LEARNING MIXED MEMBERSHIP FROM ADJACENCY GRAPH VIA SYSTEMATIC EDGE QUERY: IDENTIFIABILITY AND ALGORITHM

*Shahana Ibrahim and Xiao Fu**

School of Electrical Engineering & Computer Science
Oregon State University
Corvallis, OR 97331, United States
(ibrahish,xiao.fu)@oregonstate.edu

ABSTRACT

Graph clustering is a core technique for network analysis problems, e.g., community detection. This work puts forth a node clustering approach for largely incomplete adjacency graphs. Under the considered scenario, instead of having access to the complete graph, only a small amount of queries about the graph edges can be made for node clustering. This task is well-motivated in many large-scale network analysis problems, where complete graph acquisition is prohibitively costly. Prior work tackles this problem under the setting that the nodes only admit single membership and the clusters are disjoint, 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 offers a framework that provably learns the mixed membership of nodes from overlapping clusters using limited edge information. Our method is equipped with a systematic edge query pattern, which is arguably easier to implement relative to the random counterparts in certain applications, e.g., field survey based graph analysis. A lightweight scalable algorithm is proposed, and its performance characterizations are presented. Numerical experiments are used to showcase the effectiveness of our method.

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

1. 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., social network analysis [2], brain signal processing [3], and biological/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 with the full graph observed have been extensively studied in the past two decades [6–9]. Recently, GC under partial observation of the graph edges have drawn increasing attention. Notably, many network data have grown prohibitively large—e.g., social media networks from Facebook and Twitter could easily contain billions of edges (i.e., user-user links). Edge acquisition and the subsequent computational tasks at such a scale is highly nontrivial. In addition, in some networks, edges are intentionally removed or hidden (e.g., terrorist networks or radical group networks) [10], and the ability to retrieve information from such partially observed graphs is also critical. 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 [11].

A number of works [12–14] have considered the graph clustering problem under incomplete edge observation. There, the node membership identification guarantees were established under the assumption that every node is associated 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). In addition, the edges were queried randomly in these existing approaches, which may not be easy to implement in some applications; e.g., in field survey based network analysis spanning a large geographic area [15]—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 [12–14] 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) optimization 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 random edge query-based GC methods in [12–14], we propose a carefully designed systematic edge query principle to enable provable GC. 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 that only consists of the truncated singular value decomposition (SVD) of small matrices and a Gram-Schmidt-like greedy algorithm. We also provide theoretical guarantees, i.e., membership identifiability and estimation accuracy, to support our design. We conduct numerical evaluation on synthetic and real-world datasets to demonstrate the effectiveness of the proposed approach.

2. 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 belongs to cluster k with

*This work is supported in part by the National Science Foundation under Project NSF IIS-2007836 and in part by the Army Research Office (ARO) under Project ARO W911NF-19-1-0407.

probability $m_{k,n}$, where $\sum_{k=1}^{K} m_{k,n} = 1,\ m_{k,n} \geq 0$. Then, the vector $\boldsymbol{m}_n = [m_{1,n}, \ldots, m_{K,n}]^\top$ is referred to as the membership vector of entity n. All such vectors together constitute the membership matrix $\boldsymbol{M} = [\boldsymbol{m}_1, \ldots, \boldsymbol{m}_N] \in \mathbb{R}^{K \times N}$. In GC, the entities are the nodes of the graph, and the relationship between the nodes are represented by the edges of the graph. In this work, we consider graphs that are symmetric adjacency matrices, i.e., $\boldsymbol{A} \in \{0,1\}^{N \times N}$, where each entry $\boldsymbol{A}(i,j)$ encodes the pairwise relationship between nodes i and j. Our goal is to learn $\boldsymbol{M}$ using a limited number of edges in $\boldsymbol{A}$; i.e, the number of edges queried is much smaller than total number of edges, i.e., $N(N-1)/2$.

Our problem setting is motivated by a number of important applications. The mixed membership learning is a core task in *overlapped community detection* (OCD) [16]. In the OCD frameworks proposed in [17–19], the mixed membership was provably learned using a fully observed $\boldsymbol{A}$. However, OCD under partially observed edges is of great interest for applications like field survey based OCD [15] and hidden edge-robust network analysis [10]. In both cases, one cannot observe the entire $\boldsymbol{A}$ due to various reasons, e.g., resource constraints and difficulty of edge acquisition.

