Learning Mixed Membership from Adjacency
Graph via Systematic Edge Query: Identifiability
and Algorithm

Shahana Ibrahim and Xiao Fu

Abstract—Graph clustering is one of the key techniques in
network data analysis, which aims at associating the nodes of
a graph with different clusters. This work proposes a graph
clustering paradigm performed over largely incomplete adja-
cency matrices—where only a small amount of queries about
the edges can be made. Prior works tackled this problem under
the setting of single membership nodes and disjoint clusters,
yet multiple-membership nodes and overlapping clusters often
arise in practice. Existing approaches also rely on random
edge query patterns and convex optimization-based formulations,
which give rise to a number of implementation and scalability
challenges. This work aims at learning mixed membership of
the nodes from overlapping clusters. Our method is equipped
with a systematic edge query pattern, which is arguably easier
to implement relative to random ones in a series of realistic
scenarios. The framework is accompanied by a scalable algorithm
which entails system designer, which can be easily adjusted to accommodate
challenging scenarios in situ. We also model the adjacency
graphs using a mixed membership model, and thus naturally
cover the overlapping cluster case. In terms of algorithm, we
propose a scalable procedure to learn the mixed membership
of the nodes that only consists of the truncated singular value
decomposition (SVD) of small matrices and a Gram–Schmidt-
type greedy algorithm. Theoretical guarantees for the proposed
framework are offered to support our design. We conduct
numerical evaluation on synthetic and real-world datasets to
demonstrate the effectiveness of our method.

Index Terms—Graph clustering, mixed membership, sampled
edge query, nonnegative matrix factorization

I. INTRODUCTION

Graph clustering (GC) aims at associating the nodes of
a graph with different clusters in an unsupervised manner
[1]. GC is a core technique in data science since network
data frequently arise in various applications (e.g., in social
network analysis [2], brain signal processing [3], and biolog-
ical/ecological data mining [4]). GC techniques are also used
as nonlinear dimensionality reduction tools; see e.g., spectral
clustering [5].

Theory and methods of GC have been extensively studied in
the past two decades [6]–[9]. However, some new challenges
arise in the era of big data. Notably, many network data
have grown prohibitively large—e.g., social media networks
from Facebook and Twitter could easily contain billions of
edges. Edge acquisition and subsequent computational tasks
at such a scale is highly nontrivial. On the other hand, the
availability of edge information is limited in many networks—
e.g., in community detection of networks where edges are
intentionally removed or hidden (e.g., terrorist networks or
radical group networks) [10] and in biological/ecological net-
works where acquiring the complete edge information is too
resource-consuming [11], [12]. Under these scenarios, instead
of collecting edge information of the entire network, data
analysts (have to) sample some edges of interest, and use the
sampled network to perform graph clustering [13].

A number of works [14]–[16] have considered the graph
clustering problem under incomplete edge observation. In
these works, the node membership identification guarantees
were established under the assumption that every node is asso-
ciated with a single cluster. However, in real-world networks,
the nodes often admit mixed membership and the clusters are
usually overlapped (e.g., a person in a co-author network could
belong to both the signal processing and machine learning
communities simultaneously). Also, in existing works, the
edges were queried randomly, which may not be easy to
implement in some applications; e.g., in field survey based
network analysis spanning a large geographic area [17]—
surveys are easier to be conducted within local communities,
other than randomly scattered geographically. Random query
is also not suitable for handling networks with hidden or
intentionally removed edges [10]. In terms of computation,
the existing works in [14]–[16] recast the edge query-based
GC task as nuclear norm-based convex optimization problems,
which entails $N^2$ (where $N$ is the number of nodes) optimiza-
tion variables—making it hard to scale up for real-world large
graphs.

In this work, we offer an alternative framework for learning
the node membership from incomplete graph. Unlike existing
graph clustering methods in [14]–[16], our framework features
a systematic edge query principle. Using systematic
edge query makes the query pattern under control of the
system designer, which can be easily adjusted to accommodate
challenging scenarios in situ. We also model the adjacency
graphs using a mixed membership model, and thus naturally
cover the overlapping cluster case. In terms of algorithm, we
propose a scalable procedure to learn the mixed membership
of the nodes that only consists of the truncated singular value
decomposition (SVD) of small matrices and a Gram–Schmidt-
type greedy algorithm. Theoretical guarantees for the proposed
framework are offered to support our design. We conduct
numerical evaluation on synthetic and real-world datasets to
demonstrate the effectiveness of the proposed approach.

II. PROBLEM STATEMENT

Consider $N$ data entities that are from $K$ clusters. We
consider the case where the clusters have overlaps and the node
admits mixed membership. Assume that the $n$-th entity be-
ongs to cluster $k$ with probability $m_{k,n}$, where $\sum_{k=1}^{K} m_{k,n} = 1$. For each graph $G$, we want to
find a set of $N$ mixed membership variables $\mathbf{m}_n = (m_{k,n})_{k=1}^{K}$
for each $n$. These membership values should be inferred
from the observed edges. Each edge $e_{ij}$ has a posterior
probability $P(e_{ij} = 1 | \mathbf{m}_n)$ which is given by the
mixing probability $m_{i,n} m_{j,n}$. For each $n$, the edge
query probability $P(e_{ij} | \mathbf{m}_n)$ is also a function of
the edge posterior probability $P(e_{ij} = 1 | \mathbf{m}_n)$.

X. Fu and S. Ibrahim are with the School of Electrical Engineering and
Computer Science, Oregon State University, Corvallis, OR 97331, United
States. email (xiao.fu,ibrahim)@oregonstate.edu
where $P\in[0,1]^{N\times N}$, where each entry $A(i,j)$ encodes the pairwise relationship between nodes $i$ and $j$. Our goal is to learn $M$ with a limited number of edges in $A$ observed; i.e., the number of edges to be queried is much smaller than $O(N^2)$.

Our problem setting is motivated by a number of important applications. The mixed membership learning is a core task in overlapped community detection (OCD) [18]. In OCD frameworks proposed in [19]–[21], the mixed membership was provably learned using a fully observed $A$. However, OCD under partially observed edges is of great interest for applications like field survey based community analysis [17] or community detection involving hidden edges [10]. In both cases, one cannot observe the entire $A$ due to reasons such as resource limitations and difficulty of edge acquisition.

To handle the graph clustering problem under partial edge observations, the work in [14]–[16] models the generating process of $A$ using a single cluster membership-based model, namely stochastic block model (SBM). The SBM can be summarized as follows. Each node $n$ belongs to a single cluster $k$, i.e., the membership vector $m_n$ is $k$th unit vector. The matrix $B \in \mathbb{R}^{K \times K}$ represents a cluster-cluster similarity matrix, where $B(p,q)$ represents the probability that cluster $p$ is connected with cluster $q$. Then, the probability that $A(i,j)$ is labeled as “1” is $P(i,j) = m_i^\top B m_j$, i.e., $A(i,j) \sim \text{Bernoulli}(m_i^\top B m_j)$. The adjacency matrix $A$ is sampled from Bernoulli distributions specified by the entries of the matrix $P = MBM$. In [14]–[16], convex optimization based matrix completion criteria were proposed, and the SBM model was utilized to establish the recoverability of $A$ using random edge observations. The results from [14]–[16] are insightful. However, in many applications, mixed membership is of more interest and/or random edge queries are not easy to implement.

III. PROPOSED APPROACH

Our approach relaxes the single cluster membership assumption in SBM by allowing the nodes to have mixed membership. Hence, the membership matrix $M$ satisfies

$$1^\top M = 1^\top, \quad M \succeq 0; \quad (1)$$

i.e., $m_n$ resides in the probability simplex, instead of being the unit vectors as in SBM. Under the mixed membership assumption in (1), the Bernoulli model used in the SBM, i.e.,

$$A(i,j) \sim \text{Bernoulli}(P(i,j)), \quad (2)$$

where $P(i,j) = m_i^\top B m_j$ is adopted in our generative model for the adjacency matrix $A$. Overall, (1) and (2) present a model that is reminiscent of the *mixed membership stochastic block* (MMSB) model in OCD [22], [23].

