w.h.p., where both (a) and (b) hold by definition of the operator norm and (c) comes from the triangle inequality. On the other hand, from Lemma A.8 one can see that $\|\boldsymbol{\Phi} - \boldsymbol{\Psi}\boldsymbol{O}\| \lesssim \tau \lesssim \eta/\log(nd)$ w.h.p. Combining this with (A.9) yields

$$\|\boldsymbol{\Phi}_{i\cdot} - \boldsymbol{\Psi}_{i\cdot}\boldsymbol{O}\| = \frac{1}{\sqrt{m}}\left(O(\eta) + O\left(\frac{\eta}{\log(nd)}\right)\right) \lesssim \frac{\eta}{\sqrt{n}}$$

w.h.p. The bound $\|\boldsymbol{\Phi}_{i\cdot} - \boldsymbol{\Psi}_{i\cdot}\boldsymbol{O}\|$ for $i \in C_2$ is obtained in the same way as above. ∎

A.3. Proof of Theorem 5.2. In this section, we prove Theorem 5.2 which provides the performance guarantee of Algorithm 1. Again, we start by listing several important inequalities, where most of them directly come from Theorem 5.1. All the proofs of the lemmas are deferred to Appendix B.

To begin with, recall that Theorem 5.1 bounds the difference between the noisy eigenvector block $\boldsymbol{\Phi}_{i\cdot}$ and the clean one $\boldsymbol{\Psi}_{i\cdot}\boldsymbol{O}$; then the following lemma further bounds the difference between other statistics such as the polar factors $\mathcal{P}(\boldsymbol{\Phi}_{i\cdot})$ and $\mathcal{P}(\boldsymbol{\Psi}_{i\cdot}\boldsymbol{O})$.

LEMMA A.13. *Under the condition* (5.3) *and as n is sufficiently large, for any $i, j = 1, \ldots, n$ such that $\kappa(i) = \kappa(j)$, the following satisfy:*

$$\|\sigma_l(\boldsymbol{\Phi}_{i\cdot}) - \sqrt{2/n}\| \lesssim \eta/\sqrt{n}, \quad l = 1, \ldots, d, \tag{A.10}$$

$$\|\mathcal{P}(\boldsymbol{\Phi}_{i\cdot}) - \mathcal{P}(\boldsymbol{\Psi}_{i\cdot}\boldsymbol{O})\| \lesssim \eta, \tag{A.11}$$

$$\begin{aligned} \|\boldsymbol{\Phi}_{j\cdot}\mathcal{P}(\boldsymbol{\Phi}_{i\cdot})^\top - \boldsymbol{\Psi}_{j\cdot}\mathcal{P}(\boldsymbol{\Psi}_{i\cdot})^\top\| &\lesssim \eta/\sqrt{n}, \\ \|\mathcal{P}(\boldsymbol{\Phi}_{j\cdot}\mathcal{P}(\boldsymbol{\Phi}_{i\cdot})^\top) - \mathcal{P}(\boldsymbol{\Psi}_{j\cdot}\mathcal{P}(\boldsymbol{\Psi}_{i\cdot})^\top)\| &\lesssim \eta \end{aligned} \tag{A.12}$$

with probability $1 - O(n^{-1})$, where $\boldsymbol{O} = \mathcal{P}(\boldsymbol{\Psi}^\top\boldsymbol{\Phi})$.

Lemma A.13 is sufficient for showing the exact recovery of the cluster memberships by Algorithm 1. For showing the stable recovery of the orthogonal group elements $\{\boldsymbol{O}_i\}_{i=1}^n$, we need the following result.

LEMMA A.14. *Under the condition* (5.3) *and the assumption that $p_1 \in C_1$ is the first pivot selected by Algorithm* 1*, as n is sufficiently large, for $\boldsymbol{Q}_{\cdot 1}$ and $\boldsymbol{Q}_{\cdot 2}$ defined in* (5.16)*, the following satisfies for any $j \in C_2$:*

$$\begin{aligned} \|\boldsymbol{Q}_{\cdot 2} - \mathcal{P}(\boldsymbol{\Psi}_{j\cdot}\boldsymbol{O})^\top\bar{\boldsymbol{O}}_2\| &\lesssim \eta, \\ \|\boldsymbol{\Phi}_{j\cdot}\boldsymbol{Q}_{\cdot 2} - \boldsymbol{\Psi}_{j\cdot}\mathcal{P}(\boldsymbol{\Psi}_{j\cdot})^\top\bar{\boldsymbol{O}}_2\| &\lesssim \eta/\sqrt{n}, \\ \|\mathcal{P}(\boldsymbol{\Phi}_{j\cdot}\boldsymbol{Q}_{\cdot 2}) - \mathcal{P}(\boldsymbol{\Psi}_{j\cdot}\mathcal{P}(\boldsymbol{\Psi}_{j\cdot})^\top)\bar{\boldsymbol{O}}_2\| &\lesssim \eta \end{aligned} \tag{A.13}$$

with probability $1 - O(n^{-1})$, where $\boldsymbol{O} = \mathcal{P}(\boldsymbol{\Psi}^\top\boldsymbol{\Phi})$ and $\bar{\boldsymbol{O}}_2 = \mathcal{P}(\mathcal{P}(\boldsymbol{\Phi}_{j\cdot}\boldsymbol{O})\boldsymbol{Q}_{\cdot 2})$.

Proof of Theorem 5.2. Exact recovery of the cluster memberships. According to the blockwise CPQR in section 4.2, without loss of generality we assume $p_1 \in C_1$ is the first pivot column selected by Algorithm 1. Then for any node j, the two blocks $\boldsymbol{R}_{1j}$ and $\boldsymbol{R}_{2j}$ satisfy

$$\|\boldsymbol{R}_{1j}\|_{\mathrm{F}}^2 = \|\mathcal{P}(\boldsymbol{\Phi}_{p_1\cdot})(\boldsymbol{\Phi}_{j\cdot})^\top\|_{\mathrm{F}}^2, \quad \|\boldsymbol{R}_{2j}\|_{\mathrm{F}}^2 = \|(\boldsymbol{\Phi}_{j\cdot})^\top\|_{\mathrm{F}}^2 - \|\mathcal{P}(\boldsymbol{\Phi}_{p_1\cdot})(\boldsymbol{\Phi}_{j\cdot})^\top\|_{\mathrm{F}}^2,$$

which correspond to the projection of $(\boldsymbol{\Phi}_{j\cdot})^\top$ onto the column space $\mathcal{R}((\boldsymbol{\Phi}_{p_1\cdot})^\top)$ and the complement of $\mathcal{R}((\boldsymbol{\Phi}_{p_1\cdot})^\top)$, respectively. According to Algorithm 1, node j would be assigned to C_1 if $\|\boldsymbol{R}_{1j}\|_{\mathrm{F}} > \|\boldsymbol{R}_{2j}\|_{\mathrm{F}}$, which is equivalent to

$$\|\mathcal{P}(\boldsymbol{\Phi}_{p_1\cdot})(\boldsymbol{\Phi}_{j\cdot})^\top\|_{\mathrm{F}} > \frac{\sqrt{2}}{2}\|\boldsymbol{\Phi}_{j\cdot}\|_{\mathrm{F}}. \tag{A.14}$$

We first show that, for any $j \in C_1$, (A.14) is satisfied w.h.p. To this end, by using $\boldsymbol{\Psi}_{i\cdot}\boldsymbol{O}$ as a surrogate of $\boldsymbol{\Phi}_{i\cdot}$ for each node i, $\|\mathcal{P}(\boldsymbol{\Phi}_{p_1\cdot})(\boldsymbol{\Phi}_{j\cdot})^\top\|_{\mathrm{F}}$ can be bounded as

$$\begin{aligned}
&\|\mathcal{P}(\boldsymbol{\Phi}_{p_1\cdot})(\boldsymbol{\Phi}_{j\cdot})^\top\|_{\mathrm{F}} = \|\left[\mathcal{P}(\boldsymbol{\Psi}_{p_1\cdot}\boldsymbol{O}) + \mathcal{P}(\boldsymbol{\Phi}_{p_1\cdot}) - \mathcal{P}(\boldsymbol{\Psi}_{p_1\cdot}\boldsymbol{O})\right](\boldsymbol{\Phi}_{j\cdot})^\top\|_{\mathrm{F}} \\
&\geq \|\mathcal{P}(\boldsymbol{\Psi}_{p_1\cdot}\boldsymbol{O})(\boldsymbol{\Psi}_{j\cdot}\boldsymbol{O} + \boldsymbol{\Phi}_{j\cdot} - \boldsymbol{\Psi}_{j\cdot}\boldsymbol{O})^\top\|_{\mathrm{F}} - \|\left[\mathcal{P}(\boldsymbol{\Phi}_{p_1\cdot}) - \mathcal{P}(\boldsymbol{\Psi}_{p_1\cdot}\boldsymbol{O})\right](\boldsymbol{\Phi}_{j\cdot})^\top\|_{\mathrm{F}} \\
&\geq \underbrace{\|\mathcal{P}(\boldsymbol{\Psi}_{p_1\cdot})(\boldsymbol{\Psi}_{j\cdot})^\top\|_{\mathrm{F}}}_{=:T_1} - \underbrace{\|\mathcal{P}(\boldsymbol{\Psi}_{p_1\cdot})\|_{\mathrm{F}}\|\boldsymbol{\Phi}_{j\cdot} - \boldsymbol{\Psi}_{j\cdot}\boldsymbol{O}\|}_{=:T_2} - \underbrace{\|\mathcal{P}(\boldsymbol{\Phi}_{p_1\cdot}) - \mathcal{P}(\boldsymbol{\Psi}_{p_1\cdot}\boldsymbol{O})\|\|\boldsymbol{\Phi}_{j\cdot}\|_{\mathrm{F}}}_{=:T_3} \\
&= T_1 - T_2 - T_3,
\end{aligned} \tag{A.15}$$

where we use the fact $\|\boldsymbol{X}\boldsymbol{Y}\|_{\mathrm{F}} \leq \|\boldsymbol{X}\|\|\boldsymbol{Y}\|_{\mathrm{F}}$ for any $\boldsymbol{X}$ and $\boldsymbol{Y}$. To proceed, T_1, T_2, and T_3 can be bounded separately as

$$\begin{aligned}
&T_1 = \|\mathcal{P}(\boldsymbol{\Psi}_{p_1\cdot})(\boldsymbol{\Psi}_{j\cdot})^\top\|_{\mathrm{F}} = \sqrt{m}\|\boldsymbol{\Psi}_{p_1\cdot}(\boldsymbol{\Psi}_{j\cdot})^\top\|_{\mathrm{F}} = \sqrt{2d/n}, \\
&T_2 = \|\mathcal{P}(\boldsymbol{\Psi}_{p_1\cdot})\|_{\mathrm{F}}\|\boldsymbol{\Phi}_{j\cdot} - \boldsymbol{\Psi}_{j\cdot}\boldsymbol{O}\| \leq \sqrt{d}\cdot O\left(\eta/\sqrt{n}\right) \lesssim \eta\sqrt{d/n}, \\
&T_3 \leq \sqrt{d}\|\boldsymbol{\Phi}_{j\cdot}\|\|\mathcal{P}(\boldsymbol{\Phi}_{p_1\cdot}) - \mathcal{P}(\boldsymbol{\Psi}_{p_1\cdot}\boldsymbol{O})\| \lesssim \sqrt{d}\cdot\left(\sqrt{2/n} + O\left(\eta/\sqrt{n}\right)\right)\cdot\eta \lesssim \eta\sqrt{d/n}.
\end{aligned}$$