To handle GC under partial edge observations, the work in [12–14] adopted a generative model of $\boldsymbol{A}$ where every node admits a single cluster membership, i.e., the *stochastic block model* (SBM) [20]. The SBM can be summarized as follows. Each node n belongs to a single cluster k, i.e., the membership vector $\boldsymbol{m}_n$ is the kth unit vector. Let $\boldsymbol{B} \in \mathbb{R}^{K \times K}$ represent a cluster-cluster interaction matrix, where $\boldsymbol{B}(p,q)$ represents the probability that cluster p connects with cluster q. Then, the probability that $\boldsymbol{A}(i,j) = 1$ (i.e., nodes i,j are connected) is $\boldsymbol{P}(i,j) = \boldsymbol{m}_i^\top \boldsymbol{B} \boldsymbol{m}_j$, i.e., $\boldsymbol{A}(i,j) \sim \mathsf{Bernoulli}\left(\boldsymbol{m}_i^\top \boldsymbol{B} \boldsymbol{m}_j\right)$. In other words, the adjacency matrix $\boldsymbol{A}$ is sampled from Bernoulli distributions specified by the corresponding entries of $\boldsymbol{P} = \boldsymbol{M}^\top \boldsymbol{B} \boldsymbol{M}$. In the random edge query-based GC methods [12–14], convex optimization based matrix completion criteria were proposed to impute the unobserved edges, and the recoverability of $\boldsymbol{A}$ was established under the SBM. The results from [12–14] are insightful. However, as discussed, in many applications, mixed membership is of more interest and/or random edge queries are not easy to implement.

3. PROPOSED APPROACH

Our approach relaxes the single membership assumption in SBM by allowing the nodes to be associated with multiple clusters, i.e., the mixed membership case. We assume that the membership matrix $\boldsymbol{M}$ satisfies

$$\mathbf{1}^\top \boldsymbol{M} = \mathbf{1}^\top, \quad \boldsymbol{M} \geq \mathbf{0}; \tag{1}$$

i.e., $\boldsymbol{m}_n$ resides in the probability simplex, instead of being the unit vectors as in SBM. Under (1), the Bernoulli model used in SBM, i.e.,

$$\boldsymbol{A}(i,j) \sim \mathsf{Bernoulli}\left(\boldsymbol{P}(i,j)\right), \tag{2}$$

where $\boldsymbol{P}(i,j) = \boldsymbol{m}_i^\top \boldsymbol{B} \boldsymbol{m}_j$, is adopted in our generative model for the adjacency matrix $\boldsymbol{A}$. Overall, (1) and (2) present a model that is reminiscent of the *mixed membership stochastic block* (MMSB) model in OCD [21].

3.1. Systematic Edge Query

Our goal is to learn $\boldsymbol{M}$ from systematically sampled edges of $\boldsymbol{A}$. To proceed, we first divide the nodes into L disjoint groups $\mathcal{S}_1, \ldots, \mathcal{S}_L$ such that $\mathcal{S}_1 \cup \cdots \cup \mathcal{S}_L = [N]$ (where $[N] = \{1, \ldots, N\}$). Let $\boldsymbol{A}_{\ell,m} \in \mathbb{R}^{|\mathcal{S}_\ell| \times |\mathcal{S}_m|}$ denote the adjacency submatrix between groups $\mathcal{S}_\ell$ and $\mathcal{S}_m$, where $|\mathcal{S}_m|$ denotes the cardinality of the set $\mathcal{S}_m$. We propose an edge query principle as follows:

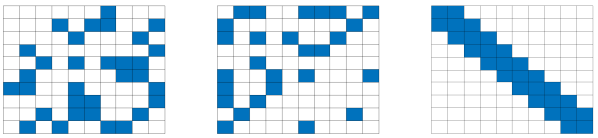

Fig. 1. Some patterns for $\boldsymbol{A}$ following EQP with $N = 1000$, $K = 5$ and $L = 10$. The shaded blue region represents the blocks queried.

Edge Query Principle (EQP):
- For every $\ell \in [L]$, $K \leq |\mathcal{S}_\ell|$ holds.
- Let $m_r \in [L]$ and $\{\ell_r\}_{r=1}^{L} = [L]$. For every ℓ_r, there exists a pair of indices m_r and ℓ_{r+1} where $\ell_{r+1} \neq \ell_r$ such that the edges from the blocks $\boldsymbol{A}_{\ell_r,m_r}$ and $\boldsymbol{A}_{\ell_{r+1},m_r}$ 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. Furthermore, instead of sampling individual edges, we sample *blocks* of edges under the EQP. This allows us to design a provable and lightweight algorithm for mixed membership learning.