The truncated top-$K$ SVD of $C_1$ and $C_2$ can be represented as follows:

$$C_1 = [U_1^\top, U_2^\top]^\top \Sigma V^\top, \quad C_2 = [\bar{U}_2^\top, \bar{U}_3^\top]^\top \bar{\Sigma} \bar{V}^\top. \quad (5)$$

**A. Systematic Edge Query**

Our goal is to learn $M$ from systematically sampled edges of $A$. To proceed, we first divide the nodes into $L$ disjoint groups $S_1, S_2, \ldots, S_L$, such that $S_1 \cup \cdots \cup S_L = [N]$ (where $[N] = \{1, \ldots, N\}$). Let $A_{\ell,m} \in \mathbb{R}^{|S_\ell| \times |S_m|}$ denote the adjacency matrix between groups $S_\ell$ and $S_m$, where $|S_m|$ denotes the cardinality of the set $S_m$. We propose an edge query principle as follows:

**Edge Query Principle (EQP):**

- For every $\ell \in [L]$, $K \leq |S_\ell|$ holds.
- Let $m_r \in [L]$ and $\ell_r \in [L]$. For every $\ell_r$, there exists a pair of indices $m_r$ and $\ell_{r+1}$ where $\ell_{r+1} \neq \ell_r$ such that the edges from the blocks $A_{\ell_r,m_r}$ and $A_{\ell_{r+1},m_r}$ are queried.

The proposed EQP covers a large variety of query ‘masks’—some examples are shown in Fig. 1. Since the query pattern can be by design instead of random, this entails the flexibility to avoid querying edges that are known a priori hard to acquire, e.g., edges that may have been intentionally removed to conceal information or edges that correspond to interactions between groups that are hard to survey. Note that, instead of sampling individual edges, we sample blocks of edges under the proposed EQP. This allows us to design a provable and lightweight algorithm for mixed membership learning.

**B. Algorithm Design**

In this section, we propose an algorithm that consists of simple SVD operations to estimate $U \in \mathbb{R}^{N \times K}$ such that range$(U) = \text{range}(M^\top)$ and a subsequent structured matrix factorization (SMF) to estimate $M$. We name this systematic edge query based SVD procedure as SEQ-SVD and is presented in Algorithm 1.

To shed some light on how Algorithm 1 identifies $U$, let us consider the ideal case where $A_{\ell,m} = P_{\ell,m} = M_{\ell}^\top BM_m$. We start by analyzing a toy example with $L = 3$ where only the following blocks are acquired following EQP (also see Fig. 2):

$$P_{1,2} = M_{1}^\top BM_2, \quad P_{2,2} = M_{2}^\top BM_2, \quad (3)$$

$$P_{2,1} = M_{2}^\top BM_1, \quad P_{3,1} = M_{3}^\top BM_1. \quad (4)$$

Define $C_1 := [P_{1,2}^\top \ , \ P_{2,2}^\top]^\top$ and $C_2 := [P_{2,1}^\top \ , \ P_{3,1}^\top]^\top$. The truncated top-$K$ SVD of $C_1$ and $C_2$ can be represented as follows:

$$C_1 = [U_1^\top, U_2^\top]^\top \Sigma V^\top, \quad C_2 = [\bar{U}_2^\top, \bar{U}_3^\top]^\top \bar{\Sigma} \bar{V}^\top. \quad (5)$$
Combining (3)-(5), and under the assumption that \( \text{rank}(M) = \text{rank}(B) = K \) and \( K \leq |S|_1 \), for all \( \ell \), one can express the bases of \( \text{range}(M_\ell^1), \text{range}(M_\ell^2), \) and \( \text{range}(M_\ell^3) \) as follows:

\[
U_1 = M_1^\top B \Theta, \quad U_2 = M_2^\top B \Theta, \quad U_3 = M_3^\top B \Phi, \quad \Theta \in \mathbb{R}^{K \times K} \text{ and } \Phi \in \mathbb{R}^{K \times K} \text{ are certain nonsingular matrices.}
\]

Our hope is to “stitch” the bases above to have:

\[
\text{range}(U) = \text{range}(U_1, U_2, U_3) = \text{range}(M_1, M_2, M_3) = \text{range}(U_1, U_2, U_3).
\]

Note that \( U_3 \) cannot be directly combined with \( U_1 \) and \( U_2 \) to attain the above, since \( \Theta = \Phi \) does not generally hold. To fix this, we define the following: \( U := \tilde{U}_3 \tilde{U}_2 U_2 \). It is not hard to see that:

\[
\tilde{U}_3 \tilde{U}_2 U_2 = M_3^\top B \Phi \times (M_2^\top B \Phi)^\dagger \times M_2^\top B \Theta = M_3^\top B \Theta,
\]

which leads to (6). To handle a general \( L \) case, the “subspace stitching” idea conveyed by this simple example is recursively applied by Algorithm 1 which iterates over the queried blocks \( A_{\ell,r,m} \) and \( A_{r+1,m} \) for \( r = 1, \ldots, L - 1 \). We start the iterations from \( r = \lfloor L/2 \rfloor \) and perform the subspace stitching of the blocks in the ascending and descending order—this helps to reduce the subspace estimation error from the overall procedure when noise is present (critical in the binary observation case in the next subsection). We have the following proposition:

**Proposition 1.** (Ideal Case) Assume that \( A_{\ell,m} = P_{\ell,m} = M_\ell^\top B M_m \in \mathbb{R}^{|S|_1 \times |S|_m} \) holds true for all \( \ell, m \in [L] \) and \( \text{rank}(M) = \text{rank}(B) = K \). Suppose that the \( A_{\ell,m} \)’s are queried according to the proposed EQP. Then, the output \( \hat{U} \) by Algorithm 1 satisfies the following \( \hat{U} = M^\top B \Theta \), where \( \Theta \in \mathbb{R}^{K \times K} \) is an invertible matrix.

The proof can be found in supplemental material in Sec. A. Once \( U \) is estimated, the second stage boils down to estimating \( M \) from the following model (assuming that \( U \) is perfectly estimated):

\[
U^\top = GM, \quad M \geq 0, \quad 1^\top M = 1^\top,
\]

where \( G \in \mathbb{R}^{K \times K} \) is nonsingular. Learning \( M \) from the model (7) poses a so-called simplex-structured matrix factorization (SSMF) problem [24], [25]. Algorithm 1 employs a Gram-Schmidt like scalable algorithm known as successive projection algorithm (SPA) [25] for this task. From the model in (7), SPA can provably identify \( M \) in \( K \) steps, if \( G \) is nonsingular and if there exists

\[
A = \{n_1, \ldots, n_K\}
\]

such that \( M(:, n_k) = e_k \), where \( e_k \in \mathbb{R}^K \) is the \( k \)th unit vector. The existence of \( A \) translates to the existence of the so-called pure nodes in the graph [20], [21], i.e., nodes that are only associated with a single cluster. This assumption is not unreasonable when clustering a large number of data samples.

**Remark 1.** (Algorithm Complexity) Algorithm 1 only consists of top-\( K \) truncated SVD on small blocks which has a complexity of \( O((N/L)K^2) \) flops, assuming \( |S|_l = N/L \) for all \( \ell \) and the per-iteration complexity of SPA is also of similar order, i.e., \( O(NK^2) \) operations. In terms of memory, the matrices involved are of size \( N \times K \) and often, \( K \ll N \)—thereby enjoying lower memory footprint compared to the convex optimization methods.