For T_1 we use $\boldsymbol{\Psi}_{p_1\cdot}(\boldsymbol{\Psi}_{j\cdot})^\top = \boldsymbol{I}_d/m$; for T_2 we bound $\|\boldsymbol{\Phi}_{j\cdot} - \boldsymbol{\Psi}_{j\cdot}\boldsymbol{O}\|$ by Theorem 5.1; for T_3, both $\|\boldsymbol{\Phi}_{j\cdot}\|$ and $\|\mathcal{P}(\boldsymbol{\Phi}_{p_1\cdot}) - \mathcal{P}(\boldsymbol{\Psi}_{p_1\cdot}\boldsymbol{O})\|$ are bounded by Lemma A.13. Plugging these into (A.15) yields

$$\|\mathcal{P}(\boldsymbol{\Phi}_{p_1\cdot})(\boldsymbol{\Phi}_{j\cdot})^\top\|_{\mathrm{F}} = \sqrt{\frac{2d}{n}} - O\left(\sqrt{\frac{d}{n}}\eta\right) \tag{A.16}$$

w.h.p. On the other hand, the RHS of (A.14) satisfies

$$\frac{\sqrt{2}}{2}\|\boldsymbol{\Phi}_{j\cdot}\|_{\mathrm{F}} \leq \frac{\sqrt{2d}}{2}\|\boldsymbol{\Phi}_{j\cdot}\| \leq \frac{\sqrt{2d}}{2}(\|\boldsymbol{\Psi}_{j\cdot}\boldsymbol{O}\| + \|\boldsymbol{\Phi}_{j\cdot} - \boldsymbol{\Psi}_{j\cdot}\boldsymbol{O}\|) = (1 + O(\eta))\sqrt{\frac{d}{n}}$$

w.h.p. Therefore, as η is sufficiently small, (A.14) is satisfied, and $j \in C_1$ is correctly assigned to C_1. On the other hand, for any $j \in C_2$, similar to (A.15) we have

$$\begin{aligned}
&\|\mathcal{P}(\boldsymbol{\Phi}_{p_1\cdot})(\boldsymbol{\Phi}_{j\cdot})^\top\|_{\mathrm{F}} = \|\left[\mathcal{P}(\boldsymbol{\Psi}_{p_1\cdot}\boldsymbol{O}) + \mathcal{P}(\boldsymbol{\Phi}_{p_1\cdot}) - \mathcal{P}(\boldsymbol{\Psi}_{p_1\cdot}\boldsymbol{O})\right](\boldsymbol{\Phi}_{j\cdot})^\top\|_{\mathrm{F}} \\
&\leq \|\mathcal{P}(\boldsymbol{\Psi}_{p_1\cdot}\boldsymbol{O})(\boldsymbol{\Psi}_{j\cdot}\boldsymbol{O} + \boldsymbol{\Phi}_{j\cdot} - \boldsymbol{\Psi}_{j\cdot}\boldsymbol{O})^\top\|_{\mathrm{F}} + \|\left[\mathcal{P}(\boldsymbol{\Phi}_{p_1\cdot}) - \mathcal{P}(\boldsymbol{\Psi}_{p_1\cdot}\boldsymbol{O})\right](\boldsymbol{\Phi}_{j\cdot})^\top\|_{\mathrm{F}} \\
&\leq \underbrace{\|\mathcal{P}(\boldsymbol{\Psi}_{p_1\cdot})(\boldsymbol{\Psi}_{j\cdot})^\top\|_{\mathrm{F}}}_{=0} + \underbrace{\|\mathcal{P}(\boldsymbol{\Psi}_{p_1\cdot})\|_{\mathrm{F}}\|\boldsymbol{\Phi}_{j\cdot} - \boldsymbol{\Psi}_{j\cdot}\boldsymbol{O}\|_2}_{=T_2} + \underbrace{\|\mathcal{P}(\boldsymbol{\Phi}_{p_1\cdot}) - \mathcal{P}(\boldsymbol{\Psi}_{p_1\cdot}\boldsymbol{O})\|_2\|\boldsymbol{\Phi}_{j\cdot}\|_{\mathrm{F}}}_{=T_3} \\
&= O\left(\eta\sqrt{d/n}\right)
\end{aligned}$$

w.h.p., where T_2 and T_3 are defined in (A.15). On the other hand,

$$\frac{\sqrt{2}}{2}\|\boldsymbol{\Phi}_{j\cdot}\|_{\mathrm{F}} = \frac{\sqrt{2}}{2}\sqrt{\sum_{l=1}^{d}\sigma_l^2(\boldsymbol{\Phi}_{j\cdot})} \geq \frac{\sqrt{2}}{2}\cdot\sqrt{d}\left(\sqrt{\frac{2}{n}} - O\left(\frac{\eta}{\sqrt{n}}\right)\right) = \sqrt{\frac{d}{n}} - O\left(\sqrt{\frac{d}{n}}\eta\right)$$

w.h.p., where $\sigma_l(\boldsymbol{\Phi}_{j\cdot})$ is bounded by Lemma A.13. Therefore, as η is small, the inequality in (A.14) does not hold, and node j is assigned to C_2. This leads to the exact recovery of the cluster memberships.

• *Stable recovery of orthogonal transformations.* To bound the estimation error of $\{\boldsymbol{O}_i\}_{i=1}^n$, recall the orthogonal matrix $\boldsymbol{Q}$ from the blockwise CPQR in (4.1); then in the case of two clusters, it can be written as $\boldsymbol{Q}=[\boldsymbol{Q}_{\cdot 1},\boldsymbol{Q}_{\cdot 2}]$, where $\boldsymbol{Q}_{\cdot 1},\boldsymbol{Q}_{\cdot 2}\in\mathbb{R}^{2d\times d}$. We follow the previous assumption that $p_1\in C_1$ is the first pivot; then $\boldsymbol{Q}_{\cdot 1}\in\mathbb{R}^{2d\times d}$ is the polar factor of $(\boldsymbol{\Phi}_{p_1\cdot})^\top$ up to some orthogonal transformation, and $\boldsymbol{Q}_{\cdot 2}\in\mathbb{R}^{2d\times d}$ is orthogonal to $\boldsymbol{Q}_{\cdot 1}$. As a result, for any node $j\in C_1$ we have

$$\boldsymbol{R}_{1j}=(\boldsymbol{Q}_{\cdot 1})^\top(\boldsymbol{\Phi}_{j\cdot})^\top,\quad \boldsymbol{Q}_{\cdot 1}=\mathcal{P}(\boldsymbol{\Phi}_{p_1\cdot}^\top)\bar{\boldsymbol{O}}_1, \tag{A.17}$$

where $\bar{\boldsymbol{O}}_1\in\mathbb{R}^{d\times d}$ is some orthogonal matrix. Then, according to (4.2), our estimation $\hat{\boldsymbol{O}}_j$ is given as $\hat{\boldsymbol{O}}_j=\mathcal{P}(\boldsymbol{R}_{1j})^\top=\mathcal{P}(\boldsymbol{\Phi}_{j\cdot}\mathcal{P}(\boldsymbol{\Phi}_{p_1\cdot})^\top)\bar{\boldsymbol{O}}_1$. Meanwhile, the ground truth $\boldsymbol{O}_j$ satisfies $\boldsymbol{O}_j=\mathcal{P}(\boldsymbol{\Psi}_{j\cdot}\mathcal{P}(\boldsymbol{\Psi}_{p_1\cdot})^\top)=\boldsymbol{I}_d$ by assumption; then the estimation error of $\boldsymbol{O}_j$ can be bounded as

$$\|\hat{\boldsymbol{O}}_j-\boldsymbol{O}_j\bar{\boldsymbol{O}}_1\|=\|\hat{\boldsymbol{O}}_j\bar{\boldsymbol{O}}_1^\top-\boldsymbol{O}_j\|=\|\mathcal{P}(\boldsymbol{\Phi}_{j\cdot}\mathcal{P}(\boldsymbol{\Phi}_{i\cdot})^\top)-\mathcal{P}(\boldsymbol{\Psi}_{j\cdot}\mathcal{P}(\boldsymbol{\Psi}_{i\cdot})^\top)\|\overset{(a)}{\lesssim}\eta$$

w.h.p., where (a) comes from Lemma A.13. This completes the proof for $j\in C_1$. Next, we check $\hat{\boldsymbol{O}}_j$ for $j\in C_2$; similar to (A.17) we have

$$\boldsymbol{R}_{2j}=(\boldsymbol{Q}_{\cdot 2})^\top(\boldsymbol{\Phi}_{j\cdot})^\top,\quad (\boldsymbol{Q}_{\cdot 1})^\top\boldsymbol{Q}_{\cdot 2}=\boldsymbol{0}. \tag{A.18}$$

Also the ground truth $\boldsymbol{O}_j$ satisfies $\boldsymbol{O}_j=\mathcal{P}(\boldsymbol{\Psi}_{j\cdot}\mathcal{P}(\boldsymbol{\Psi}_{i\cdot})^\top)=\boldsymbol{I}_d$. Then our estimation $\hat{\boldsymbol{O}}_j$ is given as $\hat{\boldsymbol{O}}_j=\mathcal{P}(\boldsymbol{R}_{2j})^\top=\mathcal{P}(\boldsymbol{\Phi}_{j\cdot}\boldsymbol{Q}_{\cdot 2})$. Furthermore, by defining $\bar{\boldsymbol{O}}_2=\mathcal{P}(\mathcal{P}(\boldsymbol{\Phi}_{j\cdot}\boldsymbol{O})\boldsymbol{Q}_{\cdot 2})$, where $\boldsymbol{O}=\mathcal{P}(\boldsymbol{\Psi}^\top\boldsymbol{\Phi})$, we have

$$\|\hat{\boldsymbol{O}}_j-\boldsymbol{O}_j\bar{\boldsymbol{O}}_2\|=\|\mathcal{P}(\boldsymbol{\Phi}_{j\cdot}\boldsymbol{Q}_{\cdot 2})-\mathcal{P}(\boldsymbol{\Psi}_{j\cdot}\mathcal{P}(\boldsymbol{\Psi}_{j\cdot})^\top)\bar{\boldsymbol{O}}_2\|\overset{(a)}{\lesssim}\eta$$

w.h.p., where (a) comes from (A.13) in Lemma A.14. This completes the proof. ∎

Appendix B. Proof of the Lemmas in Appendix A.

B.1. Proof of Lemma A.6.

The proof relies on the following two theorems.

THEOREM B.1. *Let $\boldsymbol{A}\in\mathbb{R}^{n\times n}$ be a symmetric matrix whose entries a_{ij} for $1\le i<j\le n$ are independent and identically distributed (i.i.d.) and satisfy*

$$a_{ij}=\begin{cases}1-p, & \text{w.p. } p,\\ -p, & \text{w.p. } 1-p.\end{cases}$$

Then, for any $c>0$, there exists constants $c_1,c_2>0$ such that

$$\|\boldsymbol{A}\|\le c_1\sqrt{p(1-p)n}+c_2\sqrt{\log n}$$

with probability $1-n^{-c}$.

Proof. We resort to the moment method which is commonly used in the random matrix theory (see, e.g., [5, 67]). The idea is to bound $\mathbb{E}[\|\boldsymbol{A}\|^{2k}]$ for any $k\in\mathbb{N}$ and apply Markov inequality for getting a tail bound. We start by denoting

$$\sigma_k = \left(\sum_{i=1}^{n} \left(\sum_{j=1}^{n} \mathbb{E}\left[a_{ij}^2\right] \right)^k \right)^{1/2k} = n^{1/2k} \sqrt{np(1-p)},$$
$$\sigma_k^* = \left(\sum_{i=1}^{n} \sum_{j=1}^{n} \|a_{ij}\|_\infty^{2k} \right)^{1/2k} \leq n^{1/k}.$$

Then by applying [49, Theorem 4.8], we obtain

$$\mathbb{E}\left[\|\boldsymbol{A}\|^{2k}\right]^{1/2k} \leq 2\sigma_k + C\sqrt{k}\sigma_k^* \leq 2n^{1/2k}\sqrt{np(1-p)} + C\sqrt{k}n^{1/k}$$

for some universal constant $C > 0$. By further setting $k = \lceil \gamma \log n \rceil$ for some constant $\gamma > 0$, we have

$$\mathbb{E}\left[\|\boldsymbol{A}\|^{2k}\right]^{1/2k} \leq 2e^{1/2\gamma}\sqrt{np(1-p)} + C'\sqrt{\log n}$$

for some $C' > 0$. To get a tail bound, by Markov inequality [10] we obtain

$$\begin{aligned}\mathbb{P}\{\|\boldsymbol{A}\| \geq t\} &= \mathbb{P}\{\|\boldsymbol{A}\|^{2k} \geq t^{2k}\} \leq t^{-2k}\mathbb{E}\left[\|\boldsymbol{A}\|^{2k}\right] \\ &\leq \left(\frac{2e^{1/2\gamma}\sqrt{np(1-p)} + C'\sqrt{\log n}}{t} \right)^{2\lceil \gamma \log n \rceil}\end{aligned}$$

for any $t > 0$. By setting $\gamma = \frac{c}{2} + 1$, for any constant $c > 0$, we can identify some constants $c_1, c_2 > 0$ such that $t = c_1\sqrt{np(1-p)} + c_2\sqrt{\log n}$ and $\mathbb{P}\{\|\boldsymbol{A}\| \geq t\} \leq n^{-c}$, which completes the proof. ∎

THEOREM B.2. *Let $\boldsymbol{S} \in \mathbb{R}^{n_1 d \times n_2 d}$ be an $n_1 \times n_2$ random block matrix where each block $\boldsymbol{S}_{ij}$ is i.i.d. and satisfies*

$$\boldsymbol{S}_{ij} = \begin{cases} \boldsymbol{O}_{ij} & \text{w.p. } q, \\ \boldsymbol{0} & \text{w.p. } 1-q, \end{cases}$$

where $\boldsymbol{O}_{ij}$ is uniformly drawn from $\mathrm{O}(d)$. Let $n = n_1 + n_2$. Then, for any $c > 0$, there exists $c_1, c_2 > 0$ such that

$$\|\boldsymbol{S}\| \leq c_1(\sqrt{qn_1} + \sqrt{qn_2}) + c_2\sqrt{\log n}$$

with probability $1 - n^{-c}$.

