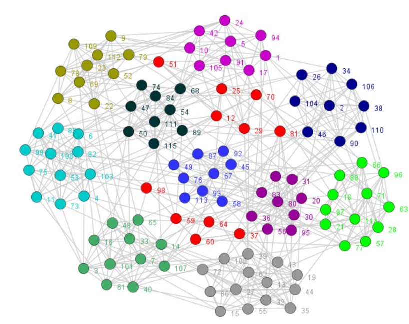
Learning Mixed Membership from Adjacency Graph via Systematic Edge Query: Identifiability and Algorithm [Poster ID : 3972]

Shahana Ibrahim, Xiao Fu

School of EECS, Oregon State University, Corvallis, OR, USA

Graph Clustering (GC)

- **Graph Clustering (GC)** is a core analysis technique used for network data:
  - Social Networks, Ecological Networks, Transportation Networks, Brain Networks etc.
- Real networks are often available with **partial observation of its edges** due to:
  - Massive Data, Cost, Security/Privacy



Existing Work with Provable Guarantees

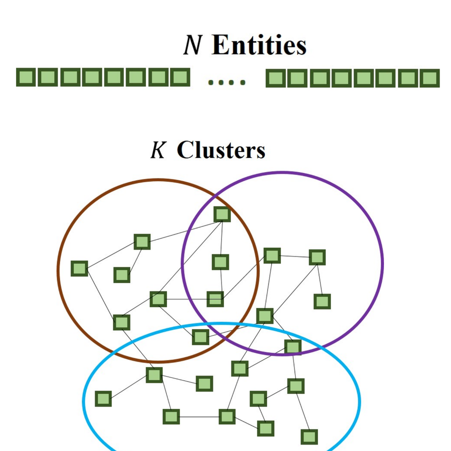
A number of works [Korlakai Vinayak et al., 2014; Korlakai Vinayak and Hassibi, 2016; Chen et al., 2014], which proposed GC under partial edge observation, features

- ☐ **single membership identification**
  - the entities often admit mixed membership in real-world networks
- ☐ **random query based edge acquisition scheme**
  - may not be easy to implement in some applications; e.g., in field surveys and in networks with hidden or intentionally removed edges
- ☐ **convex optimization based problem formulation**
  - hard to scale up for real-world large graphs

We aim to design a **systematic edge query scheme** for **mixed membership identification** via a **lightweight algorithm** with **provable guarantees.**

Mixed Membership Model

- The nth entity belongs to kth cluster with prob. m_{kn}
  - $\Sigma_{k=1}^{K} m_{k,n} = 1,\ m_{k,n} \geq 0.$
- $\boldsymbol{m}_n = [m_{1,n}, \dots, m_{K,n}]^\top$ is called as the **membership vector** of n.
- $\boldsymbol{M} = [\boldsymbol{m}_1, \dots, \boldsymbol{m}_N] \in \mathbb{R}^{K\times N}$ is called as the **membership matrix**.
- $\boldsymbol{B} \in \mathbb{R}^{K\times K}$ is **cluster-cluster interaction matrix**.

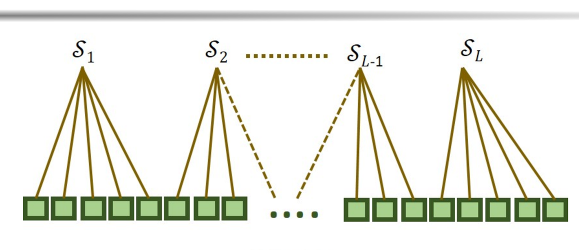


- The edges of the graph are represented using **adjacency matrix** $\boldsymbol{A} \in \{0,1\}^{N\times N}$:

$$\boldsymbol{A}(i,j) \sim \text{Bernoulli}\left(\boldsymbol{P}(i,j)\right), \quad \boldsymbol{P} = \boldsymbol{M}^\top \boldsymbol{B}\boldsymbol{M}, \quad \boldsymbol{1}^\top\boldsymbol{M} = \boldsymbol{1}^\top, \quad \boldsymbol{M} \geq \boldsymbol{0}.$$

Proposed Systematic Edge Query

$$\mathcal{S}_1 \cup \dots \cup \mathcal{S}_L = \{1,\dots,N\}$$
$$\mathcal{S}_\ell \cap \mathcal{S}_m = \emptyset,\ \forall \ell \neq m$$



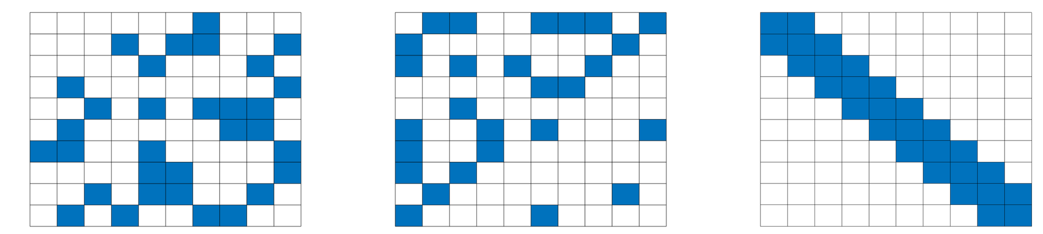
Adjacency Submatrix between $\mathcal{S}_\ell$ **and** $\mathcal{S}_m \implies \boldsymbol{A}_{\ell,m} \in \mathbb{R}^{|\mathcal{S}_\ell|\times|\mathcal{S}_m|}$

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.