3.2. Main Idea: Block Subspace Stitching

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

To shed some light on how Algorithm 1 identifies $\boldsymbol{U}$, let us consider the ideal case where $\boldsymbol{A}_{\ell,m} = \boldsymbol{P}_{\ell,m} = \boldsymbol{M}_\ell^\top \boldsymbol{B} \boldsymbol{M}_m$. We show the main idea by analyzing a toy example with $L = 3$ and the following blocks are queried following the EQP (also see Fig. 2):

$$\boldsymbol{P}_{1,2} = \boldsymbol{M}_1^\top \boldsymbol{B} \boldsymbol{M}_2\ ,\ \boldsymbol{P}_{2,2} = \boldsymbol{M}_2^\top \boldsymbol{B} \boldsymbol{M}_2\ , \tag{3}$$

$$\boldsymbol{P}_{2,1} = \boldsymbol{M}_2^\top \boldsymbol{B} \boldsymbol{M}_1\ ,\ \boldsymbol{P}_{3,1} = \boldsymbol{M}_3^\top \boldsymbol{B} \boldsymbol{M}_1\ . \tag{4}$$

Define $\boldsymbol{C}_1 := [\boldsymbol{P}_{1,2}^\top, \boldsymbol{P}_{2,2}^\top]^\top$ and $\boldsymbol{C}_2 := [\boldsymbol{P}_{2,1}^\top, \boldsymbol{P}_{3,1}^\top]^\top$. The truncated top-$K$ SVD of $\boldsymbol{C}_1$ and $\boldsymbol{C}_2$ can be represented as follows:

$$\boldsymbol{C}_1 = [\boldsymbol{U}_1^\top, \boldsymbol{U}_2^\top]^\top \boldsymbol{\Sigma} \boldsymbol{V}^\top,\ \boldsymbol{C}_2 = [\widetilde{\boldsymbol{U}}_2^\top, \widetilde{\boldsymbol{U}}_3^\top]^\top \widetilde{\boldsymbol{\Sigma}} \widetilde{\boldsymbol{V}}^\top. \tag{5}$$

Combining (3)-(5), and under the assumption that $\mathrm{rank}(\boldsymbol{M}) = \mathrm{rank}(\boldsymbol{B}) = K$ and $K \leq |\mathcal{S}_\ell|$, one can express the bases of $\mathsf{range}(\boldsymbol{M}_1^\top)$, $\mathsf{range}(\boldsymbol{M}_2^\top)$ and $\mathsf{range}(\boldsymbol{M}_3^\top)$ as $\boldsymbol{U}_1 = \boldsymbol{M}_1^\top \boldsymbol{B} \boldsymbol{\Theta}$, $\boldsymbol{U}_2 = \boldsymbol{M}_2^\top \boldsymbol{B} \boldsymbol{\Theta}$ and $\widetilde{\boldsymbol{U}}_3 = \boldsymbol{M}_3^\top \boldsymbol{B} \boldsymbol{\Phi}$, respectively, where $\boldsymbol{\Theta} \in \mathbb{R}^{K \times K}$ and $\boldsymbol{\Phi} \in \mathbb{R}^{K \times K}$ are certain nonsingular matrices. Our hope is to "stitch" the bases above to have

$$\mathsf{range}([\boldsymbol{U}_1^\top, \boldsymbol{U}_2^\top, \boldsymbol{U}_3^\top]^\top) = \mathsf{range}([\boldsymbol{M}_1, \boldsymbol{M}_2, \boldsymbol{M}_3]^\top). \tag{6}$$

with $\boldsymbol{U}_1$ and $\boldsymbol{U}_2$ in (5) and a certain $\boldsymbol{U}_3$. Note that the $\widetilde{\boldsymbol{U}}_3$ from (5) cannot be directly combined with $\boldsymbol{U}_1$ and $\boldsymbol{U}_2$ to attain the above, since $\boldsymbol{\Theta} = \boldsymbol{\Phi}$ does not hold in general. To fix this, we define $\boldsymbol{U}_3 := \widetilde{\boldsymbol{U}}_3 \widetilde{\boldsymbol{U}}_2^\dagger \boldsymbol{U}_2$. It is not hard to see that

$$\widetilde{\boldsymbol{U}}_3 \widetilde{\boldsymbol{U}}_2^\dagger \boldsymbol{U}_2 = \boldsymbol{M}_3^\top \boldsymbol{B} \boldsymbol{\Phi} \times \left(\boldsymbol{M}_2^\top \boldsymbol{B} \boldsymbol{\Phi}\right)^\dagger \times \boldsymbol{M}_2^\top \boldsymbol{B} \boldsymbol{\Theta} = \boldsymbol{M}_3^\top \boldsymbol{B} \boldsymbol{\Theta}.$$

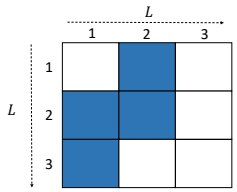

Fig. 2. An illustrative case for the subspace identifiability analysis.