**C. Performance Characterization under Binary Observations**

Proposition 1 presents the identifiability claims under the ideal case, i.e., \( A_{\ell,m} = P_{\ell,m} \). In practice, however, \( P_{\ell,m} \)’s are not observed. Instead, one observes \( A_{\ell,m} \)’s such that \( A(i,j) \sim \text{Bernoulli}(P(i,j)) \). The Bernoulli observations can be considered as a noisy data acquisition process. To characterize the performance under such noisy case, let us recall the degree of node \( i \) is the number of “similar nodes” it has in the adjacency graph; i.e., degree(\( i \)) = \( \sum_{j=1}^{N} A(i,j) \), where \( A \in \{0,1\}^{N \times N} \) [6]. Using this notion, we have the following proposition:

**Proposition 2** (Binary Observation Case). Assume that \( \text{rank}(M) = \text{rank}(B) = K \) and the matrix \( A \) is generated following (1) and (2), and that \( A_{\ell,m} \)’s are queried following the EQP. Let \( \rho := \max_{i,j} P(i,j) \). Suppose that \( \rho = \Omega(L\log(NK/L)/N) = L = O(\rho/N/d) \) where \( d \) is the maximal degree of all the observed sub-blocks \( A_{\ell,k} \). Also assume that \( N = \Omega \left( \frac{LK\rho N^2}{\rho_m(B)} \right) \). Then, the output \( \hat{U} \) by
Algorithm 1 satisfies the following with probability at least $1 - O(L^2/N)$:

$$
\|\hat{U} - UO\|_F = O\left(\frac{K^{L/4}K(B)\sqrt{p}}{\sigma_{\min}(B)\sqrt{N/L}}\right),
$$

(9)

where $U$ is an orthogonal basis of range$(M^T)$ and $O \in \mathbb{R}^{K \times K}$ is an orthogonal matrix.

The proof can be found in supplemental material in Sec. B. The key idea utilized to derive the subspace estimation error bound (9) is the fact that principal components of the binary adjacency sub-graphs returns the target range space up to bounded errors [26]. Leveraging this result, and combining with the proposed recursive “subspace stitching” technique, one can show (9) with careful derivations. Nevertheless, Proposition 2 has some important practical implications. First, the number of blocks $L$ plays a critical role. On one hand, $L$ cannot be too large since then the EQP condition $K \leq S_\ell$ will be violated. In addition, larger $L$ also makes the error bound looser, however, larger $L$ means fewer queries need to be made, and essentially less resource consuming.

**Remark 2.** (Estimation accuracy of $M$) Note that the estimated $\hat{U}$ can be represented as $\hat{U} = GM + N$ where $M$ satisfies (1), $G \in \mathbb{R}^{K \times K}$ is nonsingular and $N$ represents the noise which is shown to be bounded by Proposition 2. In order to extract $M$ from the estimated $\hat{U}$, Algorithm 1 employs SPA which is provably robust to bounded noise [25]. Hence, levering Proposition 2 and the noise robustness of SPA, one can show that the estimated $\hat{M}$ by Algorithm 1 is not far away from $M_{\Pi}$ for a certain permutation matrix $\Pi$.

**Remark 3.** Under the EQP, an alternate solution to estimate $M$ is to apply existing mixed membership learning algorithms [19]–[21] on the small blocks (e.g., $A_{\ell,m}$) individually to learn the corresponding part of $M$. Then, the entire $M$ can be recovered by fixing some intrinsic ambiguities between each part of $M$. This is doable, but may have relatively poor identifiability guarantees. The reason is that learning $M$ via these methods requires that the convex hull of $M^T$ to be well spread in the probability simplex (e.g., the existence of pure nodes [20], [21] implies $\text{conv}\{M^T\} = \{x \in \mathbb{R}^{K} | x^T1 = 1, x \geq 0\}$—see in-depth discussion in [27]). When the methods are applied on small subblocks $A_{\ell,m}$, this assumption is less likely to hold for the corresponding membership submatrix $M_\ell$ or $M_m$ [28]. Hence, applying existing mixed membership learning methods on small blocks is less preferable—as one will see in the experiments.

### IV. Experiments

**Baselines.** We employ two state-of-the-art mixed membership learning algorithms, namely, GeoNMF [21] and CD-MVS [19] as baselines. For real data experiments, we additionally use the normalized spectral clustering algorithm (denoted as SC-Norm) [7]. The baseline algorithms are not designed to directly handle the sampled adjacency matrix. Therefore, we use a procedure that applies these baselines to each block and aligns the estimated block membership matrices, i.e., $M_\ell$’s.

**TABLE I**

<table>
<thead>
<tr>
<th>Graph Size</th>
<th>Ideal Case</th>
<th>Binary Observation Case</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Proposed</td>
<td>Proposed</td>
</tr>
<tr>
<td>$N$</td>
<td>Dist</td>
<td>MSE</td>
</tr>
<tr>
<td>$1 \times 10^4$</td>
<td>$7.34 \times 10^{-13}$</td>
<td>0.0475</td>
</tr>
<tr>
<td>$2 \times 10^4$</td>
<td>$2.80 \times 10^{-13}$</td>
<td>0.0198</td>
</tr>
<tr>
<td>$4 \times 10^4$</td>
<td>$1.22 \times 10^{-13}$</td>
<td>0.0123</td>
</tr>
<tr>
<td>$8 \times 10^4$</td>
<td>$1.12 \times 10^{-13}$</td>
<td>0.0066</td>
</tr>
</tbody>
</table>

The details of this alignment procedure can be found in Sec. C of the supplemental material.

**Synthetic Data.** We consider $N$ data entities (where $N \in [1 \times 10^4, 8 \times 10^4]$) whose membership vectors are associated with $K = 5$ clusters. The membership vectors $m_n$ are drawn from a Dirichlet distribution with parameter $(1/K)$. The entries of matrix $B$ are drawn from $0$ to $1$. We first test the identifiability claims under the ideal case. The blocks of the adjacency matrix with a non-diagonal query pattern (the leftmost pattern in Fig. 1) is generated by assuming $A_{\ell,m} = P_{\ell,m} = M^T \cdot BM_m$. We fix the number of groups $L = 10$. Each case is evaluated for 20 random trials. Table I shows the averaged subspace estimation accuracy of our method measured using subspace distance measure (denoted as Dist) under different $N$’s. One can observe that the proposed method estimates the subspace of the membership matrix $M$ very accurately, which verifies our subspace identifiability analysis in Proposition 1.

Next, we consider the diagonal query pattern (the rightmost pattern in Fig 1) to evaluate the proposed algorithm and the baselines in the binary observation cases $A(i,j) \sim \text{Bernoulli}(P(i,j))$. Table I shows the subspace distance and mean squared error (MSE) of the estimated $M$ averaged over 20 random trials (see definition in [29]). One can see that, the subspace estimation error of the proposed algorithm gets smaller as $N$ grows, as Proposition 2 indicates. In all the cases, the proposed method outperforms the baseline methods in terms of MSE of $M$.

**Real Data.** We test the algorithms using the co-authorship network data from Microsoft Academic Graph (MAG) [30]. We use the version of the adjacency graphs, named as MAG1 and MAG2, published by the authors of [21]. The networks are provided with ground-truth mixed membership of the nodes—nodes represent the authors of the research papers from different fields of study and some nodes exhibit memberships in more than one clusters. From the original dataset, we select only those nodes which have degree at least 5. With this, the MAG1 and MAG2 networks have 37,680 and 19,457 authors, respectively; all the authors are from 3 different fields of study. We let all the algorithms access only part of the network under the diagonal query pattern. We randomize the node order in each of the 20 trials and present the averaged performance.
that the proposed algorithm outperforms the baselines for both datasets. In addition, the proposed algorithm takes runs very fast on these large-scale clustering tasks.

Table III shows the clustering accuracy of the algorithm under different $L$’s on MAG2 which is measured via applying $k$-means to the learned membership vectors. One can see that the performance decreases along with $L$’s increase—which is consistent with our analysis in Proposition 2. Notably, when $L = 50$, i.e. only 5.25% of $A$ is observed—but the proposed method still outputs a reasonable clustering accuracy, which demonstrates promising balance between low query sample complexity and clustering accuracy for large-scale graph clustering problems.

To conclude, we proposed a graph query scheme that enables provable graph clustering with partially observed binary edges. Unlike previous works relying on random edge query and computationally heavy convex programming, our method features a lightweight algorithm and works with systematic edge query patterns that are arguably more realistic in some applications. Our method also learns mixed membership of the nodes, which improves upon existing provable graph query methods that work with single membership.