Proof. The proof is similar to the one of [28, Theorem A.7], with the only difference on the orthogonal group $\mathrm{O}(d)$ rather than $\mathrm{SO}(d)$ considered in [28]. ∎

Proof of Lemma A.6. By definition, $\boldsymbol{\Delta}$ can be written into four blocks as

$$\boldsymbol{\Delta} = \begin{bmatrix} \widetilde{\boldsymbol{\Delta}}_{11} & \widetilde{\boldsymbol{\Delta}}_{12} \\ \widetilde{\boldsymbol{\Delta}}_{12}^\top & \widetilde{\boldsymbol{\Delta}}_{22} \end{bmatrix} = \begin{bmatrix} \widetilde{\boldsymbol{\Delta}}_{11} & \boldsymbol{0} \\ \boldsymbol{0} & \widetilde{\boldsymbol{\Delta}}_{22} \end{bmatrix} + \begin{bmatrix} \boldsymbol{0} & \widetilde{\boldsymbol{\Delta}}_{12} \\ \widetilde{\boldsymbol{\Delta}}_{12}^\top & \boldsymbol{0} \end{bmatrix} =: \boldsymbol{\Delta}_{\text{in}} + \boldsymbol{\Delta}_{\text{out}},$$

where $\widetilde{\boldsymbol{\Delta}}_{11}, \widetilde{\boldsymbol{\Delta}}_{22} \in \mathbb{R}^{md \times md}$ correspond to the two clusters C_1, C_2, respectively. By using the fact $\|\boldsymbol{\Delta}_{\text{out}}\| = \|\widetilde{\boldsymbol{\Delta}}_{12}\|$ and the triangle inequality that $\|\boldsymbol{\Delta}_{\text{in}}\| \leq \|\widetilde{\boldsymbol{\Delta}}_{11}\| + \|\widetilde{\boldsymbol{\Delta}}_{22}\|$ and $\|\boldsymbol{\Delta}\| \leq \|\boldsymbol{\Delta}_{\text{in}}\| + \|\boldsymbol{\Delta}_{\text{out}}\|$, we have $\|\boldsymbol{\Delta}\| \leq \|\widetilde{\boldsymbol{\Delta}}_{11}\| + \|\widetilde{\boldsymbol{\Delta}}_{22}\| + \|\widetilde{\boldsymbol{\Delta}}_{12}\|$. For $\|\widetilde{\boldsymbol{\Delta}}_{11}\|$, by definition in (A.2) we can denote $\boldsymbol{\Delta}_{ij} = r_{ij}\boldsymbol{I}_d$, where r_{ij} is a random variable such that $\mathbb{P}\{r_{ij} = 1-p\} = p$ and $\mathbb{P}\{r_{ij} = -p\} = 1-p$. Then let $\boldsymbol{E}_1 \in \mathbb{R}^{m \times m}$

be a matrix that contains all r_{ij} for $i,j \in C_1$ such that $r_{ij} = r_{ji}$ and $r_{ii} = 0$, and one can see that $\widetilde{\boldsymbol{\Delta}}_{11} = \boldsymbol{E}_1 \otimes \boldsymbol{I}_d$, where $\otimes$ denotes the Kronecker product, and further

$$\|\widetilde{\boldsymbol{\Delta}}_{11}\| = \|\boldsymbol{E}_1 \otimes \boldsymbol{I}_d\| = \|\boldsymbol{E}_1\|.$$

Then from Theorem B.1 we get $\|\widetilde{\boldsymbol{\Delta}}_{11}\| \lesssim \sqrt{p(1-p)n}$ w.h.p., where the residual term $c_2\sqrt{\log m}$ is absorbed owing to the assumption $p = \Omega(\log n/n)$ and $1-p = \Omega(\log n/n)$. Similarly $\|\widetilde{\boldsymbol{\Delta}}_{22}\|$ is bounded as $\|\widetilde{\boldsymbol{\Delta}}_{22}\| \lesssim \sqrt{p(1-p)n}$ w.h.p. For $\|\widetilde{\boldsymbol{\Delta}}_{12}\|$, applying Theorem B.2 gives $\|\widetilde{\boldsymbol{\Delta}}_{12}\| \lesssim \sqrt{qn}$ w.h.p. Putting all together yields the bound on $\|\boldsymbol{\Delta}\|$.

For the block row $\boldsymbol{\Delta}_{i\cdot}$, let us consider another matrix $\boldsymbol{\Delta}' \in \mathbb{R}^{nd\times nd}$ which is all zero, but only the ith block row is equal to $\boldsymbol{\Delta}_{i\cdot}$ (i.e., $\boldsymbol{\Delta}'_{i\cdot} = \boldsymbol{\Delta}_{i\cdot}$). Notice that $\boldsymbol{\Delta}'$ is of the same size as $\boldsymbol{\Delta}$; then by the definition of the operator norm $\|\boldsymbol{X}\| = \max_{\|v\|_2=1}\|\boldsymbol{X}v\|_2$, one can see that $\|\boldsymbol{\Delta}_{i\cdot}\| = \|\boldsymbol{\Delta}'\| \le \|\boldsymbol{\Delta}\|$.

For $\boldsymbol{\Delta}^{(i)}$, let us consider another matrix $\boldsymbol{\Delta}^{(i),\mathrm{col}}$ which is identical to $\boldsymbol{\Delta}$, but only the ith block column is all zero, i.e., $\boldsymbol{\Delta}^{(i),\mathrm{col}}_{\cdot i} = \boldsymbol{0}$. Then by definition of the operator norm we have $\|\boldsymbol{\Delta}^{(i),\mathrm{col}}\| \le \|\boldsymbol{\Delta}\|$. Furthermore, since $\boldsymbol{\Delta}^{(i)}$ is identical to $\boldsymbol{\Delta}^{(i),\mathrm{col}}_{\cdot i}$, but only the ith block row is all zero, then by following the same idea as above we have $\|\boldsymbol{\Delta}^{(i)}\| \le \|\boldsymbol{\Delta}^{(i),\mathrm{col}}_{\cdot i}\|$, which completes the proof. ∎

B.2. Proof of Lemma A.7. The proof relies on the following inequality.

THEOREM B.3 (matrix Bernstein inequality [69]). *Let $\boldsymbol{X}_1,\dots,\boldsymbol{X}_n \in \mathbb{R}^{d_1\times d_2}$ be an independent matrix such that $\mathbb{E}[\boldsymbol{X}_i] = \boldsymbol{0}$ and $\|\boldsymbol{X}_i\| \le L$ for $i = 1,\dots,n$. Let $\boldsymbol{Z} = \sum_{i=1}^n \boldsymbol{X}_i$ and $v(\boldsymbol{Z}) := \max\{\|\mathbb{E}(\boldsymbol{Z}\boldsymbol{Z}^\top)\|, \|\mathbb{E}(\boldsymbol{Z}^\top\boldsymbol{Z})\|\}$. Then for any $t > 0$,*

$$\mathbb{P}\{\|\boldsymbol{Z}\| \ge t\} \le (d_1+d_2)\exp\left(\frac{-t^2/2}{v(\boldsymbol{Z}) + Lt/3}\right). \tag{B.1}$$

In other words, for any $c > 0$,

$$\|\boldsymbol{Z}\| \le \sqrt{2cv(\boldsymbol{Z})\log(n(d_1+d_2))} + \frac{2cL\log(n(d_1+d_2))}{3} \tag{B.2}$$

with probability $1 - n^{-c}$.

To obtain (B.2), by setting the RHS of (B.1) to be n^{-c} for some $c > 0$ we get

$$t = \sqrt{2v(\boldsymbol{Z})\gamma}\left(\sqrt{1 + \frac{L^2\gamma}{18v(\boldsymbol{Z})}} + \sqrt{\frac{L^2\gamma}{18v(\boldsymbol{Z})}}\right), \quad \gamma := c\log n + \log(d_1+d_2).$$

Then when $c > 1$ it satisfies $\gamma \le c\log(n(d_1+d_2))$ and $t \le \sqrt{2v(\boldsymbol{Z})\gamma}(1 + 2\sqrt{\frac{L^2\gamma}{18v(\boldsymbol{Z})}}) \le \sqrt{2v(\boldsymbol{Z})\gamma} + \frac{2L\gamma}{3}$, which leads to (B.2).

Proof of Lemma A.7. We directly apply Theorem B.3 by letting $\boldsymbol{X}_j = \boldsymbol{\Delta}_{ij}\boldsymbol{M}_{j\cdot}$. Clearly, $\mathbb{E}[\boldsymbol{X}_j] = 0$ and $\|\boldsymbol{X}_j\| \le \|\boldsymbol{\Delta}_{ij}\|\|\boldsymbol{M}_{j\cdot}\| \le \max_j\|\boldsymbol{M}_{j\cdot}\|$ for $j = 1,\dots,n$, where $\|\boldsymbol{\Delta}_{ij}\| \le 1$ by definition in (A.2). Then $L = \max_j\|\boldsymbol{M}_{j\cdot}\|$. For $v(\boldsymbol{Z})$ in Theorem B.3, where $\boldsymbol{Z} := \sum_{j=1}^n \boldsymbol{X}_j$, we have

$$\begin{aligned}\mathbb{E}(\boldsymbol{Z}\boldsymbol{Z}^\top) &= \sum_{j=1}^n \mathbb{E}\left[\boldsymbol{\Delta}_{ij}\boldsymbol{M}_{j\cdot}\boldsymbol{M}_{j\cdot}^\top\boldsymbol{\Delta}_{ij}^\top\right] \overset{(a)}{\preceq} \sum_{j=1}^n \mathbb{E}\left[\boldsymbol{\Delta}_{ij}\boldsymbol{\Delta}_{ij}^\top\right]\|\boldsymbol{M}_{j\cdot}\|_2^2 \\ &\preceq \sum_{j=1}^n \mathbb{E}\left[\boldsymbol{\Delta}_{ij}\boldsymbol{\Delta}_{ij}^\top\right]\max_j\|\boldsymbol{M}_{j\cdot}\|_2^2 \overset{(b)}{=} \left[\frac{n}{2}(p(1-p)+q)\max_j\|\boldsymbol{M}_{j\cdot}\|_2^2\right]\boldsymbol{I}_d,\end{aligned}$$

where (a) holds since $\boldsymbol{M}_{j\cdot}\boldsymbol{M}_{j\cdot}^\top \preceq \|\boldsymbol{M}_{j\cdot}\|^2\boldsymbol{I}_d$ and (b) holds by (A.2). Similarly, one can see that $\mathbb{E}(\boldsymbol{Z}^\top\boldsymbol{Z}) = [\frac{n}{2}(p(1-p)+q)\max_j\|\boldsymbol{M}_{j\cdot}\|_2^2]\boldsymbol{I}_d$. Then, by definition, $v(\boldsymbol{Z}) = \frac{n}{2}(p(1-p)+q)\max_j\|\boldsymbol{M}_{j\cdot}\|^2$. In this way, Theorem B.3 gives us

$$\begin{aligned}\|\boldsymbol{\Delta}_{i\cdot}\boldsymbol{M}\| &\le \sqrt{2c(p(1-p)+q)n\log(2nd)}\max_j\|\boldsymbol{M}_{j\cdot}\| + \frac{2c}{3}\log(2nd)\max_j\|\boldsymbol{M}_{j\cdot}\| \\ &\lesssim \sqrt{(p(1-p)+q)n\log(nd)}\max_j\|\boldsymbol{M}_{j\cdot}\|\end{aligned}$$

with probability $1-n^{-c}$ for $c>0$, where the last inequality holds because of the assumption $p(1-p)+q=\Omega(\log n/n)$, and then the second term $\frac{2c}{3}\log(2nd)\max_j\|\boldsymbol{M}_{j\cdot}\|$ is grouped into the first one. This completes the proof. ∎