This estimated $\boldsymbol{U}_3$ can be combined with $\boldsymbol{U}_1$ and $\boldsymbol{U}_2$ to attain (6). To handle the more general $L > 3$ case, the "subspace stitching" idea conveyed by this simple example is recursively applied in Algorithm 1 over the queried blocks $\boldsymbol{A}_{\ell_r,m_r}$ and $\boldsymbol{A}_{\ell_{r+1},m_r}$ for $r = 1,\ldots,L-1$. Note that we start the iterations from $r = \lfloor L/2 \rfloor$ and perform the subspace stitching of the blocks in ascending and descending orders of r, respectively, because this helps reduce the overall subspace estimation error when noise is present (critical in the binary observation case in the next subsection). In terms of subspace identifiability, we have the following proposition:

Proposition 1. *(Ideal Case) Assume that* $\boldsymbol{A}_{\ell,m} = \boldsymbol{P}_{\ell,m} = \boldsymbol{M}_\ell^\top \boldsymbol{B} \boldsymbol{M}_m \in \mathbb{R}^{|\mathcal{S}_\ell| \times |\mathcal{S}_m|}$ *holds true for all* $\ell, m \in [L]$ *and* $\mathrm{rank}(\boldsymbol{M}) = \mathrm{rank}(\boldsymbol{B}) = K$. *Suppose that the* $\boldsymbol{A}_{\ell,m}$*'s are queried according to the proposed EQP. Then, the output* $\widehat{\boldsymbol{U}}$ *by Algorithm 1 satisfies* $\mathsf{range}(\widehat{\boldsymbol{U}}) = \mathsf{range}(\boldsymbol{M}^\top)$.

Once $\boldsymbol{U}$ is estimated, the second stage boils down to estimating $\boldsymbol{M}$ from the following model:

$$\boldsymbol{U}^\top = \boldsymbol{G}\boldsymbol{M},\ \boldsymbol{M} \geq \boldsymbol{0},\ \boldsymbol{1}^\top \boldsymbol{M} = \boldsymbol{1}^\top, \tag{7}$$

where $\boldsymbol{G} \in \mathbb{R}^{K \times K}$ is nonsingular. Learning $\boldsymbol{M}$ from the model (7) is the so-called *simplex-structured matrix factorization* (SSMF) problem [22–25]. Algorithm 1 employs a Gram-Schmidt-like scalable algorithm known as successive projection algorithm (SPA) [23] for this task. From the model in (7), SPA can provably identify $\boldsymbol{M}$ in K steps, if $\boldsymbol{G}$ is nonsingular and if there exists

$$\boldsymbol{\Lambda} = \{n_1, \ldots, n_K\} \tag{8}$$

such that $\boldsymbol{M}(:, n_k) = \boldsymbol{e}_k$, where $\boldsymbol{e}_k \in \mathbb{R}^K$ is the kth unit vector. The existence of $\boldsymbol{\Lambda}$ translates to the existence of the so-called *pure nodes* (i.e., nodes belong to a single cluster) [18, 19]. This assumption is considered reasonable when the graph is large.

3.3. Performance Characterization under Binary Observations

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

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

$$\|\widehat{\boldsymbol{U}} - \boldsymbol{U}\boldsymbol{O}\|_{\mathrm{F}} = O\left(\frac{K^{L/4}\kappa(\boldsymbol{B})\sqrt{\rho}}{\sigma_{\min}(\boldsymbol{B})\sqrt{N/L}}\right), \tag{9}$$

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

Algorithm 1: `Proposed Algorithm`

input : $\{\boldsymbol{A}_{m,\ell}\}$, L, K

divide the blocks as $\{\boldsymbol{A}_{\ell_r,m_r}\}_{r=1}^L$, $\{\boldsymbol{A}_{\ell_{r+1},m_r}\}_{r=1}^{L-1}$ (where $\ell_r \neq \ell_{r+1}$, $\{\ell_r\}_{r=1}^L = [L]$, $m_r \in [L]$;

$T \leftarrow \lfloor L/2 \rfloor$;

$\boldsymbol{C}_T \leftarrow [\boldsymbol{A}_{\ell_T,m_T}^\top,\ \boldsymbol{A}_{\ell_{T+1},m_T}^\top]^\top$;

$[\boldsymbol{U}_{\ell_T}^\top, \boldsymbol{U}_{\ell_{T+1}}^\top]^\top \boldsymbol{\Sigma}\boldsymbol{V}^\top \leftarrow \mathrm{svd}_K(\boldsymbol{C}_T)$;