REFERENCES


Algorithm 1 defines the following matrices in the $A$\immediately get for a certain nonsingular $\Theta \in \mathbb{R}^{K \times K}$. The above also implies that
\[
\text{range}(\tilde{U}) = \text{range}([M_1^T, M_2^T, \ldots, M_L^T]^T).
\]

To see how Algorithm 1 obtains such $\tilde{U}$, we consider the below four blocks for any $r \in \{1, \ldots, L - 2\}$:
\[
P_{\ell,m} = M_\ell^T B M_m, \quad P_{\ell+1,m} = M_{\ell+1}^T B M_m, \quad P_{\ell+1,m+1} = M_{\ell+1}^T B M_{m+1}, \quad P_{\ell+2,m+1} = M_{\ell+2}^T B M_{m+1}.
\] (10)

Algorithm 1 defines the following matrices in the $r$-th and $(r+1)$-th iterations, respectively (assuming the noiseless case where $A_{\ell,m} = P_{\ell,m} = M_\ell^T B M_m$ for any $\ell, m \in \{1, \ldots, L\}$).
\[
C_r := [P_{\ell,m}, P_{\ell+1,m}]^T, \quad C_{r+1} := [P_{\ell+1,m+1}, P_{\ell+2,m+1}]^T.
\]

The truncated top-$K$ SVD of $C_r$ and $C_{r+1}$ can then be represented as follows:
\[
C_r = [U_r^T, U_{r+1}^T]^T \Sigma_r V^T, \quad C_{r+1} = [U_{r+1}^T, U_{r+2}^T]^T \Sigma_{r+1} V^T.
\] (11)

Combining (10) and (11), and using the assumption that $\text{rank}(M) = \text{rank}(B) = K$ and $K \leq |S_\ell|$, the following equations hold:
\[
\begin{align*}
U_\ell &= M_\ell^T B \Theta_\ell, \quad U_{\ell+1} = M_{\ell+1}^T B \Theta_{\ell+1}, \\
\tilde{U}_{\ell+1} &= M_{\ell+1}^T B \Phi_{\ell+1}, \quad \tilde{U}_{\ell+2} = M_{\ell+2}^T B \Phi_{\ell+1},
\end{align*}
\] (12a)

(12b)

where $\Theta_\ell \in \mathbb{R}^{K \times K}$ and $\Phi_{\ell+1} \in \mathbb{R}^{K \times K}$ are certain nonsingular matrices. The equalities in (12a) imply that $U_\ell, U_{\ell+1}$ and $\tilde{U}_{\ell+2}$ are the bases of $\text{range}(M_\ell^T), \text{range}(M_{\ell+1}^T)$ and $\text{range}(M_{\ell+2}^T)$, respectively. Since $\Theta_\ell = \Phi_{\ell+1}$ does not generally hold, the basis $\tilde{U}_{\ell+2}$ cannot be directly combined with the bases $U_\ell$ and $U_{\ell+1}$ to obtain $\text{range}([M_\ell^T, M_{\ell+1}^T, M_{\ell+2}^T]^T)$. Therefore, we are interested in obtaining the basis defined as below:
\[
U_{\ell+2} := M_{\ell+2}^T B \Theta_\ell.
\] (13)

In order to identify $U_{\ell+2}$ as defined in (13), we utilize the second equality in (12a) and the first equality in (12b) to construct the following:
\[
\tilde{U}_{\ell+2}^T U_{\ell+1} = \Phi_{\ell+1}^{-1} \Theta_\ell.
\] (14)

Combining the second equality in (12a) with (14), we can observe that
\[
\tilde{U}_{\ell+2}^T U_{\ell+1} = \left(M_{\ell+2}^T B \Phi_{\ell+1}\right) \left(\Phi_{\ell+1}^{-1} \Theta_\ell\right)
\]
\[
= M_{\ell+2}^T B \Theta_\ell.
\] (15)

Hence, we have
\[
U_{\ell+2} = \tilde{U}_{\ell+2}^T U_{\ell+1} = M_{\ell+2}^T B \Theta_\ell.
\] (16)

Thus, performing top-$K$ SVD as in (11) followed by (16) allows one to identify the bases $U_\ell, U_{\ell+1}$ and $U_{\ell+2}$ such that
\[
\text{range}([U_\ell^T, U_{\ell+1}^T, U_{\ell+2}^T]^T) = \text{range}([M_\ell^T, M_{\ell+1}^T, M_{\ell+2}^T]^T).
\]

Applying (11) and (16) for $r = T = \lfloor L/2 \rfloor$, the bases $U_T, U_{T+1}$ and $U_{T+2}$ can be identified. By letting $\Theta_r = \Theta$, we immediately get
\[
U_T = M_T^T B \Theta, \quad U_{T+1} = M_{T+1}^T B \Theta, \quad U_{T+2} = M_{T+2}^T B \Theta.
\]

Next, we use induction to show how the iterations in Algorithm 1 (lines 5-8) identify $U_{t+3}, \ldots, U_L$. The induction hypothesis is that before any iteration at $r \in \{T + 1, \ldots, L - 1\}$, the following bases are identified from the previous iterations:
\[
U_t = M_t^T B \Theta, \quad \text{for } t = T + 1, \ldots, r - 1.
\] (17)
Note that (18) is satisfied for \( r = T + 1 \) from lines 1-3 of Algorithm 1 by performing top-\( K \) svd of \( C_T \).

By performing (11) and (16) for \( r = T + 1, \ldots, L - 1 \), we get

\[
U_{t_r} = M_{t_r}^T B \Theta, \quad \text{for } t = T + 1, \ldots, L.
\]  

(18)

Therefore, by induction, Algorithm 1 identifies the bases \( U_1, \ldots, U_{r+2} \) such that

\[
\text{range}(U_1^T, U_2^T, \ldots, U_{r+2}^T) = \text{range}(M_1^T, M_2^T, \ldots, M_{r+2}^T).
\]

Repeating this until \( r = L - 1 \), Algorithm 1 outputs \( \hat{U} \).

### A. Estimation of the subspace

In Theorem 1, the subspace identifiability is established via successively applying the truncated top-\( K \) SVD on \( C_r := [P_{r,n,0}^T, \ldots, P_{r,n,m_r}^T]^T \) where \( \{\ell_r\}_{r=1}^L = \{1, \ldots, L\} \) and \( m_r \in \{1, \ldots, L\} \).

To simplify the notations in the proof, we first analyze a simpler pattern where the truncated top-

\[
\hat{\sigma}_r := \min_{\ell_r = 1}^{K, r} C, \quad \text{where } C = \min_{r \in \{1, \ldots, L\}} \sigma_{\min}(C_r). \quad \text{Assume that the following holds:}
\]

\[
\text{rank}(C_r) = K, \quad \|\hat{C}_r - C_r\|_2 \leq \|C_r\|_2. \quad (20)
\]

Then, there exists an orthogonal matrix \( O_r \),

\[
\|\hat{U}_r - U_r O_r\|_F \leq \frac{6\sqrt{2K} \sigma_1 \|\hat{C}_r - C_r\|_2}{\sigma_K}. \quad (21)
\]

In addition, \( \|\hat{U}_{r+1} - U_{r+1} O_r\|_F \) is also upper bounded by the R.H.S. of (21).

The proof of the lemma is given in Sec. D.

To utilize Lemma 1 in our problem, we need to characterize \( \|\hat{C}_r - C_r\|_2 \). To this end, we present the following lemma:

### Lemma 2

Let \( \rho := \max_{i,j} P(i,j) \). Assume that the below holds true

\[
L \leq \frac{4\rho N}{d}, \quad (N/L) \geq \frac{c_0 \log(2N/L)}{2\rho}. \quad (22)
\]

Then, there exists a constant \( c_1 = c(c_0) \) such that for every \( r \in \{1, \ldots, L - 1\} \),

\[
\|\hat{C}_r - C_r\|_2 \leq 2c_1 \sqrt{\rho N/L}. \quad (23)
\]

holds with probability at least \( 1 - L(L - 1)/(2N) \).