B.3. Proof of the lemmas in Appendix A.2. *Proof of Lemma* A.8. For $\|\boldsymbol{\Phi}-\boldsymbol{\Psi}\boldsymbol{O}\|$, applying Theorem A.3 with (A.1) yields

$$\|\boldsymbol{\Phi}-\boldsymbol{\Psi}\boldsymbol{O}\| \lesssim \frac{\|\boldsymbol{\Delta}\boldsymbol{\Psi}\|}{\delta} \le \frac{\|\boldsymbol{\Delta}\|\|\boldsymbol{\Psi}\|}{\delta},$$

where $\delta = |\sigma_{2d}(\mathbb{E}[\boldsymbol{A}]) - \sigma_{2d+1}(\boldsymbol{A})|$. Here, $\sigma_{2d}(\mathbb{E}[\boldsymbol{A}]) = pn/2$ by definition, and $\sigma_{2d+1}(\boldsymbol{A}) \le \sigma_{2d+1}(\mathbb{E}[\boldsymbol{A}]) + \|\boldsymbol{\Delta}\| = \|\boldsymbol{\Delta}\|$ by Theorem A.2. Then applying Lemma A.6 gives $\delta \ge \sigma_{2d}(\mathbb{E}[\boldsymbol{A}]) - \sigma_{2d+1}(\boldsymbol{A}) = \Omega(pn)$ w.h.p. Also, by definition $\|\boldsymbol{\Phi}\| = 1$. Combining these gives $\|\boldsymbol{\Phi}-\boldsymbol{\Psi}\boldsymbol{O}\| \lesssim \tau$ w.h.p. For $1-\sigma_{\min}(\boldsymbol{\Psi}^\top\boldsymbol{\Phi})$, notice that

$$\|\boldsymbol{\Psi}^\top\boldsymbol{\Phi}-\boldsymbol{O}\| = \|\boldsymbol{\Psi}^\top(\boldsymbol{\Phi}-\boldsymbol{\Psi}\boldsymbol{O})\| \le \|\boldsymbol{\Psi}\|\|\boldsymbol{\Phi}-\boldsymbol{\Psi}\boldsymbol{O}\| \lesssim \tau$$

w.h.p.; this leads to

$$\sigma_{\min}(\boldsymbol{O}) - \sigma_{\min}(\boldsymbol{\Psi}^\top\boldsymbol{\Phi}) = 1-\sigma_{\min}(\boldsymbol{\Psi}^\top\boldsymbol{\Phi}) \overset{(a)}{\le} \|\boldsymbol{\Psi}^\top\boldsymbol{\Phi}-\boldsymbol{O}\| \lesssim \tau$$

w.h.p., where (a) comes from Theorem A.2. The bounds on $\|\boldsymbol{\Phi}^{(i)}-\boldsymbol{\Psi}\boldsymbol{O}^{(i)}\|$ and $1-\sigma_{\min}(\boldsymbol{\Psi}^\top\boldsymbol{\Phi}^{(i)})$ are derived in a similar manner; therefore, we do not repeat. ∎

Proof of Lemma A.9. Applying Theorem A.3 on $\|\boldsymbol{\Phi}-\boldsymbol{\Phi}^{(i)}\boldsymbol{O}^{(i)}\|$ yields

$$\|\boldsymbol{\Phi}-\boldsymbol{\Phi}^{(i)}\boldsymbol{O}^{(i)}\| \le \frac{\|(\boldsymbol{A}-\boldsymbol{A}^{(i)})\boldsymbol{\Phi}^{(i)}\|}{\delta},$$

where $\delta = |\sigma_{2d}(\boldsymbol{A}) - \sigma_{2d+1}(\boldsymbol{A}^{(i)})|$. To bound δ, similar to the proof in Lemma A.8, we have $\sigma_{2d}(\boldsymbol{A}) \ge \sigma_{2d}(\mathbb{E}[\boldsymbol{A}]) - \|\boldsymbol{\Delta}\|$ and $\sigma_{2d+1}(\boldsymbol{A}^{(i)}) \le \sigma_{2d+1}(\mathbb{E}[\boldsymbol{A}]) + \|\boldsymbol{\Delta}^{(i)}\| = \|\boldsymbol{\Delta}^{(i)}\|$ w.h.p.; then applying Lemma A.6 gives $\delta \ge \sigma_{2d}(\mathbb{E}[\boldsymbol{A}]) - 2\|\boldsymbol{\Delta}\| = \Omega(pn)$ w.h.p. To bound $\|(\boldsymbol{A}-\boldsymbol{A}^{(i)})\boldsymbol{\Phi}^{(i)}\|$, notice that $\boldsymbol{A}-\boldsymbol{A}^{(i)} = \boldsymbol{\Delta}-\boldsymbol{\Delta}^{(i)}$ which is only nonzero on the ith block row and the ith block column; then

$$\begin{aligned}\|(\boldsymbol{A}-\boldsymbol{A}^{(i)})\boldsymbol{\Phi}^{(i)}\| = \|(\boldsymbol{\Delta}-\boldsymbol{\Delta}^{(i)})\boldsymbol{\Phi}^{(i)}\| &\overset{(a)}{\le} \|\boldsymbol{\Delta}_{i\cdot}\boldsymbol{\Phi}^{(i)}\| + \|\boldsymbol{\Delta}_{i\cdot}^\top\boldsymbol{\Phi}_{i\cdot}^{(i)}\| \\ &\le \|\boldsymbol{\Delta}_{i\cdot}\boldsymbol{\Phi}^{(i)}\| + \|\boldsymbol{\Delta}_{i\cdot}\|\|\boldsymbol{\Phi}_{i\cdot}^{(i)}\|,\end{aligned}$$

where (a) holds by separating the ith block row and block column with the triangle inequality following. To proceed, since $\boldsymbol{\Delta}_{i\cdot}$ is independent of $\boldsymbol{\Phi}^{(i)}$, $\|\boldsymbol{\Delta}_{i\cdot}\boldsymbol{\Phi}^{(i)}\|$ is bounded by Lemma A.7 and yields

$$\|\boldsymbol{\Delta}_{i\cdot}\boldsymbol{\Phi}^{(i)}\| \lesssim \sqrt{(p(1-p)+q)n\log(nd)}\max_j\|\boldsymbol{\Phi}_{j\cdot}^{(i)}\|$$

w.h.p. Also, $\|\boldsymbol{\Delta}_{i\cdot}\|\|\boldsymbol{\Phi}_{i\cdot}^{(i)}\| \le (\sqrt{p(1-p)n}+\sqrt{qn})\max_j \|\boldsymbol{\Phi}_{j\cdot}^{(i)}\|$ w.h.p., where $\|\boldsymbol{\Delta}_{i\cdot}\|$ is bounded by Lemma A.6. Combining the result above gives

$$\|\boldsymbol{\Phi}-\boldsymbol{\Phi}^{(i)}\boldsymbol{O}^{(i)}\| \lesssim \eta \max_j \|\boldsymbol{\Phi}_{j\cdot}^{(i)}\| \tag{B.3}$$

w.h.p. Next we show that $\max_j \|\boldsymbol{\Phi}_{j\cdot}^{(i)}\| \lesssim \max_j \|\boldsymbol{\Phi}_{j\cdot}\|$. Let j' be the index such that $\|\boldsymbol{\Phi}_{j'\cdot}^{(i)}\| = \max_j \|\boldsymbol{\Phi}_{j\cdot}^{(i)}\|$; then we have

$$\|\boldsymbol{\Phi}_{j'\cdot}^{(i)}\| - \|\boldsymbol{\Phi}_{j'\cdot}\| \overset{(a)}{\le} \|\boldsymbol{\Phi}_{j'\cdot}-\boldsymbol{\Phi}_{j'\cdot}^{(i)}\boldsymbol{O}^{(i)}\| \overset{(b)}{\le} \|\boldsymbol{\Phi}-\boldsymbol{\Phi}^{(i)}\boldsymbol{O}^{(i)}\| \lesssim \eta\max_j\|\boldsymbol{\Phi}_{j\cdot}^{(i)}\| = \eta\|\boldsymbol{\Phi}_{j'\cdot}^{(i)}\| \tag{B.4}$$

w.h.p., where (a) comes from the triangle inequality, and (b) holds since $\boldsymbol{\Phi}_{j'\cdot}-\boldsymbol{\Phi}_{j'\cdot}^{(i)}\boldsymbol{O}^{(i)}$ is the j'th block row of $\boldsymbol{\Phi}-\boldsymbol{\Phi}^{(i)}\boldsymbol{O}^{(i)}$. This implies if the condition (5.3) that $\eta \le c_0$ holds for a sufficiently small c_0, then $\max_j \|\boldsymbol{\Phi}_{j\cdot}^{(i)}\| = \|\boldsymbol{\Phi}_{j'\cdot}^{(i)}\| \lesssim \|\boldsymbol{\Phi}_{j'\cdot}\| \le \max_j \|\boldsymbol{\Phi}_{j\cdot}\|$ w.h.p. Plugging this into (B.3) completes the proof. ∎

Proof of Lemma A.10. From the triangle inequality we obtain

$$\begin{aligned}\max_j \|\boldsymbol{\Phi}_{j\cdot}^{(i)}-\boldsymbol{\Psi}_{j\cdot}\boldsymbol{S}^{(i)}\| &\le \max_j(\|\boldsymbol{\Phi}_{j\cdot}^{(i)}\|+\|\boldsymbol{\Psi}_{j\cdot}\boldsymbol{S}^{(i)}\|) \le \max_j\|\boldsymbol{\Phi}_{j\cdot}^{(i)}\| + \max_j\|\boldsymbol{\Psi}_{j\cdot}\| \\ &\overset{(a)}{\lesssim} \max_j\|\boldsymbol{\Phi}_{j\cdot}\| + (1/\sqrt{n})\end{aligned} \tag{B.5}$$

w.h.p., where (a) holds since $\max_j \|\boldsymbol{\Phi}_{j\cdot}^{(i)}\| \lesssim \max_j \|\boldsymbol{\Phi}_{j\cdot}\|$ in Lemma A.9, and $\|\boldsymbol{\Psi}_{j\cdot}\| = 1/\sqrt{m}$ by definition. It remains to show $\max_j \|\boldsymbol{\Phi}_{j\cdot}\| \ge 1/\sqrt{n}$, which comes from

$$1 = \|\boldsymbol{\Phi}\| \le \sqrt{\sum_{i=1}^n \|\boldsymbol{\Phi}_{i\cdot}\|^2} \le \sqrt{n}\max_j\|\boldsymbol{\Phi}_{j\cdot}\|.$$

Plugging $\max_j \|\boldsymbol{\Phi}_{j\cdot}\| \ge 1/\sqrt{n}$ back into (B.5) completes the proof. ∎