$\boldsymbol{U}_{\mathrm{ref}} \leftarrow \boldsymbol{U}_{\ell_{T+1}}$;

for $r = T+1 : 1 : L-1$ **do**

  $\boldsymbol{C}_r \leftarrow [\boldsymbol{A}_{\ell_r,m_r}^\top,\ \boldsymbol{A}_{\ell_{r+1},m_r}^\top]^\top$;

  $[\widetilde{\boldsymbol{U}}_{\ell_r}^\top, \widetilde{\boldsymbol{U}}_{\ell_{r+1}}^\top]^\top \boldsymbol{\Sigma}_r \boldsymbol{V}_{m_r}{}^\top \leftarrow \mathrm{svd}_K(\boldsymbol{C}_r)$;

  $\boldsymbol{U}_{\ell_{r+1}} \leftarrow \widetilde{\boldsymbol{U}}_{\ell_{r+1}} \widetilde{\boldsymbol{U}}_{\ell_r}^\dagger \boldsymbol{U}_{\mathrm{ref}}$;

  $\boldsymbol{U}_{\mathrm{ref}} \leftarrow \boldsymbol{U}_{\ell_{r+1}}$;

end

$\boldsymbol{U}_{\mathrm{ref}} \leftarrow \boldsymbol{U}_{\ell_T}$;

for $r = T : -1 : 2$ **do**

  $\boldsymbol{C}_r \leftarrow [\boldsymbol{A}_{\ell_r,m_r}^\top,\ \boldsymbol{A}_{\ell_{r-1},m_r}^\top]^\top$;

  $[\widetilde{\boldsymbol{U}}_{\ell_r}^\top, \widetilde{\boldsymbol{U}}_{\ell_{r-1}}^\top]^\top \boldsymbol{\Sigma}_r \boldsymbol{V}_{m_r}{}^\top \leftarrow \mathrm{svd}_K(\boldsymbol{C}_r)$;

  $\boldsymbol{U}_{\ell_{r-1}} \leftarrow \widetilde{\boldsymbol{U}}_{\ell_{r-1}} \widetilde{\boldsymbol{U}}_{\ell_r}^\dagger \boldsymbol{U}_{\mathrm{ref}}$;

  $\boldsymbol{U}_{\mathrm{ref}} \leftarrow \boldsymbol{U}_{\ell_{r-1}}$;

end

$\widehat{\boldsymbol{U}} \leftarrow [\boldsymbol{U}_1^\top, \ldots, \boldsymbol{U}_L^\top]^\top$;

apply SPA on $\widehat{\boldsymbol{U}}$ to estimate $\widehat{\boldsymbol{M}}$.

output: Estimated membership matrix $\widehat{\boldsymbol{M}}$.

The proof can be found in a longer version[1]. The key idea behind the proof is the fact that the principal components of the binary adjacency sub-graphs return the target range space up to bounded errors [26]. Leveraging this result, and combining with the proposed recursive "subspace stiching" technique (with extensive care to manage error propagation among iterations), one can show (9). 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 $K \leq |\mathcal{S}_\ell|$ will be violated. In addition, larger L also makes the error bound looser. On the other hand, larger L means that only fewer queries need to be made, and thus less resource consuming.

Remark 1. *Note that the estimated* $\widehat{\boldsymbol{U}}$ *can be represented as* $\widehat{\boldsymbol{U}}^\top = \boldsymbol{G}\boldsymbol{M} + \boldsymbol{N}$ *where* $\boldsymbol{M}$ *satisfies* (1), $\boldsymbol{G} \in \mathbb{R}^{K \times K}$ *is nonsingular and* $\boldsymbol{N}$ *represents the noise which is shown to be bounded by Proposition 2. In order to extract* $\boldsymbol{M}$ *from the estimated* $\widehat{\boldsymbol{U}}$, *Algorithm 1 employs SPA which is provably robust to bounded noise [23]. Hence, leveraging Proposition 2 and the noise robustness of SPA, one can show*

[1] http://people.oregonstate.edu/ ibrahish/graphclusteringicassp2021.pdf

Table 1. The subspace distance between $\widehat{\boldsymbol{U}}$ and $\boldsymbol{U}$ and MSE of $\boldsymbol{M}$ for ideal case and binary observation case.