The proof of the lemma is given in Sec. E.

Combining Lemma 2 with (21) in Lemma 1, the following two inequalities, i.e.,

\[
\|\hat{U}_r - U_r O_r\|_F \leq \phi, \quad \|\hat{U}_{r+1} - U_{r+1} O_r\|_F \leq \phi \quad \text{hold with probability at least } 1 - L(L - 1)/(2N) \text{ where } \phi = \frac{12c_1 \sqrt{2K \rho(N/L) \sigma_1}}{\sigma_K^2}.
\]
Next, we characterize the estimation accuracy of the bases $U_1, \ldots, U_L$. The iterative steps (lines 5-10) in Algorithm 1 performs the below operation to identify the basis $U_{r+1}$ for $r = T + 1, \ldots, L$ (where $T = \lceil L/2 \rceil$):

$$U_{r+1} = \underbrace{U_{r+1}}_{\Theta_r} - U_{r+1} O_r.$$ 

Therefore for certain orthogonal matrices $O_r$ and $O_r^*$, we can write

$$U_{r+1} O_r = (U_{r+1} O_r)(O_r^T \Theta_r O_r), \quad (24)$$

where $\Theta_r = U_r^T U_r$.

In order to characterize the estimation accuracy of the basis $U_{r+1}$, we need to characterize each term in the R.H.S of (24). The term $U_{r+1} O_r$ can be characterized by applying Lemma 1. The other term can be analyzed by using the following lemma:

**Lemma 3.** Consider the noiseless model $U = V \Theta$ where $\Theta$ is an orthogonal matrix and the columns of $U, V$ are orthonormal. Consider the noisy estimates $\hat{V}$ and $\hat{U}$. Assume that there exist constants $\phi_1, \phi_2 > 0$ such that $\|\hat{V} - VO_1\|_2 < \phi_1$, and $\|\hat{U} - UO_2\|_2 < \phi_2$ and $\phi_1 \leq 1/2$ for certain orthogonal matrices $O_1$ and $O_2$, Then we have

$$\|\hat{\Theta} - O_1^T \Theta O_2^T\|_2 < 2\phi_1 + 2\phi_2,$$

where $\hat{\Theta} = \arg \min Z \|\hat{U} - \hat{V} Z\|_F^2$.

The proof of the lemma can be found in Sec. F.

Next, we need to characterize the accuracy of (24). For this, we have the following lemma:

**Lemma 4.** Suppose that there exists a set of $I$ matrices $G_1, G_2, \ldots, G_I$ such that the matrix multiplication $G_1 G_2 \ldots G_I$ exists. For each $i \in \{1, \ldots, I\}$ consider another matrix of the same size $\hat{G}_i$. Assume that $\|G_i\|_2 \leq 1$ for any $i \in \{2, \ldots, I\}$, Also assume that $\|\hat{G}_i\|_2 \leq 1$ for any $i \in \{1, \ldots, I-1\}$, then

$$\|\hat{G}_1 \hat{G}_2 \ldots \hat{G}_I - G_1 G_2 \ldots G_I\|_2 \leq \sum_{i=1}^I \|\hat{G}_i - G_i\|_2$$

(25)

The proof can be found in Sec. G.

Note that since $\hat{U}_r$ and $U_r$ are semi-orthogonal matrices, $\Theta_r$ is orthogonal. Therefore, for orthogonal $O_r$ and $O_r^*$

$$\|O_r^T \Theta_r O_r\|_2 = 1. \quad (26)$$

Also, since $\hat{U}_{r+1}, \hat{U}_r$ and $U_r$ are obtained by truncated top-$K$ SVD operation, we have

$$\|\hat{U}_{r+1}\|_2 = 1, \quad \|\hat{\Theta}_r\|_2 = \|\hat{U}_r U_r\|_2 = 1 \quad (27)$$

Eqs. (26) and (27) make sure that the conditions in Lemma 4 are satisfied.

Therefore, invoking Lemma 4 on (24), we get for any $i \geq 2$,

$$\|\hat{U}_{r+1} - U_{r+1} O_r\|_F \leq \sqrt{K} \|\hat{U}_{r+1} - U_{r+1} O_r\|_2$$

$$\leq \sqrt{K} \|\hat{U}_{r+1} - U_{r+1}\|_2 + \sqrt{K} \|\hat{\Theta}_r - O_r^T \Theta_r O_r^T\|_2$$

$$\leq \sqrt{K} \phi + 3\sqrt{K} \phi + 2\sqrt{K} \|\hat{U}_r - U_r O_r\|_F$$

$$\leq 4\sqrt{K} \phi + 2\sqrt{K} \|\hat{U}_r - U_r O_r\|_F,$$

where $\phi = \frac{12K\sqrt{2Kp[N/L]p_1}}{\sigma_{\epsilon}^2}$, the first inequality is obtained by using norm equivalence, the second inequality is obtained by applying Lemma 4 and the third inequality is by applying Lemmas 1 and 3.

We further have

$$\|\hat{U}_{r+1} - U_{r+1} O_r\|_F \leq 4\sqrt{K} \phi + 2\sqrt{K} \left(4\sqrt{K} \phi + 2\sqrt{K} \|\hat{U}_{r-1} - U_{r-1} O_r\|_F\right)$$

$$\leq 4\sqrt{K} \phi + 2\sqrt{K} \left(4\sqrt{K} \phi + 2\sqrt{K} \left(4 \sqrt{K} \phi + 2\sqrt{K} \|\hat{U}_{r-2} - U_{r-2} O_r\|_F\right)\right)$$

$$\leq 4\sqrt{K} \phi \left(1 + 2\sqrt{K} + (2\sqrt{K})^2 + \ldots + (2\sqrt{K})^{r-1}\right).$$

Therefore, for $r = T + 1, \ldots, L - 1$, we have

$$\|\hat{U}_{r+1} - U_{r+1} O_r\|_F \leq \frac{4\sqrt{K} \phi (2\sqrt{K})^{r-T}}{2\sqrt{K} - 1}. \quad (28)$$
Algorithm 1 uses the below operation iteratively (lines 12-18) in order to estimate the basis $U_1, \ldots, U_{T-1}$ and is executed in the reverse order:

$$U_{r-1} = U_{r-1}U_r^\dagger U_r,$$

Using the similar steps as before, we can easily show that, for $r = 2, \ldots, T$,

$$\|\hat{U}_{r-1} - U_{r-1}O\|_F \leq \frac{4\sqrt{K}\phi(2\sqrt{K})^{T-r+1}}{2\sqrt{K} - 1}. \quad (29)$$

Note that Algorithm 1 first estimates $U_T$ and $U_{T+1}$ by performing truncated top-$K$ SVD to $\hat{C}_T$. Therefore, by directly invoking Lemma 1, we have $\|\hat{U}_T - U_TO\|_F \leq \phi$ and $\|\hat{U}_{T+1} - U_{T+1}O\|_F \leq \phi$.

Then, we have

$$\|\hat{U} - UO\|_F \leq \sum_{r=1}^L \|\hat{U}_r - U_rO\|_F$$

$$= \sum_{r=1}^{T-1} \|\hat{U}_r - U_rO\|_F + \|\hat{U}_T - U_TO\|_F + \|\hat{U}_{T+1} - U_{T+1}O\|_F + \sum_{r=T+2}^L \|\hat{U}_r - U_rO\|_F$$

$$\leq 2\phi(2\sqrt{K})^{L/2} + 2\phi \leq 3\phi(2\sqrt{K})^{L/2} \leq (2\sqrt{K} - 1)^2.$$

Substituting $\phi = \frac{12\sqrt{2K}\rho(N/L)\sigma_1}{\sigma_K}$ into the above, we get the below with probability at least with probability at least $1 - L(L - 1)/(2N)$,

$$\|\hat{U} - UO\|_F \leq \frac{48\sqrt{2K}c(2\sqrt{K})^{(L/2)}\sigma_1\sqrt{\rho(N/L)}}{(2\sqrt{K} - 1)^2\sigma_K^2}. \quad (30)$$

Next, we obtain upper and lower bounds for $\sigma_1$ and $\sigma_K$, respectively. To this end, we have the following lemma:

**Lemma 5.** Assume that the columns of $M$ are generated from a continuous distribution. Also assume that $(N/L) \geq \frac{2}{\epsilon} \log(NK/L)$. Then, there exists constants $c$ and $C$ where $c \leq C$ depending only on distribution of the columns of $M$ such that the following holds true with probability at least $1 - (3L(L - 1)/N)$.

$$\sigma_K \geq \sqrt{2\sigma_{\min}(B)c(N/L)}, \quad (31)$$

$$\sigma_1 \leq \sqrt{2\sigma_{\max}(B)C(N/L)}. \quad (32)$$

The proof can be found in Sec. H.