Proof of Lemma A.11. By definition,

$$\begin{aligned}\|\boldsymbol{S}^{(i)}\boldsymbol{O}^{(i)}-\boldsymbol{O}\| &= \|\mathcal{P}(\boldsymbol{\Psi}^\top\boldsymbol{\Phi}^{(i)})\boldsymbol{O}^{(i)}-\mathcal{P}(\boldsymbol{\Psi}^\top\boldsymbol{\Phi})\| = \|\mathcal{P}(\boldsymbol{\Psi}^\top\boldsymbol{\Phi}^{(i)}\boldsymbol{O}^{(i)})-\mathcal{P}(\boldsymbol{\Psi}^\top\boldsymbol{\Phi})\| \\ &\overset{(a)}{\le} 2\sqrt{2}\min\{\sigma_{\min}(\boldsymbol{\Psi}^\top\boldsymbol{\Phi}^{(i)}),\,\sigma_{\min}(\boldsymbol{\Psi}^\top\boldsymbol{\Phi})\}\|\boldsymbol{\Psi}^\top(\boldsymbol{\Phi}^{(i)}\boldsymbol{O}^{(i)}-\boldsymbol{\Phi})\|,\end{aligned} \tag{B.6}$$

where (a) comes from Lemma A.1. To proceed, from Lemma A.8 we have $1-\sigma_{\min}(\boldsymbol{\Psi}^\top\boldsymbol{\Phi}^{(i)}) \lesssim \tau$ and $1-\sigma_{\min}(\boldsymbol{\Psi}^\top\boldsymbol{\Phi}) \lesssim \tau$ w.h.p. Moreover, under the condition (5.3) that $\eta \le c_0$, it satisfies

$$\tau = \frac{\sqrt{p(1-p)}+\sqrt{q}}{p\sqrt{n}} \overset{(a)}{\le} \frac{\sqrt{2(p(1-p)+q)}}{p\sqrt{n}} \le \frac{\sqrt{2}c_0}{\log(nd)} = o(1),$$

where (a) uses the fact $\sqrt{x}+\sqrt{y} \le \sqrt{2(x+y)}$ for any $x, y \ge 0$. This implies $\sigma_{\min}(\boldsymbol{\Psi}^\top\boldsymbol{\Phi}^{(i)}) = \Omega(1), \sigma_{\min}(\boldsymbol{\Psi}^\top\boldsymbol{\Phi}) = \Omega(1)$ w.h.p. Plugging this back into (B.6) yields

$$\|\boldsymbol{S}^{(i)}\boldsymbol{O}^{(i)}-\boldsymbol{O}\| \lesssim \|\boldsymbol{\Psi}^\top(\boldsymbol{\Phi}^{(i)}\boldsymbol{O}^{(i)}-\boldsymbol{\Phi})\| \le \|\boldsymbol{\Psi}\|\|\boldsymbol{\Phi}-\boldsymbol{\Phi}^{(i)}\boldsymbol{O}^{(i)}\| \lesssim \eta\max_j\|\boldsymbol{\Phi}_{j\cdot}\|$$

w.h.p., where Lemma A.9 is applied to bound $\|\boldsymbol{\Phi}-\boldsymbol{\Phi}^{(i)}\boldsymbol{O}^{(i)}\|$. ∎

Proof of Lemma A.12. Let $i' \in C_1$ and $j' \in C_2$ be two nodes such that

$$\|\boldsymbol{\Phi}_{i'\cdot}\| = \max_{i \in C_1} \|\boldsymbol{\Phi}_{i\cdot}\|, \quad \|\boldsymbol{\Phi}_{j'\cdot}\| = \max_{j \in C_2} \|\boldsymbol{\Phi}_{j\cdot}\|.$$

Without loss of generality we assume $\|\boldsymbol{\Phi}_{j'\cdot}\| = \max_j \|\boldsymbol{\Phi}_{j\cdot}\|$ which is the largest block among all nodes. Then, we define a matrix $\boldsymbol{\Phi}' \in \mathbb{R}^{nd \times 2d}$ which has the same size of $\boldsymbol{\Phi}$ and is formed by $\boldsymbol{\Phi}_{i'\cdot}, \boldsymbol{\Phi}_{j'\cdot}$ as

$$(\boldsymbol{\Phi}')^\top = [\underbrace{\boldsymbol{\Phi}_{i'\cdot}^\top \quad \cdots \quad \boldsymbol{\Phi}_{i'\cdot}^\top}_{m \times \boldsymbol{\Phi}_{i'\cdot}^\top} \quad \underbrace{\boldsymbol{\Phi}_{j'\cdot}^\top \quad \cdots \quad \boldsymbol{\Phi}_{j'\cdot}^\top}_{m \times \boldsymbol{\Phi}_{j'\cdot}^\top}].$$

As a result, $\boldsymbol{\Phi}'$ is close to $\boldsymbol{\Phi}$ such that

$$\|\boldsymbol{\Phi} - \boldsymbol{\Phi}'\| \overset{(a)}{\leq} \sqrt{\sum_{i=1}^{m} \|\boldsymbol{\Phi}_{i\cdot} - \boldsymbol{\Phi}_{i'\cdot}\|^2 + \sum_{j=m+1}^{n} \|\boldsymbol{\Phi}_{j\cdot} - \boldsymbol{\Phi}_{j'\cdot}\|^2} \overset{(b)}{\lesssim} \sqrt{n}\eta \|\boldsymbol{\Phi}_{j'\cdot}\| \tag{B.7}$$

w.h.p., where (a) holds by definition of the operator norm and (b) comes from (A.8) in the proof of Lemma 5.4. This leads to

$$|\sigma_{\max}(\boldsymbol{\Phi}) - \sigma_{\max}(\boldsymbol{\Phi}')| \overset{(a)}{=} |1 - \sigma_{\max}(\boldsymbol{\Phi}')| \overset{(b)}{\leq} \|\boldsymbol{\Phi} - \boldsymbol{\Phi}'\| \lesssim \sqrt{n}\eta \|\boldsymbol{\Phi}_{j'\cdot}\| \tag{B.8}$$

w.h.p., where (a) uses $\sigma_{\max}(\boldsymbol{\Phi}) = 1$ since $\boldsymbol{\Phi}^\top \boldsymbol{\Phi} = \boldsymbol{I}_{2d}$ by definition and (b) comes from Weyl's inequality. On the other hand, by definition of the operator norm,

$$\begin{aligned}\sigma_{\max}(\boldsymbol{\Phi}') &= \max_{\|\boldsymbol{x}\|=1} \|\boldsymbol{\Phi}'\boldsymbol{x}\| = \max_{\|\boldsymbol{x}\|=1} \sqrt{\sum_{i=1}^{m} \|\boldsymbol{\Phi}_{i'\cdot}\boldsymbol{x}\|^2 + \sum_{j=m+1}^{n} \|\boldsymbol{\Phi}_{j'\cdot}\boldsymbol{x}\|^2} \\ &\geq \max_{\|\boldsymbol{x}\|=1} \sqrt{\sum_{j=m+1}^{n} \|\boldsymbol{\Phi}_{j'\cdot}\boldsymbol{x}\|^2} = \sqrt{m}\|\boldsymbol{\Phi}_{j'\cdot}\|.\end{aligned}$$

Combining this with (B.8) gives

$$\sqrt{m}\|\boldsymbol{\Phi}_{j'\cdot}\| - 1 \leq \sigma_{\max}(\boldsymbol{\Phi}') - 1 \lesssim \sqrt{n}\eta \|\boldsymbol{\Phi}_{j'\cdot}\|$$

w.h.p. This implies as long as the condition (5.3) that $\eta \leq c_0$ holds for a sufficiently small c_0, it satisfies $\|\boldsymbol{\Phi}_{j'\cdot}\| = O(1/\sqrt{n})$, which completes the proof. ∎

B.4. Proof of the lemmas in Appendix A.3. *Proof of Lemma* A.13. By definition, $\sigma_l(\boldsymbol{\Psi}_{i\cdot}\boldsymbol{O}) = \sqrt{2/n}$ for $l = 1, \ldots, d$. Then,

$$\|\sigma_l(\boldsymbol{\Phi}_{i\cdot}) - \sqrt{2/n}\| \overset{(a)}{\leq} \|\boldsymbol{\Phi}_{i\cdot} - \boldsymbol{\Psi}_{i\cdot}\boldsymbol{O}\| \overset{(b)}{\lesssim} \eta/\sqrt{n}, \quad l = 1, \ldots, d,$$

w.h.p., where (a) holds by Weyl's inequality in Theorem A.2 and (b) comes from Theorem 5.1. $\|\mathcal{P}(\boldsymbol{\Phi}_{i\cdot}) - \mathcal{P}(\boldsymbol{\Psi}_{i\cdot}\boldsymbol{O})\|$ is bounded by Lemma A.1 as

$$\begin{aligned}\|\mathcal{P}(\boldsymbol{\Phi}_{i\cdot}) - \mathcal{P}(\boldsymbol{\Psi}_{i\cdot}\boldsymbol{O})\| &\leq 2\sqrt{2}\min\{\sigma_{\min}^{-1}(\boldsymbol{\Phi}_{i\cdot}), \sigma_{\min}^{-1}(\boldsymbol{\Psi}_{i\cdot}\boldsymbol{O})\}\|\boldsymbol{\Phi}_{i\cdot} - \boldsymbol{\Psi}_{i\cdot}\boldsymbol{O}\| \\ &\lesssim \min\left\{(\sqrt{2/n} - O(\eta/\sqrt{n}))^{-1}, (\sqrt{2/n})^{-1}\right\} \cdot O(\eta/\sqrt{n}) = O(\eta)\end{aligned}$$

w.h.p. For $\|\boldsymbol{\Phi}_{j\cdot}\mathcal{P}(\boldsymbol{\Phi}_{i\cdot})^\top - \boldsymbol{\Psi}_{j\cdot}\mathcal{P}(\boldsymbol{\Psi}_{i\cdot})^\top\|$, by substituting $\boldsymbol{\Phi}_{j\cdot}$ and $\mathcal{P}(\boldsymbol{\Psi}_{i\cdot})$ with $\boldsymbol{\Psi}_{j\cdot}\boldsymbol{O} + \boldsymbol{\Phi}_{j\cdot} - \boldsymbol{\Psi}_{j\cdot}\boldsymbol{O} = \boldsymbol{\Psi}_{j\cdot}\boldsymbol{O}$ and $\mathcal{P}(\boldsymbol{\Psi}_{i\cdot}\boldsymbol{O}) + \mathcal{P}(\boldsymbol{\Phi}_{i\cdot}) - \mathcal{P}(\boldsymbol{\Psi}_{i\cdot}\boldsymbol{O})$, respectively, we denote $\boldsymbol{\Delta}_{\boldsymbol{\Phi}_{j\cdot}} := \boldsymbol{\Phi}_{j\cdot} - \boldsymbol{\Psi}_{j\cdot}\boldsymbol{O}$ and $\boldsymbol{\Delta}_{\mathcal{P}(\boldsymbol{\Phi}_{i\cdot})} := \mathcal{P}(\boldsymbol{\Phi}_{i\cdot}) - \mathcal{P}(\boldsymbol{\Psi}_{i\cdot}\boldsymbol{O})$; then we get

$$\begin{aligned}
&\|\boldsymbol{\Phi}_{j\cdot}\mathcal{P}(\boldsymbol{\Phi}_{i\cdot})^\top - \boldsymbol{\Psi}_{j\cdot}\mathcal{P}(\boldsymbol{\Psi}_{i\cdot})^\top\| \\
&\quad = \|(\boldsymbol{\Psi}_{j\cdot}\boldsymbol{O} + \boldsymbol{\Delta}_{\boldsymbol{\Phi}_{j\cdot}})(\mathcal{P}(\boldsymbol{\Psi}_{i\cdot}\boldsymbol{O}) + \boldsymbol{\Delta}_{\mathcal{P}(\boldsymbol{\Phi}_{i\cdot})})^\top - \boldsymbol{\Psi}_{j\cdot}\mathcal{P}(\boldsymbol{\Psi}_{i\cdot})^\top\| \\
&\quad = \|\boldsymbol{\Delta}_{\boldsymbol{\Phi}_{j\cdot}}\mathcal{P}(\boldsymbol{\Psi}_{j\cdot}\boldsymbol{O})^\top + \boldsymbol{\Psi}_{j\cdot}\boldsymbol{O}\boldsymbol{\Delta}_{\mathcal{P}(\boldsymbol{\Phi}_{i\cdot})}^\top + \boldsymbol{\Delta}_{\boldsymbol{\Phi}_{j\cdot}}\boldsymbol{\Delta}_{\mathcal{P}(\boldsymbol{\Phi}_{i\cdot})}^\top\| \\
&\quad \le \|\boldsymbol{\Delta}_{\boldsymbol{\Phi}_{j\cdot}}\|\|\mathcal{P}(\boldsymbol{\Psi}_{j\cdot})\| + \|\boldsymbol{\Psi}_{j\cdot}\|\|\boldsymbol{\Delta}_{\mathcal{P}(\boldsymbol{\Phi}_{i\cdot})}\| + \|\boldsymbol{\Delta}_{\boldsymbol{\Phi}_{j\cdot}}\|\|\boldsymbol{\Delta}_{\mathcal{P}(\boldsymbol{\Phi}_{i\cdot})}\| \\
&\quad \lesssim \frac{\eta}{\sqrt{n}}\cdot 1 + \sqrt{\frac{2}{n}}\cdot\eta + \frac{\eta}{\sqrt{n}}\cdot\eta \lesssim \frac{\eta}{\sqrt{n}}
\end{aligned}$$