Graph Size	Ideal Case	Binary Observation Case			
	Proposed	Proposed		GeoNMF	CD-MVS
N	Dist	Dist	MSE	MSE	MSE
1×10^4	7.34×10^{-13}	0.342	**0.0475**	0.0554	0.0839
2×10^4	2.80×10^{-13}	0.209	**0.0198**	0.0386	0.0943
4×10^4	1.22×10^{-13}	0.194	**0.0123**	0.0341	0.0955
8×10^4	1.12×10^{-13}	0.101	**0.0066**	0.0261	0.0924

that the estimated $\widehat{\boldsymbol{M}}$ by Algorithm 1 is not far away from $\boldsymbol{M}\boldsymbol{\Pi}$ for a certain permutation matrix $\boldsymbol{\Pi}$.

Remark 2. *Under the EQP, another solution for estimating $\boldsymbol{M}$ is to apply existing mixed membership learning algorithms, e.g., those in [17–19] on the small blocks $\boldsymbol{A}_{\ell,m}$ and individually learn the corresponding parts of $\boldsymbol{M}$. Then, the entire $\boldsymbol{M}$ can be recovered by unifying the intrinsic column permutation ambiguity between blocks of $\boldsymbol{M}$. This is doable, but may have relatively poor identifiability guarantees. The reason is that learning $\boldsymbol{M}_\ell$ from $\boldsymbol{A}_{\ell,m}$ via the methods in [17–19] requires that the convex hull of $\boldsymbol{M}_\ell^\top$ to be well spread in the probability simplex (e.g., the existence of pure nodes [18, 19] implies $\mathrm{conv}\{\boldsymbol{M}^\top\} = \{\boldsymbol{x} \in \mathbb{R}^K | \mathbf{1}^T\boldsymbol{x} = 1, \boldsymbol{x} \geq 0\}$)—see detailed discussion in [27]. When the methods are applied on small subblocks $\boldsymbol{A}_{\ell,m}$, this assumption is less likely to hold by the corresponding submatrix $\boldsymbol{M}_\ell$ or $\boldsymbol{M}_m$ [28]. Hence, this seemingly natural approach is less preferable—as one will see in the next section.*

Remark 3. *Algorithm 1 only consists of top-K truncated SVD on small blocks which has a complexity of $O((N/L)K^2)$ flops, assuming $|\mathcal{S}_\ell| = N/L$ for all ℓ 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 $K \ll N$ often holds. Instead, the convex optimization approaches for single membership learning in [12–14] use $O(N^2)$ memory to instantiate the optimization variable, which is heavily memory consuming.*

4. EXPERIMENTS AND CONCLUSION

Baselines. We employ two state-of-the-art mixed membership learning algorithms, namely, GeoNMF [19] and CD-MVS [17] 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 queried adjacency matrix. We use a procedure that applies these baselines to each block and aligns the estimated block membership matrices, i.e., $\boldsymbol{M}_\ell$'s, as mentioned in Remark 2. The details of this alignment procedure can be found in the longer version of this submission.

Synthetic Data. We consider N nodes (where $N \in [1 \times 10^4, 8 \times 10^4]$) and $K = 5$ clusters. The membership vectors $\boldsymbol{m}_n$ are drawn from the Dirichlet distribution with parameters being $(1/K)\mathbf{1}$, where $\mathbf{1}$ is a K-dimensional all-one vector. The entries of matrix $\boldsymbol{B}$ are drawn from 0 to 1 uniformly at random. We first test the identifiability claims under the ideal case (i..e, $\boldsymbol{A} = \boldsymbol{P}$). The blocks of the adjacency matrix with the leftmost query pattern in Fig. 1 is used. We fix the number of groups $L = 10$. The results are averaged over 20 random trials. Table 1 shows the averaged subspace estimation accuracy of our method measured using the subspace distance measure (denoted as Dist) (see definition in [29]) under different N's. One can see that the proposed method estimates the subspace of the membership matrix $\boldsymbol{M}$ very accurately, which verifies our subspace identifiability analysis in Proposition 1.

Table 2. Averaged SRC and runtime in seconds for MAG1 ($N = 37680, K = 3$) and MAG2 ($N = 19457, K = 3$) fixing $L = 10$.

Datasets	Proposed		GeoNMF		CDMVS	
	SRC	Time(s.)	SRC	Time(s.)	SRC	Time(s.)
MAG1	**0.125**	**0.26**	0.122	1.79	0.089	0.59
MAG2	**0.441**	**0.23**	0.240	4.66	0.249	0.53

Table 3. Clustering accuracy (%) of MAG2. $N = 19457$, $K = 3$.