Applying (31), (32) in (30), we have with probability at least $1 - (7L(L - 1)/(2N))$,

$$\|\hat{U} - UO\|_F \leq \frac{48\sqrt{2K}c1(2\sqrt{K})^{(L/2)}\sigma_1\sqrt{\rho(N/L)}}{(2\sqrt{K} - 1)^2\sigma_K^2} \leq \frac{48c1\sqrt{K}(2\sqrt{K})^{(L/2)}\kappa(B)\sqrt{\rho}}{c^2\sigma_{\min}(B)(2\sqrt{K} - 1)^2\sqrt{(N/L)}}. \quad (33)$$

To make the above derivation legitimate, we need the assumptions in Lemma 1-5 hold true. The assumptions on $C_r$ in Lemma 1 given by the equalities in (20) are satisfied if

$$\text{rank}(M_r) = \text{rank}(B) = K, \quad r = 1, \ldots, L \quad (34a)$$

$$2c_1\sqrt{\rho N/L} \leq \sigma_K \quad (34b).$$

Using (31), The condition (34b) can be further written as

$$2c_1\sqrt{\rho N/L} \leq \sqrt{2\sigma_{\min}(B)c(N/L)} \quad \Rightarrow \sqrt{\rho N/L} \geq \frac{c_1\sqrt{2\rho}}{\sigma_{\min}(B)c}. \quad (35)$$

The assumption in Lemma 3 can be written as:

$$\phi = \frac{12\sqrt{2K}\rho(N/L)\sigma_1}{\sigma_K^2} \leq \frac{12c_1\sigma_1\kappa(B)\sqrt{K\rho}}{c^2\sigma_{\min}(B)\sqrt{(N/L)}} \leq 1/2 \quad \Rightarrow \frac{(24c_1)^2\kappa^2(B)K\rho}{c^2\sigma_{\min}(B)} \quad (36)$$
We can see that the condition in (35) is satisfied if (36) holds.

Also, the below conditions from Lemma 2 and 5 have to be satisfied:

\[
L \leq \frac{4pN}{d}, \quad (N/L) \geq \frac{c_0 \log(2N/L)}{2\rho} \\
(N/L) \geq \frac{4}{e} \log(NK/L)
\]

Hence, if the conditions (34a) and (36) are satisfied along with the conditions in Lemma 2 and 5, the bound given by (33) holds true with probability at least \(1 - (7L(L - 1)/(2N))\).

Next, we will show that the analysis can be easily extended to all the patterns according Criterion 3.

\section*{Appendix C}

\textbf{Permutation Fixing Procedure for Baseline Algorithms}

For synthetic and real data experiments, we employ the following procedure for the baseline algorithms to align the permutations of the estimated \(\hat{M}_t\) over different blocks.

\begin{algorithm}[h]
\caption{Permutation Fixing Procedure}
\begin{algorithmic}[1]
\State \textbf{input:} Observed adjacency matrix blocks \(\{A_{t, \ell}\}\) and \(\{A_{t, \ell+1}\}\), \(K\) and \(L\) (where \(K \leq N/L\)), BaselineAlgorithm (SPOC/GeoNMF/CD-MVS/SC-Unnorm/SC-Norm).
\State Initialize permutation matrix \(\Pi \leftarrow I_K\) (where \(I_k\) is identity matrix of size \(K\)).
\For {\(\ell = 1\) to \(L - 1\)}
\State \(H_\ell \leftarrow \begin{bmatrix} A_{t, \ell}, A_{t, \ell+1} \\ A_{t+1, \ell}, A_{t+1, \ell+1} \end{bmatrix}\).
\State run BaselineAlgorithm with inputs \(H_\ell\) and \(K\) to output \([\hat{M}_t, \hat{M}_{t+1}]\);
\If {\(\ell > 0\)}
\State apply Hugarian algorithm \([31]\) with inputs \(\hat{M}_\ell\) and \(M_{\ell}^{(\text{prev})}\) which updates permutation matrix \(\Pi\);
\EndIf
\State \([M_t, M_{t+1}] \leftarrow \Pi^T [\hat{M}_t, \hat{M}_{t+1}];\)
\State \(M_{t}^{(\text{prev})} \leftarrow M_{t+1};\)
\EndFor
\State \textbf{output:} Estimated Membership Matrix \(\hat{M} = [M_1, \ldots, M_L]\).
\end{algorithmic}
\end{algorithm}

\section*{Appendix D}

\textbf{Proof of Lemma 1}

We invoke the following lemma from \([32]\) to characterize the estimation accuracy of the factors resulting from top-\(K\) SVD of \(\hat{C}_r\):

\textbf{Lemma 6.} \([32]\) Let \(C \in \mathbb{R}^{m \times n}\) and \(\hat{C} \in \mathbb{R}^{m \times n}\) have singular values \(\alpha_1 \geq \alpha_2 \geq \cdots \geq \alpha_{\min(m,n)}\) and \(\tilde{\alpha}_1 \geq \tilde{\alpha}_2 \geq \cdots \tilde{\alpha}_{\min(m,n)}\), respectively. Fix \(1 \leq t \leq s \leq \text{rank}(C)\) and assume that \(\min(\alpha_t^2 - \alpha_{t-1}^2, \alpha_s^2 - \alpha_{s+1}^2) > 0\), where \(\alpha_0 := \infty\) and \(\alpha_{\text{rank}(C)+1} := 0\). Let \(d := s - t + 1\) and let \(U = [u_t, u_{t+1}, \ldots, u_s] \in \mathbb{R}^{p \times d}\) and \(\hat{U} = [\hat{u}_t, \hat{u}_{t+1}, \ldots, \hat{u}_s] \in \mathbb{R}^{n \times d}\) have orthonormal columns satisfying \(C^T u_j = \alpha_j v_j\) and \(\hat{C}^T \hat{u}_j = \tilde{\alpha}_j \hat{v}_j\) for \(j = t, t+1, \ldots, s\) and let \(V = [v_t, v_{t+1}, \ldots, v_s] \in \mathbb{R}^{n \times d}\) and \(\hat{V} = [\hat{v}_t, \hat{v}_{t+1}, \ldots, \hat{v}_s] \in \mathbb{R}^{n \times d}\) have orthonormal columns satisfying \(C v_j = \alpha_j u_j\) and \(\hat{C} \hat{v}_j = \tilde{\alpha}_j \hat{u}_j\) for \(j = t, t+1, \ldots, s\). Then there exists an orthogonal matrix \(O \in \mathbb{R}^{d \times d}\) such that

\[
\|\hat{U} - UO\|_F \leq \frac{2^{3/2}(2 \alpha_1 + \|\hat{C} - C\|_2) \min(\rho^{d/2}\|\hat{C} - C\|_2, \|\hat{C} - C\|_F)}{\min(\alpha_t^2 - \alpha_{t-1}^2, \alpha_s^2 - \alpha_{s+1}^2)} \tag{37}
\]

and the same upper bound holds for \(\|\hat{V} - VO\|_F\).