w.h.p., where $\|\boldsymbol{\Delta}_{\boldsymbol{\Phi}_{j\cdot}}\|$ is bounded by Theorem 5.1. It remains to bound the last term $\|\mathcal{P}(\boldsymbol{\Phi}_{j\cdot}\mathcal{P}(\boldsymbol{\Phi}_{i\cdot})^\top) - \mathcal{P}(\boldsymbol{\Psi}_{j\cdot}\mathcal{P}(\boldsymbol{\Psi}_{i\cdot})^\top)\|$; by Lemma A.1 we have

$$\begin{aligned}
&\|\mathcal{P}(\boldsymbol{\Phi}_{j\cdot}\mathcal{P}(\boldsymbol{\Phi}_{i\cdot})^\top) - \mathcal{P}(\boldsymbol{\Psi}_{j\cdot}\mathcal{P}(\boldsymbol{\Psi}_{i\cdot})^\top)\| \\
&\quad \lesssim \min\{\sigma_{\min}^{-1}(\boldsymbol{\Phi}_{j\cdot}\mathcal{P}(\boldsymbol{\Phi}_{i\cdot})^\top),\, \sigma_{\min}^{-1}(\boldsymbol{\Psi}_{j\cdot}\mathcal{P}(\boldsymbol{\Psi}_{i\cdot})^\top)\}\|\boldsymbol{\Phi}_{j\cdot}\mathcal{P}(\boldsymbol{\Phi}_{i\cdot})^\top - \boldsymbol{\Psi}_{j\cdot}\mathcal{P}(\boldsymbol{\Psi}_{i\cdot})^\top\| \\
&\quad \lesssim \min\left\{(\sqrt{2/n} - O(\eta/\sqrt{n}))^{-1},\, (\sqrt{2/n})^{-1}\right\}\cdot O(\eta/\sqrt{n}) = O(\eta)
\end{aligned}$$

w.h.p., where $\sigma_{\min}(\boldsymbol{\Psi}_{j\cdot}\mathcal{P}(\boldsymbol{\Psi}_{i\cdot})^\top) = \sqrt{2/n}$, and

$$\begin{aligned}
\sigma_{\min}(\boldsymbol{\Phi}_{j\cdot}\mathcal{P}(\boldsymbol{\Phi}_{i\cdot})^\top) &\ge \sigma_{\min}(\boldsymbol{\Psi}_{j\cdot}\mathcal{P}(\boldsymbol{\Psi}_{i\cdot})^\top) - \|\boldsymbol{\Phi}_{j\cdot}\mathcal{P}(\boldsymbol{\Phi}_{i\cdot})^\top - \boldsymbol{\Psi}_{j\cdot}\mathcal{P}(\boldsymbol{\Psi}_{i\cdot})^\top\| \\
&\ge \sqrt{2/n} - O(\eta/\sqrt{n})
\end{aligned}$$

w.h.p., which is bounded by Theorem A.2. This completes the proof. ∎

Proof of Lemma A.14. For $\|\boldsymbol{Q}_{\cdot 2} - \mathcal{P}(\boldsymbol{\Psi}_{j\cdot}\boldsymbol{O})^\top\bar{\boldsymbol{O}}_2\|$, the following satisfies:

$$\begin{aligned}
&\|\boldsymbol{Q}_{\cdot 2} - \mathcal{P}(\boldsymbol{\Psi}_{j\cdot}\boldsymbol{O})^\top\bar{\boldsymbol{O}}_2\| = \|\boldsymbol{Q}_{\cdot 1}^\perp - \mathcal{P}(\boldsymbol{\Psi}_{j\cdot}\boldsymbol{O})^\top\bar{\boldsymbol{O}}_2\| \overset{(a)}{\lesssim} \|\boldsymbol{Q}_{\cdot 1}(\boldsymbol{Q}_{\cdot 1})^\top\mathcal{P}(\boldsymbol{\Psi}_{j\cdot}\boldsymbol{O})^\top\bar{\boldsymbol{O}}_2\| \\
&= \|(\boldsymbol{Q}_{\cdot 1})^\top\mathcal{P}(\boldsymbol{\Psi}_{j\cdot}\boldsymbol{O})^\top\bar{\boldsymbol{O}}_2\| = \|\bar{\boldsymbol{O}}_1^\top\mathcal{P}(\boldsymbol{\Phi}_{i\cdot})\mathcal{P}(\boldsymbol{\Psi}_{j\cdot}\boldsymbol{O})^\top\bar{\boldsymbol{O}}_2\| = \|\mathcal{P}(\boldsymbol{\Phi}_{i\cdot})\mathcal{P}(\boldsymbol{\Psi}_{j\cdot}\boldsymbol{O})^\top\| \\
&= \|\mathcal{P}(\boldsymbol{\Phi}_{i\cdot})\mathcal{P}(\boldsymbol{\Psi}_{j\cdot}\boldsymbol{O})^\top\| \le \|\mathcal{P}(\boldsymbol{\Psi}_{i\cdot}\boldsymbol{O})\mathcal{P}(\boldsymbol{\Psi}_{j\cdot}\boldsymbol{O})^\top\| + \|(\mathcal{P}(\boldsymbol{\Psi}_{i\cdot}) - \mathcal{P}(\boldsymbol{\Psi}_{i\cdot}\boldsymbol{O}))\mathcal{P}(\boldsymbol{\Psi}_{j\cdot})^\top\| \\
&\overset{(b)}{\le} \|\mathcal{P}(\boldsymbol{\Psi}_{i\cdot})\mathcal{P}(\boldsymbol{\Psi}_{j\cdot})^\top\| + \|\mathcal{P}(\boldsymbol{\Psi}_{i\cdot}) - \mathcal{P}(\boldsymbol{\Psi}_{i\cdot}\boldsymbol{O})\|\|\mathcal{P}(\boldsymbol{\Psi}_{j\cdot})^\top\| \lesssim \eta
\end{aligned}$$

w.h.p., where (a) comes from Lemma A.5 with $\boldsymbol{X} = \boldsymbol{Q}_{\cdot 1}^\perp$, then $\boldsymbol{I}_{2d} - \boldsymbol{X}\boldsymbol{X}^\top = \boldsymbol{Q}_{\cdot 1}(\boldsymbol{Q}_{\cdot 1})^\top$, and (b) uses (A.11) in Lemma A.13. Next, for $\|\boldsymbol{\Phi}_{j\cdot}\boldsymbol{Q}_{\cdot 2} - \boldsymbol{\Psi}_{j\cdot}\mathcal{P}(\boldsymbol{\Psi}_{j\cdot})^\top\bar{\boldsymbol{O}}_2\|$ we have

$$\begin{aligned}
&\|\boldsymbol{\Phi}_{j\cdot}\boldsymbol{Q}_{\cdot 2} - \boldsymbol{\Psi}_{j\cdot}\mathcal{P}(\boldsymbol{\Psi}_{j\cdot})^\top\bar{\boldsymbol{O}}_2\| \le \|\boldsymbol{\Psi}_{j\cdot}\boldsymbol{O}\boldsymbol{Q}_{\cdot 2} - \boldsymbol{\Psi}_{j\cdot}\mathcal{P}(\boldsymbol{\Psi}_{j\cdot})^\top\bar{\boldsymbol{O}}_2\| + \|(\boldsymbol{\Phi}_{j\cdot} - \boldsymbol{\Psi}_{j\cdot}\boldsymbol{O})\boldsymbol{Q}_{\cdot 2}\| \\
&\quad \le \|\boldsymbol{\Psi}_{j\cdot}\|\|\boldsymbol{Q}_{\cdot 2} - \mathcal{P}(\boldsymbol{\Psi}_{j\cdot}\boldsymbol{O})^\top\bar{\boldsymbol{O}}_2\| + \|\boldsymbol{\Phi}_{j\cdot} - \boldsymbol{\Psi}_{j\cdot}\boldsymbol{O}\|\|\boldsymbol{Q}_{\cdot 2}\| \lesssim \eta/\sqrt{n}
\end{aligned}$$

w.h.p. The bounds on $\|\mathcal{P}(\boldsymbol{\Phi}_{j\cdot}\boldsymbol{Q}_{\cdot 2}) - \mathcal{P}(\boldsymbol{\Psi}_{j\cdot}\mathcal{P}(\boldsymbol{\Psi}_{j\cdot})^\top\bar{\boldsymbol{O}}_2)\|$ can be obtained by applying Lemma A.1 with Theorem A.2, which is similar to (A.12) in Lemma A.13, and thus we do not repeat. ∎

Appendix C. A more involved noise model. In this section, we study a more involved noise model that extends the one introduced in section 3. Recall that the orthogonal transform $\boldsymbol{A}_{ij}$ is measured exactly when nodes i and j belong to the same cluster. To extend from the measurement model in (3.1), we include additive noise perturbation to (3.1), and the new noisy measurement $\widetilde{\boldsymbol{A}}_{ij}$ for any $i < j$ follows:

$$\widetilde{\boldsymbol{A}}_{ij} = \boldsymbol{A}_{ij} + \boldsymbol{W}_{ij}, \tag{C.1}$$

where $\boldsymbol{W}_{ij}$ denotes the additive noise that is independent for each pair of nodes (i,j). For now, there is no need to specify the statistics of $\boldsymbol{W}_{ij}$ but only make sure that $\boldsymbol{W}_{ii}=\mathbf{0}$ for $i=1,\dots,n$; $\mathbb{E}[\boldsymbol{W}_{ij}]=\mathbf{0}$ for any pair of nodes, and $\boldsymbol{W}_{ij}=\boldsymbol{W}_{ji}^{\top}$ for the sake of symmetry. Notably, such an additive noise model was also broadly considered in orthogonal group synchronization problems, e.g., [52, 53].

In this case, our proposed Algorithm 1 still applies and is able to recover the cluster memberships and the orthogonal transforms, as we show both theoretically and empirically in the following.

C.1. Analysis. Our analysis for this noise model is still based on the setting of two clusters with equal cluster sizes. First, let us denote $\boldsymbol{W}=[\boldsymbol{W}_{ij}]_{i,j=1}^{n}\in\mathbb{R}^{nd\times nd}$ as the whole symmetric matrix of additive noise; then we require the following assumptions on $\boldsymbol{W}$.

Assumption C.1 (operator norm). It satisfies $\|\boldsymbol{W}\|\le\epsilon_0$ for some

$$\epsilon_0=O\big(\sqrt{p(1-p)n}+\sqrt{qn}\big),$$

with probability at least $1-O(n^{-1})$.

Assumption C.2 (block row sum concentration). Given any block matrix $\boldsymbol{M}\in\mathbb{R}^{nd\times r}$ with n block rows, for each block row $\boldsymbol{W}_{i\cdot}$ it satisfies

$$\|\boldsymbol{W}_{i\cdot}\boldsymbol{M}\|\lesssim\epsilon\max_{j}\|\boldsymbol{M}_{j\cdot}\|$$

for some $\epsilon>0$, with probability at least $1-O(n^{-1})$.

Both Assumptions C.1 and C.2 are reasonable as Assumption C.1 states an overall concentration on $\boldsymbol{W}_{ij}$, and Assumption C.2 further confines the variance of each block row of $\boldsymbol{W}$. In particular, such a block row concentration is necessary for getting a blockwise analysis. Given the above, we are able to derive the following blockwise error bound between the noisy eigenvectors $\boldsymbol{\Phi}$ and the clean ones $\boldsymbol{\Psi}$.