Alorithms	$L = 10$	$L = 25$	$L = 50$	$L = 75$	$L = 100$
Proposed	**78.70**	**77.19**	**67.81**	**61.85**	**56.98**
GeoNMF	58.16	57.87	56.88	52.68	52.33
CDMVS	53.45	21.82	14.57	13.53	11.71
SC-Norm	64.80	67.29	59.80	52.70	55.90

Next, we consider the rightmost pattern in Fig 1 to evaluate the proposed algorithm and the baselines in the binary observation case (i.e., $\boldsymbol{A}(i,j) \sim \mathsf{Bernoulli}(\boldsymbol{P}(i,j))$). The results can also be found in Table 1, which shows the subspace distance and mean squared error (MSE) of the estimated $\boldsymbol{M}$ (see definition in [25]) averaged over 20 random trials. 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 $\boldsymbol{M}$.

Real Data. We test the algorithms using the co-authorship network data from *Microsoft Academic Graph* (MAG) [30]. We use MAG1 and MAG2 versions from [19]. The networks are provided with the ground-truth mixed membership of the nodes. In MAG, the nodes represent the authors of research papers published in different fields (clusters). Some authors publish in more than one fields and thus have mixed membership. From the original dataset, we select nodes that admit a degree at least 5. Consequently, the MAG1 and MAG2 networks under test have 37,680 and 19,457 authors, respectively. All the authors are from 3 different fields ($K = 3$). We let all the algorithms access only part of the network under the diagonal query pattern in Fig. 1. We randomize the node order in each of the 20 trials and present the averaged result.

In Table 2, the proposed method and the baselines are evaluated using the averaged *Spearman's rank correlation* coefficient (SRC) (see definition in [19]) for $L = 10$. The SRC takes values between -1 and 1; SRC has a higher value if the ranking of the entries in two vectors are more similar—which is desired. From Table 2, it can be observed that the proposed algorithm outperforms the baselines for both datasets. In addition, the runtime of the proposed algorithm is appealing, considering the large scale of MAG1 and MAG2.

Table 3 shows the clustering accuracy (see definition in [24] and reference therein) of the algorithms under different L's on MAG2. The clustering accuracy is measured via applying k-means to the learned membership vectors. One can see that the performance decreases when L increases—which is consistent with our analysis in Proposition 2. Notably, when $L = 50$, i.e, only 5.25% of $\boldsymbol{A}$ is observed—but the proposed method still outputs a reasonable clustering accuracy, which demonstrates a promising balance between query sample complexity and clustering accuracy.

To conclude, we proposed a graph query scheme that enables provable graph clustering with partially observed binary edges. Unlike previous works which rely 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 with provable

guarantees, while existing graph query methods do not offer performance characterizations beyond the single membership case. Our method was tested on real co-author networks and exhibited promising performance.

5. REFERENCES

[1] Satu Elisa Schaeffer, "Graph clustering," *Computer Science Review*, vol. 1, no. 1, pp. 27 – 64, 2007.

[2] Mark Handcock, Adrian Raftery, and Jeremy Tantrum, "Model-based clustering for social networks," *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, vol. 170, pp. 301 – 354, 03 2007.

[3] Dietmar Cordes, Vic Haughton, John D Carew, Konstantinos Arfanakis, and Ken Maravilla, "Hierarchical clustering to measure connectivity in fmri resting-state data," *Magnetic resonance imaging*, vol. 20, no. 4, pp. 305—317, May 2002.

[4] Georgios Pavlopoulos, Maria Secrier, Charalampos Moschopoulos, Theodoros Soldatos, Sophia Kossida, Jan Aerts, Reinhard Schneider, and Pantelis Bagos, "Using graph theory to analyze biological networks," *BioData mining*, vol. 4, pp. 10, 04 2011.

[5] Andrew Y Ng, Michael I Jordan, and Yair Weiss, "On spectral clustering: Analysis and an algorithm," in *Advances in Neural Information Processing Systems*, 2002, pp. 849–856.

[6] Ulrike Luxburg, "A tutorial on spectral clustering," *Statistics and Computing*, vol. 17, no. 4, pp. 395–416, Dec. 2007.

[7] Jianbo Shi and J. Malik, "Normalized cuts and image segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 22, no. 8, pp. 888–905, 2000.

[8] A. Saade, M. Lelarge, F. Krzakala, and L. Zdeborová, "Clustering from sparse pairwise measurements," in *2016 IEEE International Symposium on Information Theory (ISIT)*, 2016, pp. 780–784.

[9] Hao Yin, Austin R. Benson, Jure Leskovec, and David F. Gleich, "Local higher-order graph clustering," New York, NY, USA, 2017, KDD '17, p. 555–564, Association for Computing Machinery.

[10] Matthew J. Salganik and Douglas D. Heckathorn, "Sampling and estimation in hidden populations using respondent-driven sampling," *Sociological Methodology*, vol. 34, no. 1, pp. 193–240, 2004.