By letting \(t = 1\) and \(s = K\), using the assumption that \(\text{rank}(C_r) = K\) and applying Lemma 6, we have

\[
\|\hat{U}_r - U_rO_r\|_F \leq \frac{2^{3/2}(2\sigma_{\max}(C_r) + \|\hat{C}_r - C_r\|_2) \min(\sqrt{K}\|\hat{C}_r - C_r\|_2, \|\hat{C}_r - C_r\|_F)}{\sigma_{\min}(C_r)^2} \tag{38}
\]

\[
= \frac{2^{3/2}\sqrt{K}(2\sigma_{\max}(C_r) + \|\hat{C}_r - C_r\|_2)\|\hat{C}_r - C_r\|_2}{\sigma_{\min}(C_r)^2} \tag{39}
\]

\[
\leq \frac{2^{3/2}\sqrt{K}3\sigma_{\max}(C_r)\|\hat{C}_r - C_r\|_2}{\sigma_{\min}(C_r)^2}.
\]
Lemma 7. [26] Let $A \in \mathbb{R}^{N \times N}$ be the symmetric binary adjacency matrix of a random graph. Let $A(i, j)$ denote the edge between node $i$ and node $j$. Assume that $A(i, j) \sim$ Bernoulli$(P(i, j))$ and $A(i, j)$’s are observed independently. Denote $\rho := \max_{i,j} P(i, j)$. Assume that $d \geq \max(\rho N, c_0 \log N)$ where $c_0 > 0$. Then, for any $t > 0$, there exists a constant $c_t = c(t, c_0)$ such that

$$\|A - P\|_2 \leq c_t \sqrt{d},$$

with probability greater than $1 - N^{-t}$.

To apply Lemma 7 for the proposed pattern, we first consider the below sub-adjacency matrix for any $r \in \{1, \ldots, L - 1\}$:

$$\tilde{A}_r = \begin{bmatrix} A_{r,r} & A_{r,r+1}^T \\ A_{r+1,r} & A_{r+1,r+1}^T \end{bmatrix},$$

where $\tilde{A}_r \in \mathbb{R}^{2N/L \times 2N/L}$.

Now, we apply Lemma 7 onto $\tilde{P}_r$, where

$$\tilde{P}_r = \begin{bmatrix} P_{r,r} & P_{r,r+1}^T \\ P_{r+1,r} & P_{r+1,r+1}^T \end{bmatrix}.$$

Hence, by assuming

$$d \geq \max\left(\frac{2\rho N}{L}, c_0 \log(2N/L)\right),$$

we apply Lemma 7 and obtain $\|\tilde{A}_r - \tilde{P}_r\|_2 \leq c_t \sqrt{d}$ with probability greater than $1 - (2N/L)^{-t}$. By using the assumptions

$$L \leq \frac{4\rho N}{d} \quad \text{and} \quad \frac{c_0 \log(2N/L)}{2\rho},$$

in (39) and applying Lemma 7 with $t = 1$, we get $\|\tilde{A}_r - \tilde{P}_r\|_2 \leq 2c_1 \sqrt{\rho N/L}$ with probability at least $1 - (2N/L)^{-1}$.

Then, we have

$$\|\tilde{C}_r - C_r\|_2 = \|\tilde{A}_r, A_{r+1,r+1}^T - [P_{r,r+1}^T, P_{r+1,r+1}^T]^T\|_2$$

$$\leq \|\tilde{A}_r - \tilde{P}_r\|_2$$

$$\leq 2c_1 \sqrt{\rho N/L} \quad (41)$$

holds with probability greater than $1 - (2N/L)^{-1}$. Applying union bound, (41) holds with probability greater than $1 - L(L - 1)/(2N)$ for every $r \in \{1, \ldots, L - 1\}$.
Lemma 8. [33] Consider the noiseless model $U = V \Theta$. Suppose that there exists measurement errors in the matrices $V$ and $U$ such that $\tilde{V} = V + \Delta V$ and $\tilde{U} = U + \Delta U$. Assume that there exist constants $\phi_1, \phi_2 > 0$ such that $\|\Delta V\|_2 \leq \phi_1\|V\|_2$, $\|\Delta U\|_2 \leq \phi_2\|U\|_2$ and $\phi_1 \kappa(V) < 1$. Then we have

$$\frac{\|\hat{\Theta} - \Theta\|_2}{\|\Theta\|_2} \leq \frac{\kappa(V)(\phi_1 + \phi_2)}{1 - \phi_1 \kappa(V)} + \kappa(V)\phi_1,$$

where $\kappa(V) = \frac{\sigma_{\max}(V)}{\sigma_{\min}(V)}$ and $\hat{\Theta}$ is the least squares solution to the problem $\hat{U} = \tilde{V}\Theta$.

Since $\Theta, O_1$ and $O_2$ are orthogonal matrices and the columns of $U, V$ are orthonormal, we have

$$\|U\|_2 = \|U_O\|_2 = 1, \quad \|V\|_2 = \|VO_1\|_2 = 1, \quad \|\Theta\|_2 = \|O_1^T \Theta O_2\|_2 = 1, \quad \kappa(V) = 1.$$

Then, using the assumption that $\phi_1 \leq 1/2$ and applying Lemma 8, we finally have

$$\|\hat{\Theta} - O_1^T \Theta O_2\|_2 \leq 3\phi_1 + 2\phi_2,$$

where $\hat{\Theta}$ is the least squares solution to the problem $\hat{U} = \tilde{V}\Theta'$.

APPENDIX G

PROOF OF LEMMA 4

Consider the left hand side of (25):

$$\|\hat{G}_1 \hat{G}_2 \ldots \hat{G}_I - G_1 G_2 \ldots G_I\|_2 = \|\hat{G}_1 \hat{G}_2 \ldots \hat{G}_I - \hat{G}_1 G_2 \ldots G_I + \hat{G}_1 G_2 \ldots G_I\|_2$$

$$= \|\hat{G}_1 (\hat{G}_2 - G_2) \ldots \hat{G}_I - G_2 \ldots G_I + \hat{G}_1 G_2 \ldots G_I\|_2$$

$$\leq \|\hat{G}_1 (\hat{G}_2 - G_2) \ldots \hat{G}_I\|_2 + \|\hat{G}_1 G_2 \ldots G_I\|_2$$

$$\leq \|\hat{G}_1 (\hat{G}_2 - G_2) \ldots \hat{G}_I + \hat{G}_1 G_2 \ldots G_I\|_2 + \|\hat{G}_1 G_2 \ldots G_I\|_2$$

$$\leq \|\hat{G}_1 (\hat{G}_2 - G_2) \ldots \hat{G}_I\|_2 + \|\hat{G}_1 G_2 \ldots G_I\|_2 + \ldots$$

$$+ \|\hat{G}_1 G_2 \ldots G_{I-2} (\hat{G}_{I-1} - G_{I-1}) G_I\|_2$$

where we have applied triangle inequality to obtain the first and last inequalities. Using the assumption that $\|G_i\|_2 \leq 1$ for $i \in \{2, \ldots, I\}$ and $\|\hat{G}_i\|_2 \leq 1$ for $i \in \{1, \ldots, I-1\}$ and also by triangle inequality, we finally have

$$\|\hat{G}_1 \hat{G}_2 \ldots \hat{G}_I - G_1 G_2 \ldots G_I\|_2 \leq \sum_{i=1}^{I} \|\hat{G}_i - G_i\|_2.$$

APPENDIX H

PROOF OF LEMMA 5

Let us first define the following matrix:

$$S_r := [P_{r,r}, P_{r,r+1}] = M_r^T B[M_r, M_{r+1}].$$

Consider the $K$-th eigenvalue of the symmetric matrix $S_r^T S_r$:

$$\lambda_K(S_r^T S_r) = \lambda_K([M_r, M_{r+1}]^T B M_r, M_{r+1}^T B[M_r, M_{r+1}])$$

$$= \lambda_K([M_r, M_{r+1}]^T B M_r, M_{r+1}^T B^{1/2} B^{1/2}[M_r, M_{r+1}]),$$

where $\lambda_K$ denotes the $K$-th eigenvalue with $\lambda_1 \geq \ldots \geq \lambda_K$.