THEOREM C.1 (blockwise error bound). *Under Assumptions* C.1 *and* C.2 *and the setting of two equal-sized clusters with model parameters* $(n,p,q,d,\epsilon_0,\epsilon)$*, for a sufficiently large* n*, suppose*

$$\tilde{\eta}:=\frac{\sqrt{(p(1-p)+q)n\log(nd)}+\epsilon}{pn}\le c_0 \tag{C.2}$$

for some small constant $c_0\ge 0$*. Then with probability* $1-O(n^{-1})$*,*

$$\max_{1\le i\le n}\|\boldsymbol{\Phi}_{i\cdot}-\boldsymbol{\Psi}_{i\cdot}\boldsymbol{O}\|\lesssim\frac{\tilde{\eta}}{\sqrt{n}}, \tag{C.3}$$

where $\boldsymbol{O}=\mathcal{P}(\boldsymbol{\Psi}^{\top}\boldsymbol{\Phi})$.

Proof. The proof structure essentially follows the one of Theorem 5.1, with the only difference on the analysis of the perturbation $\widetilde{\boldsymbol{\Delta}}=\widetilde{\boldsymbol{A}}-\mathbb{E}[\widetilde{\boldsymbol{A}}]$. In this case, $\widetilde{\boldsymbol{\Delta}}=\boldsymbol{\Delta}+\boldsymbol{W}$, where $\boldsymbol{\Delta}$ represents the original perturbation defined in (A.2). As a result, by applying the triangular inequality, the operator norm bound on $\|\widetilde{\boldsymbol{\Delta}}\|$ given in Lemma A.6 is now written as

$$\|\widetilde{\boldsymbol{\Delta}}\|\le\|\boldsymbol{\Delta}\|+\|\boldsymbol{W}\|\le\sqrt{p(1-p)n}+\sqrt{qn}+\epsilon_0.$$

Similarly, the block row sum inequality in Lemma A.7 can be given as

$$\|\widetilde{\boldsymbol{\Delta}}_{i\cdot}\boldsymbol{M}\| \le \|\boldsymbol{\Delta}_{i\cdot}\boldsymbol{M}\| + \|\boldsymbol{W}_{i\cdot}\boldsymbol{M}\| \le \sqrt{(p(1-p)+q)n\log(nd)} + \epsilon.$$

Then, the remaining proof is almost identical to the one presented in Appendix A.2, but only replacing the original bounds on $\|\widetilde{\boldsymbol{\Delta}}\|$ and $\|\widetilde{\boldsymbol{\Delta}}_{i\cdot}\boldsymbol{M}\|$ with the updated ones, and we leave the detailed proof to interested readers. ∎

As a result, the performance guarantee of Algorithm 1 on the new noise model is identical to Theorem 5.2, and we restate that here.

THEOREM C.2 (performance guarantee). *Under the assumption of Theorem* C.1, *for* $i = 1, \dots, n$, *with probability* $1 - O(n^{-1})$, *Algorithm* 1 *exactly recovers the cluster memberships* $\kappa(i)$ *defined in section* 3, *and* $\hat{\boldsymbol{O}}_i$ *satisfies*

$$\|\hat{\boldsymbol{O}}_i - \boldsymbol{O}_i\bar{\boldsymbol{O}}_{\kappa(i)}\| \lesssim \tilde{\eta}, \tag{C.4}$$

where $\bar{\boldsymbol{O}}_{\kappa(i)}$ *is orthogonal and only depends on the cluster that* i *belongs to.*

Proof. The proof is identical to the one of Theorem 5.2, and therefore we do not repeat. ∎

Remark C.5. As an example, let us consider the case of i.i.d. Gaussian noise such that all the entries in the off-diagonal blocks $\boldsymbol{W}_{ij}$ are i.i.d. Gaussian random variables with mean zero and variance σ^2. In other words, $\boldsymbol{W}_{ij} \sim \mathcal{N}(0, \sigma^2\boldsymbol{I}_d)$ for any pair of nodes (i, j) that $i \neq j$. Then, by using an existing result on the operator norm (e.g., [70, Theorem 4.4.5]) we have

$$\|\boldsymbol{W}\| \lesssim \sigma\sqrt{nd}$$

with high probability. For the bound on $\|\widetilde{\boldsymbol{\Delta}}_{i\cdot}\boldsymbol{M}\|$, we have

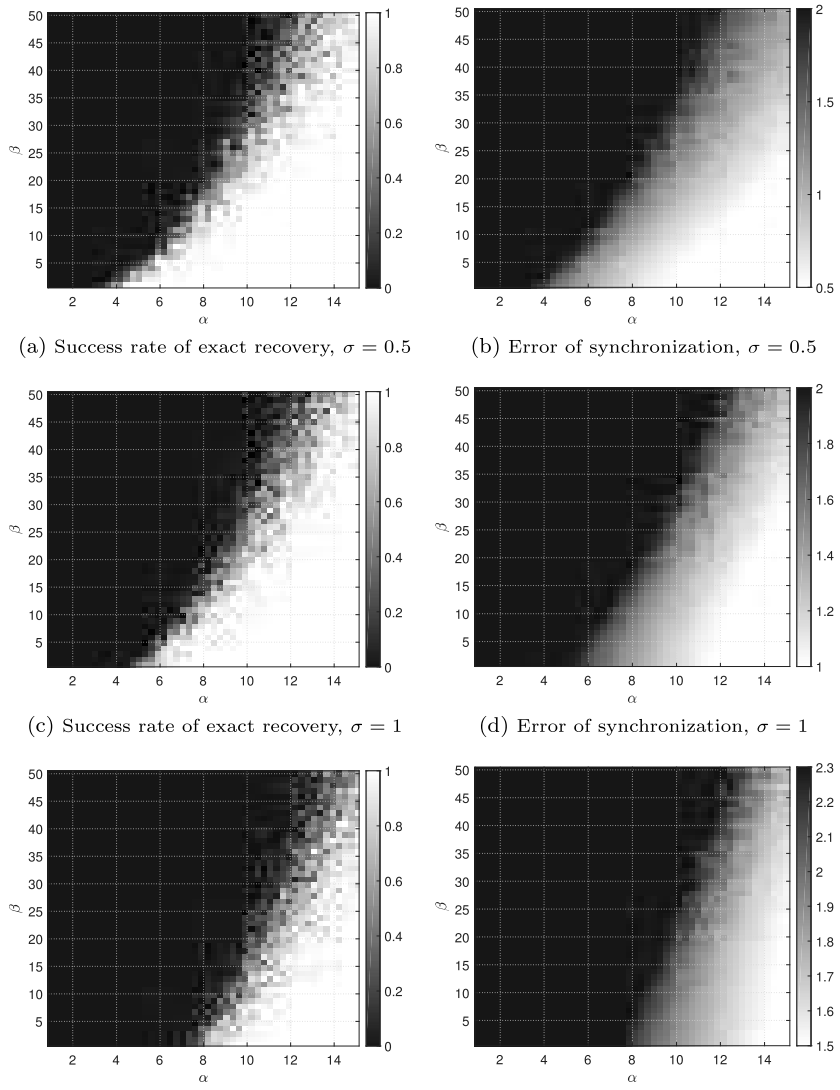
$$\|\widetilde{\boldsymbol{\Delta}}_{i\cdot}\boldsymbol{M}\| \le \|\widetilde{\boldsymbol{\Delta}}_{i\cdot}\|\|\boldsymbol{M}\| \overset{(a)}{\lesssim} \sigma\sqrt{nd}\cdot\|\boldsymbol{M}\| \overset{(b)}{\le} \sigma n\sqrt{nd}\max_j\|\boldsymbol{M}_{j\cdot}\|.,$$

where (a) holds since $\|\widetilde{\boldsymbol{\Delta}}_{i\cdot}\| \lesssim \sigma\sqrt{nd}$ and (b) comes from the fact that $\|\boldsymbol{M}\| \le n\max_j\|\boldsymbol{M}_{j\cdot}\|$. As a result, we can set $\epsilon_0 = \sigma\sqrt{nd}$ and $\epsilon = \sigma n\sqrt{nd}$; then the exact recovery condition (C.2) becomes

$$\tilde{\eta} = \frac{\sqrt{(p(1-p)+q)\log(nd)} + \sigma n\sqrt{d}}{p\sqrt{n}} \le c_0,$$

which indicates that exact recovery is available with high probability as long as σ is less than a certain threshold.

C.2. Experiments. In this part, we empirically test Algorithm 1 on the model with additive white Gaussian noise discussed in Remark C.5. We follow the same evaluation process as in section 6, and the result is shown in Figure 8. We test on the case of $m_1 = m_2 = 500$, $d = 2$, and different noise levels $\sigma \in \{0, 5, 1, 2\}$. In Figure 8 we still observe a sharp phase transition phenomenon on the exact recovery of cluster memberships, and the error of synchronization is also bounded when the clusters are perfectly identified. This agrees with our theoretical analysis and demonstrates the efficacy of our proposed algorithm.



(a) Success rate of exact recovery, $\sigma = 0.5$

(b) Error of synchronization, $\sigma = 0.5$

(c) Success rate of exact recovery, $\sigma = 1$

(d) Error of synchronization, $\sigma = 1$

(e) Success rate of exact recovery, $\sigma = 2$

(f) Error of synchronization, $\sigma = 2$

FIG. 8. *Results on the probabilistic model with additive white Gaussian noise. We test under the setting* $m_1 = m_2 = 500$, $d = 2$, *the noise levels* $\sigma \in \{0.5, 1, 2\}$. (a), (c), *and* (e)*: the success rate of exact recovery by* (6.1)*, under varying* α *in* $p = \alpha \log n/n$ *and* β *in* $q = \beta \log n/n$*;* (b), (d), *and* (f)*: the synchronization error defined in* (6.2) *under varying* α *and* β.

REFERENCES

[1] E. Abbe. A. S. Bandeira, and G. Hall, *Exact recovery in the stochastic block model*, IEEE Trans. Inform. Theory, 62 (2015), pp. 471–487.

[2] E. Abbe, J. Fan, K. Wang, Y. Zhong, et al., *Entrywise eigenvector analysis of random matrices with low expected rank*, Ann. Statist., 48 (2020), pp. 1452–1474.

[3] E. Abbe and C. Sandon, *Community detection in general stochastic block models: Fundamental limits and efficient algorithms for recovery*, in Proceedings of the 2015 IEEE 56th Annual Symposium on Foundations of Computer Science, IEEE, 2015, pp. 670–688.

[4] A. A. Amini and E. Levina, *On semidefinite relaxations for the block model*, Ann. Statist., 46 (2018), pp. 149–179.

[5] G. W. Anderson, A. Guionnet, and O. Zeitouni, *An Introduction to Random Matrices*, Cambridge University Press, Cambridge, UK, 2010.

[6] F. Arrigoni, B. Rossi, and A. Fusiello, *Spectral synchronization of multiple views in SE*(3), SIAM J. Imaging Sci., 9 (2016), pp. 1963–1990.

[7] D. Arthur and S. Vassilvitskii, *k-means++: The advantages of careful seeding*, in Proceedings of the Eighteenth Annual ACM-SIAM Symposium on Discrete Algorithms, 2007.

[8] C. Bajaj, T. Gao, Z. He, Q. Huang, and Z. Liang, *SMAC: Simultaneous mapping and clustering using spectral decompositions*, in Proceedings of the International Conference on Machine Learning, 2018, pp. 324–333.

[9] A. S. Bandeira, *Random Laplacian matrices and convex relaxations*, Found. Comput. Math., 18 (2018), pp. 345–379.

[10] S. Boucheron, G. Lugosi, and P. Massart, *Concentration Inequalities: A Nonasymptotic Theory of Independence*, Oxford University Press, Oxford, 2013.

[11] P. Businger and G. H. Golub, *Linear least squares solutions by Householder transformations*, Numer. Math., 7 (1965), pp. 269–276.

[12] K. N. Chaudhury, Y. Khoo, and A. Singer, *Global registration of multiple point clouds using semidefinite programming*, SIAM J. Optim., 25 (2015), pp. 468–501.