[11] Jure Leskovec and Christos Faloutsos, "Sampling from large graphs," in *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, New York, NY, USA, 2006, p. 631–636, Association for Computing Machinery.

[12] Ramya Korlakai Vinayak, Samet Oymak, and Babak Hassibi, "Graph clustering with missing data: Convex algorithms and analysis," in *Advances in Neural Information Processing Systems 27*, 2014, pp. 2996–3004.

[13] Ramya Korlakai Vinayak and Babak Hassibi, "Crowdsourced clustering: Querying edges vs triangles," in *Advances in Neural Information Processing Systems 29*, pp. 1316–1324. 2016.

[14] Yudong Chen, Ali Jalali, Sujay Sanghavi, and Huan Xu, "Clustering partially observed graphs via convex optimization," *The Journal of Machine Learning Research*, vol. 15, no. 1, pp. 2213–2238, 2014.

[15] C.E. Särndal, B. Swensson, and J.H. Wretman, *Model Assisted Survey Sampling*, Springer series in statistics. Springer-Verlag, 1992.

[16] Jierui Xie, Stephen Kelley, and Boleslaw K. Szymanski, "Overlapping community detection in networks: The state-of-the-art and comparative study," *ACM Comput. Surv.*, vol. 45, no. 4, Aug. 2013.

[17] Kejun Huang and Xiao Fu, "Detecting overlapping and correlated communities without pure nodes: Identifiability and algorithm," in *Proceedings of the 36th International Conference on Machine Learning*, 2019, pp. 2859–2868.

[18] Maxim Panov, Konstantin Slavnov, and Roman Ushakov, "Consistent estimation of mixed memberships with successive projections," *Complex Networks Their Applications VI*, p. 53–64, Nov 2017.

[19] Xueyu Mao, Purnamrita Sarkar, and Deepayan Chakrabarti, "On mixed memberships and symmetric nonnegative matrix factorizations," in *International Conference on Machine Learning*, 2017, pp. 2324–2333.

[20] Tom Snijders and Krzysztof Nowicki, "Estimation and prediction for stochastic blockmodels for graphs with latent block structure," *Journal of Classification*, vol. 14, pp. 75–100, 01 1997.

[21] Edoardo M Airoldi, David M Blei, Stephen E Fienberg, and Eric P Xing, "Mixed membership stochastic blockmodels," *Journal of Machine Learning Research*, vol. 9, no. Sep, pp. 1981–2014, 2008.

[22] J. M. Bioucas-Dias, "A variable splitting augmented lagrangian approach to linear spectral unmixing," in *Proc. IEEE WHISPERS'09*, 2009, pp. 1–4.

[23] N. Gillis and S.A. Vavasis, "Fast and robust recursive algorithms for separable nonnegative matrix factorization," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 36, no. 4, pp. 698–714, April 2014.

[24] Xiao Fu, Kejun Huang, Bo Yang, Wing-Kin Ma, and Nicholas D Sidiropoulos, "Robust volume minimization-based matrix factorization for remote sensing and document clustering," *IEEE Trans. Signal Process.*, vol. 64, no. 23, pp. 6254–6268, 2016.

[25] X. Fu, W.-K. Ma, K. Huang, and N. D. Sidiropoulos, "Blind separation of quasi-stationary sources: Exploiting convex geometry in covariance domain," *IEEE Trans. Signal Process.*, vol. 63, no. 9, pp. 2306–2320, May 2015.

[26] Jing Lei and Alessandro Rinaldo, "Consistency of spectral clustering in stochastic block models," *Ann. Statist.*, vol. 43, no. 1, pp. 215–237, 02 2015.

[27] X. Fu, K. Huang, N. D. Sidiropoulos, and W.-K. Ma, "Nonnegative matrix factorization for signal and data analytics: Identifiability, algorithms, and applications," *IEEE Signal Process. Mag.*, vol. 36, no. 2, pp. 59–80, 2019.

[28] Shahana Ibrahim, Xiao Fu, Nikos Kargas, and Kejun Huang, "Crowdsourcing via pairwise co-occurrences: Identifiability and algorithms," in *Advances in Neural Information Processing Systems 32*, 2019, pp. 7847–7857.

[29] Gene H Golub and Charles F Van Loan, *Matrix computations*, vol. 3, JHU Press, 2012.

[30] A. Sinha, Zhihong Shen, Yang Song, Hao Ma, Darrin Eide, B. Hsu, and K. Wang, "An overview of microsoft academic service (mas) and applications," in *WWW '15 Companion*, 2015.