We utilize the below fact to proceed further:

**Fact 1.** [Theorem 1.3.22 [34]] Consider two matrices $W_1 \in \mathbb{R}^{N \times K}$ and $W_2 \in \mathbb{R}^{K \times N}$. Let $\lambda_1, \ldots, \lambda_K$ be the nonzero eigenvalues of the matrix $W_1 W_2$. Then, the matrix $W_2 W_1$ also holds the same set of eigenvalues.

Combining Fact 1 and (42), we have

$$\lambda_K(S_r^T S_r) = \lambda_K(B^{1/2}[M_r, M_{r+1}] [M_r, M_{r+1}]^T B M_r, M_{r+1}^T B^{1/2}).$$

(43)
Next, we consider the following matrix
\[ H_r := M_r M_r^\top = \sum_{n=1}^{N/L} M_r(:,n)M_r(:,n)^\top. \] (44)
Since the columns of \( M \) are assumed to be generated from a continuous distribution, we have
\[ \mathbb{E}[H_r] = \frac{N}{L} \mathbb{E}[M_r(:,1)M_r(:,1)^\top] \]
and \( \mathbb{E}[H_r] \) is positive-definite for every \( r \in \{1, \ldots, L\} \). Then we have
\[ \lambda_K(\mathbb{E}[H_r]) = O(N/L), \quad \lambda_1(\mathbb{E}[H_r]) = O(N/L). \]

To proceed further, we have the following lemma to bound \( \lambda_K(H_r) \) and \( \lambda_1(H_r) \).

**Lemma 9.** Consider the definition of \( H_r \) given by (44). Suppose \( (N/L) \geq \frac{2}{3} \log(NK/L) \). Then, there exists constants \( c \) and \( C \) where \( c \leq C \) depending only on distribution of the columns of \( M \) such that the following holds true:
\[ \Pr(\lambda_K(H_r) \geq c(N/L)) \geq 1 - (N/L)^{-1}, \]
\[ \Pr(\lambda_1(H_r) \leq C(N/L)) \geq 1 - (N/L)^{-1}. \]

The proof of the lemma is given in Section J.

Using Lemma 9, we can also have
\[ \lambda_K([M_r,M_{r+1}][M_r,M_{r+1}]^\top) \geq c(2N/L), \]
\[ \lambda_1([M_r,M_{r+1}][M_r,M_{r+1}]^\top) \leq C(2N/L), \]
each holding with probability at least \( 1 - (2N/L)^{-1} \).

Thus, applying Lemma 9 twice in (43), we get
\[ \lambda_K(S_r^\top S_r) \geq \lambda_K^2(B)\lambda_K([M_r,M_{r+1}][M_r,M_{r+1}]^\top)\lambda_K(M_r M_r^\top) \]
\[ \geq \lambda_K^2(B)\lambda_1(M_r M_r^\top) = 2\lambda_K^2(B)c^2(N/L)^2, \]
\[ = 2\sigma_{\text{min}}^2(B)c^2(N/L)^2, \]
with probability at least \( 1 - (2N/3L)^{-1} \). Note that \( \lambda_K(B) = \sigma_{\text{min}}(B) \) since \( B \) is symmetric.

Similarly, we can get the below with probability at least \( 1 - (2N/3L)^{-1} \):
\[ \lambda_1(S_r^\top S_r) \leq \lambda_1^2(B)\lambda_1([M_r,M_{r+1}][M_r,M_{r+1}]^\top)\lambda_1(M_r M_r^\top) \]
\[ \leq \lambda_1^2(B)C(2N/L)C(N/L) = 2\lambda_1^2(B)C^2(N/L)^2, \]
\[ = 2\sigma_{\text{max}}^2(B)C^2(N/L)^2. \]

Therefore, by taking union bound over \( r \in \{1, \ldots, L - 1\} \), with probability at least \( 1 - (3L(L-1)/N) \), the following equations hold simultaneously for every \( r \):
\[ \sigma_{\text{min}}(P_{r,r}, P_{r,r+1}) = \sigma_{\text{min}}(S_r) = \sqrt{\lambda_K(S_r^\top S_r)} \geq \sqrt{2}\sigma_{\text{min}}(B)c(N/L), \] (45)
\[ \sigma_{\text{max}}(P_{r,r}, P_{r,r+1}) = \sigma_{\text{max}}(S_r) = \sqrt{\lambda_1(S_r^\top S_r)} \leq \sqrt{2}\sigma_{\text{max}}(B)C(N/L). \] (46)

The equations (45) and (46) immediately follow that the below holds with probability at least \( 1 - (3L(L-1)/N) \):
\[ \sigma_K = \min_{r \in \{1, \ldots, L-1\}} \sigma_{\text{min}}(P_{r,r}, P_{r,r+1}) \geq \sqrt{2}\sigma_{\text{min}}(B)c(N/L), \]
\[ \sigma_1 = \max_{r \in \{1, \ldots, L-1\}} \sigma_{\text{max}}(P_{r,r}, P_{r,r+1}) \leq \sqrt{2}\sigma_{\text{max}}(B)C(N/L). \]

**APPENDIX I**

**PROOF OF FACT 1**

Let \( v_k \in \mathbb{R}^N \) be the eigenvector of the product \( W_1 W_2 \in \mathbb{R}^{N \times N} \) corresponding to the eigenvalue \( \lambda_k \). It follows that
\[ W_1 W_2 v_k = \lambda_k v_k. \] (47)
After multiplying \( W_2 \in \mathbb{R}^{K \times N} \) on both sides of (47), we get
\[ W_2 W_1 W_2 v_k = W_2 \lambda_k v_k, \]
\[ \Rightarrow W_2 W_1 (W_2 v_k) = \lambda_k (W_2 v_k), \]
\[ \Rightarrow W_2 W_1 w_k = \lambda_k w_k, \]
where \( w_k = W_2 v_k \in \mathbb{R}^K \) is the eigenvector of \( W_2 W_1 \) corresponding to the eigenvalue \( \lambda_k \). Note that \( w_k = W_2 v_k \neq 0 \) since \( W_1 W_2 v_k = \lambda_k v_k \neq 0 \). It implies that all the nonzero eigenvalues \( \lambda_1, \ldots, \lambda_K \) of \( W_1 W_2 \) are the eigenvalues of \( W_2 W_1 \) also.

This completes the proof.

**Appendix J**

**Proof of Lemma 9**

We apply the following lemma from [20] in order to bound \( \lambda_1(H_r) \) and \( \lambda_K(H_r) \):

**Lemma 10.** [20] There exists constants \( c \) and \( C \) depending only on distribution of the columns of \( M \) such that

\[
\Pr \left( \lambda_K \left( \sum_{n=1}^{N/L} M_r(:,n)M_r(:,n)^T \right) \leq c(N/L) \right) \leq K \exp(-c(N/4L)),
\]

\[
\Pr \left( \lambda_1 \left( \sum_{n=1}^{N/L} M_r(:,n)M_r(:,n)^T \right) \geq C(N/L) \right) \leq \frac{K}{2CN/L}.
\]

To simplify the bound on the probability, let us find the conditions at which the following is satisfied:

\[
K \exp(-c(N/4L)) \leq (N/L)^{-1},
\]

\[
\frac{K}{2CN/L} \leq (N/L)^{-1}.
\]

The conditions (48) and (49) are equivalent to the following:

\[
(N/L) \geq \frac{4}{c} \log(NK/L),
\]

\[
(N/L) \geq \log_2 e \frac{C}{c} \log(NK/L).
\]

We can immediately see that if (50) is satisfied, (51) is also satisfied since \( c \leq C \).

Therefore, we get that if \( (N/L) \geq \frac{4}{c} \log(NK/L) \), with probability at least \( 1 - (N/L)^{-1} \),

\[
\lambda_K \left( \sum_{n=1}^{N/L} M_r(:,n)M_r(:,n)^T \right) \leq c(N/L),
\]

\[
\lambda_1 \left( \sum_{n=1}^{N/L} M_r(:,n)M_r(:,n)^T \right) \geq C(N/L).
\]