[13] Y. Chen, J. Fan, C. Ma, and K. Wang, *Spectral method and regularized MLE are both optimal for top-K ranking*, Ann. Statist., 47 (2019), pp. 2204–2235.

[14] A. Damle, L. Lin, and L. Ying, *Computing localized representations of the Kohn–Sham subspace via randomization and refinement*, SIAM J. Sci. Comput., 39 (2017), pp. B1178–B1198.

[15] A. Damle, L. Lin, and L. Ying, *SCDM-k: localized orbitals for solids via selected columns of the density matrix*, J. Comput. Phys., 334 (2017), pp. 1–15.

[16] A. Damle, V. Minden, and L. Ying, *Simple, direct and efficient multi-way spectral clustering*, Inf. Inference, 8 (2019), pp. 181–203.

[17] A. Damle and Y. Sun, *Uniform bounds for invariant subspace perturbations*, SIAM J. Matrix Anal. Appl., 41 (2020), pp. 1208–1236.

[18] C. Davis and W. M. Kahan, *The rotation of eigenvectors by a perturbation.* III, SIAM J. Numer. Anal., 7 (1970), pp. 1–46.

[19] A. Decelle. F. Krzakala, C. Moore, and L. Zdeborová, *Asymptotic analysis of the stochastic block model for modular networks and its algorithmic applications*, Phys. Rev. E, 84 (2011), 066106.

[20] S. Deng, S. Ling, and T. Strohmer, *Strong consistency, graph laplacians, and the stochastic block model*, J. Mach. Learn. Res., 22 (2021), pp. 5210–5253.

[21] P. Doreian, V. Batagelj, and A. Ferligoj, *Generalized Blockmodeling*, Struct. Anal. Soc. Sci. 25, Cambridge University Press, Cambridge, UK, 2005.

[22] M. E. Dyer and A. M. Frieze, *The solution of some random NP-hard problems in polynomial expected time*, J. Algorithm, 10 (1989), pp. 451–489.

[23] J. Eldridge, M. Belkin, and Y. Wang, *Unperturbed: Spectral analysis beyond Davis-Kahan*, in Proceedings of Algorithmic Learning Theory, 2018, pp. 321–358.

[24] J. Fan, W. Wang, and Y. Zhong, *An l_∞ eigenvector perturbation bound and its application to robust covariance estimation*, J. Mach. Learn. Res., 18 (2018), pp. 1–42.

[25] K. Fan and A. J. Hoffman, *Some metric inequalities in the space of matrices*, Proc. Amer. Math. Soc., 6 (1955), pp. 111–116.

[26] Y. Fan, T. Gao, and Z. Zhao, *Unsupervised co-learning on G-manifolds across irreducible representations*, in Proceedings of the Advances in Neural Information Processing Systems, 2019, pp. 9041–9053.

[27] Y. Fan, T. Gao, and Z. Zhao, *Representation theoretic patterns in multi-frequency class averaging for three-dimensional cryo-electron microscopy*, Inf. Inference, 10 (2021), pp. 723–771.

[28] Y. Fan, Y. Khoo, and Z. Zhao, *Joint community detection and rotational synchronization via semidefinite programming*, SIAM J. Math. Data Sci., 4 (2022), pp. 1052–1081.

[29] Y. Fan and Z. Zhao, *Multi-frequency vector diffusion maps*, in Proceedings of the International Conference on Machine Learning, 2019, pp. 1843–1852.

[30] U. Feige and E. Ofek, *Spectral techniques applied to sparse random graphs*, Random Structures Algorithms, 27 (2005), pp. 251–275.

[31] S. E. Fienberg, M. M. Meyer, and S. S. Wasserman, *Statistical analysis of multiple sociometric relations*, J. Amer. Statist. Assoc., 80 (1985), pp. 51–67.

[32] J. Frank, *Three-Dimensional Electron Microscopy of Macromolecular Assemblies: Visualization of Biological Molecules in Their Native State*, 2nd ed., Oxford University Press, New York, 2006.

[33] B. J. Frey and D. Dueck, *Clustering by passing messages between data points*, Science, 315 (2007), pp. 972–976.
[34] C. Gao, Z. Ma, A. Y. Zhang, and H. H. Zhou, *Achieving optimal misclassification proportion in stochastic block models*, J. Mach. Learn. Res., 18 (2017), pp. 1980–2024.
[35] C. Gao and A. Y. Zhang, *Optimal ortghogonal group synchronization and rotation group synchronization*, Inf. Inference, 12 (2023), pp. 591–632.
[36] T. Gao and Z. Zhao, *Multi-frequency phase synchronization*, in Proceedings of the International Conference on Machine Learning, 2019, pp. 2132–2141.
[37] G. H. Golub and C. F. Van Loan, *Matrix Computations. Johns Hopkins Studies in the Mathematical Sciences*, Johns Hopkins University Press, Baltimore, MD, 1996.
[38] M. Gu and S. C. Eisenstat, *Efficient algorithms for computing a strong rank-revealing QR factorization*, SIAM J. Sci. Comput., 17 (1996), pp. 848–869.
[39] O. Guédon and R. Vershynin, *Community detection in sparse networks via Grothendieck's inequality*, Probab. Theory Related Fields, 165 (2016), pp. 1025–1049.
[40] L. Hagen and A. B. Kahng, *New spectral methods for ratio cut partitioning and clustering*, IEEE Trans. Comput. Aided Des. Integr. Circuits Syst., 11 (1992), pp. 1074–1085.
[41] B. Hajek, Y. Wu, and J. Xu, *Achieving exact cluster recovery threshold via semidefinite programming*, IEEE Trans. Inform. Theory, 62 (2016), pp. 2788–2797.
[42] B. Hajek, Y. Wu, and J. Xu, *Achieving exact cluster recovery threshold via semidefinite programming: Extensions*, IEEE Trans. Inform. Theory, 62 (2016), pp. 5918–5937.
[43] N. Halko, P.-G. Martinsson, and J. A. Tropp, *Finding structure with randomness: Probabilistic algorithms for constructing approximate matrix decompositions*, SIAM Rev., 53 (2011), pp. 217–288.
[44] P. W. Holland. K. B. Laskey, and S. Leinhardt, *Stochastic blockmodels: First steps*, Soc. Networks, 5 (1983), pp. 109–137.
[45] A. S. Householder, *Unitary triangularization of a nonsymmetric matrix*, J. ACM, 5 (1958), pp. 339–342.
[46] Q.-X. Huang and L. Guibas, *Consistent shape maps via semidefinite programming*, Comput. Graph. Forum, 32 (2013), pp. 177–186.
[47] B. Karrer and M. E. Newman, *Stochastic blockmodels and community structure in networks*, Phys. Rev. E, 83 (2011), 016107.
[48] F. Krzakala, C. Moore, E. Mossel, J. Neeman. A. Sly, L. Zdeborová, and P. Zhang, *Spectral redemption in clustering sparse networks*, Proc. Natl. Acad. Sci. USA, 110 (2013), pp. 20935–20940.
[49] R. Latała, R. van Handel, and P. Youssef, *The dimension-free structure of nonhomogeneous random matrices*, Invent. Math., 214 (2018), pp. 1031–1080.
[50] R. R. Lederman and A. Singer, *A representation theory perspective on simultaneous alignment and classification*, Appl. Comput. Harmon. Anal., 49 (2020), pp. 1001–1024.
[51] J. Leskovec, K. J. Lang. A. Dasgupta, and M. W. Mahoney, *Statistical properties of community structure in large social and information networks*, in Proceedings of the 17th International Conference on World Wide Web, 2008, pp. 695–704.
[52] S. Ling, *Near-optimal performance bounds for orthogonal and permutation group synchronization via spectral methods*, Appl. Comput. Harmon. Anal., 60 (2022), pp. 20–52.
[53] S. Ling, *Solving orthogonal group synchronization via convex and low-rank optimization: tightness and landscape analysis*, Math. Program., (2022), pp. 1–40.
[54] S. Lloyd, *Least squares quantization in PCM*, IEEE Trans. Inform. Theory, 28 (1982), pp. 129–137.
[55] L. Massoulié, *Community detection thresholds and the weak Ramanujan property*, in Proceedings of the Forty-Sixth Annual ACM Symposium on Theory of Computing, 2014, pp. 694–703.
[56] F. McSherry, *Spectral partitioning of random graphs*, in Proceedings of the 42nd IEEE Symposium on Foundations of Computer Science, IEEE, 2001, pp. 529–537.
[57] E. Mossel, J. Neeman, and A. Sly, *Reconstruction and estimation in the planted partition model*, Probab. Theory Related Fields, 162 (2015), pp. 431–461.
[58] E. Mossel, J. Neeman, and A. Sly, *A proof of the block model threshold conjecture*, Combinatorica, 38 (2018), pp. 665–708.
[59] A. Y. Ng, M. I. Jordan, and Y. Weiss, *On spectral clustering: Analysis and an algorithm*, in Proceedings of the Advances in Neural Information Processing Systems, 2002, pp. 849–856,
[60] D. Pachauri, R. Kondor, and V. Singh, *Solving the multi-way matching problem by permutation synchronization*, in Proceedings of the Advances in Neural Information Processing Systems, 2013, pp. 1860–1868.

[61] A. Perry and A. S. Wein, *A semidefinite program for unbalanced multisection in the stochastic block model*, in Proceedings of the 2017 International Conference on Sampling Theory and Applications (SampTA), IEEE, 2017, pp. 64–67.
[62] Y. Shen, Q. Huang, N. Srebro, and S. Sanghavi, *Normalized spectral map synchronization*, in Proceedings of the Advances in Neural Information Processing Systems, 2016, pp. 4925–4933.
[63] A. Singer, *Angular synchronization by eigenvectors and semidefinite programming*, Appl. Comput. Harmon. Anal., 30 (2011), pp. 20–36.
[64] A. Singer, Z. Zhao, Y. Shkolnisky, and R. Hadani, *Viewing angle classification of cryo-electron microscopy images using eigenvectors*, SIAM J. Imaging Sci., 4 (2011), pp. 723–759.
[65] G. W. Stewart, *Perturbation theory for the singular value decomposition*, Techical report, University of Maryland, College Park, MD, 1998.
[66] G. W. Stewart, *A Krylov–Schur algorithm for large eigenproblems*, SIAM J. Matrix Anal. Appl., 23 (2002), pp. 601–614.
[67] T. Tao, *Topics in Random Matrix Theory*, American Mathematical Society, Providence, RI, 2012.
[68] L. N. Trefethen and D. Bau III, *Numerical Linear Algebra*, SIAM, Philadelphia, PA, 1997.
[69] J. A. Tropp, *An Introduction to Matrix Concentration Inequalities*, Found. Trends Mach. Learn. 8, Now Publishers, Hanover, MA, 2015.
[70] R. Vershynin, *High-Dimensional Probability: An Introduction with Applications in Data Science*, Cambridge University Press, Cambridge, UK, 2018.
[71] U. Von Luxburg, *A tutorial on spectral clustering*, Stat. Comput., 17 (2007), pp. 395–416.
[72] V. Vu, *A simple SVD algorithm for finding hidden partitions*, Combin. Probab. Comput., 27 (2018), pp. 124–140.
[73] F. Woolfe, E. Liberty, V. Rokhlin, and M. Tygert, *A fast randomized algorithm for the approximation of matrices*, Appl. Comput. Harmon. Anal., 25 (2008), pp. 335–366.
[74] S.-Y. Yun and A. Proutiere, *Accurate Community Detection in the Stochastic Block Model via Spectral Algorithms*, preprint, arXiv:1412.7335, 2014.
[75] H. Zha, X. He, C. Ding, M. Gu, and H. D. Simon, *Spectral relaxation for K-means clustering*, in Proceedings of the *Advances in Neural Information Processing Systems*, 2001, pp. 1057–1064.
[76] A. Y. Zhang, *Exact Minimax Optimality of Spectral Methods in Phase Synchronization and Orthogonal Group Synchronization*, preprint, arXiv:2209.04962, 2022.
[77] Z. Zhao and A. Singer, *Rotationally invariant image representation for viewing direction classification in cryo-EM*, J. Struct. Biol., 186 (2014), pp. 153–166.
[78] Y. Zhong and N. Boumal, *Near-optimal bounds for phase synchronization*, SIAM J. Optim., 28 (2018), pp. 989–1016.