Algorithm Design for Learning M under EQP

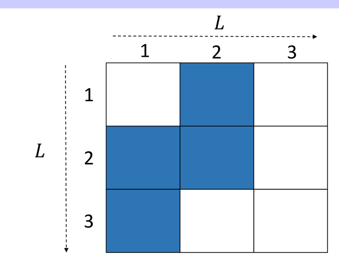
Some examples for EQP patterns with $N = 1000$, $K = 5$ and $L = 10$.



Step 1: Estimate $\boldsymbol{U} \in \mathbb{R}^{N\times K}$ such that range$(\boldsymbol{U})$ = range$(\boldsymbol{M}^\top)$
Step 2: Estimate $\boldsymbol{M}$ from $\boldsymbol{U}$ via structured matrix factorization (SMF)

Consider $L = 3$ and $\boldsymbol{A}_{\ell,m} = \boldsymbol{P}_{\ell,m} = \boldsymbol{M}_\ell^\top \boldsymbol{B}\boldsymbol{M}_m$:

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



- Define $\boldsymbol{C}_1 := [\boldsymbol{P}_{1,2}^\top, \boldsymbol{P}_{2,2}^\top]^\top$ and $\boldsymbol{C}_2 := [\boldsymbol{P}_{2,1}^\top, \boldsymbol{P}_{3,1}^\top]^\top$. Consider their top-$K$ SVD:

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

- The bases of range$(\boldsymbol{M}_1^\top)$, range$(\boldsymbol{M}_2^\top)$ and range$(\boldsymbol{M}_3^\top)$ are:

$$\boldsymbol{U}_1 = \boldsymbol{M}_1^\top \boldsymbol{B}\boldsymbol{\Theta}, \quad \boldsymbol{U}_2 = \boldsymbol{M}_2^\top \boldsymbol{B}\boldsymbol{\Theta}, \quad \boldsymbol{U}_3 = \boldsymbol{M}_3^\top \boldsymbol{B}\boldsymbol{\Phi}, \quad \boxed{\boldsymbol{\Phi} \neq \boldsymbol{\Theta} \text{ in general.}}$$

- We need a certain $\boldsymbol{U}_3$ such that range$([\boldsymbol{U}_1^\top, \boldsymbol{U}_2^\top, \boldsymbol{U}_3^\top]^\top)$ = range$([\boldsymbol{M}_1, \boldsymbol{M}_2, \boldsymbol{M}_3]^\top)$.
- We can obtain such $\boldsymbol{U}_3$ as below:

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

- To estimate $\boldsymbol{U}$, this "**subspace stitching**" idea is recursively applied over the queried blocks $\boldsymbol{A}_{\ell_r,m_r}$ and $\boldsymbol{A}_{\ell_{r+1},m_r}$ for $r = 1, \dots, L-1$.

$$\boldsymbol{U}^\top = \boldsymbol{G}\boldsymbol{M},\ \boldsymbol{M} \geq \boldsymbol{0},\ \boldsymbol{1}^\top\boldsymbol{M} = \boldsymbol{1}^\top, \quad \boldsymbol{G} \in \mathbb{R}^{K\times K} \text{ is nonsingular.}$$

- *Successive Projection Algorithm* (SPA) [Gillis and Vavasis, 2014] can provably identify $\boldsymbol{M}$ in K steps, if $\boldsymbol{G}$ is nonsingular and if there exists $\{n_1, \dots, n_K\}$ such that $\boldsymbol{M}(:, n_k) = \boldsymbol{e}_k$ (**pure nodes**).

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 $\ell_r \neq \ell_{r+1}$, $\{\ell_r\}_{r=1}^{L} = [L]$, $m_r \in [L]$;
$T \leftarrow \lfloor L/2 \rfloor$;
$\boldsymbol{C}_T \leftarrow [\boldsymbol{A}_{\ell_T,m_T}^\top\ ,\ \boldsymbol{A}_{\ell_{T+1},m_T}^\top]^\top$;
$[\boldsymbol{U}_{\ell_T}^\top, \boldsymbol{U}_{\ell_{T+1}}^\top]^\top \boldsymbol{\Sigma}\boldsymbol{V}^\top \leftarrow \text{svd}_K(\boldsymbol{C}_T)$;
$\boldsymbol{U}_{\text{ref}} \leftarrow \boldsymbol{U}_{\ell_{T+1}}$;
for $r = T+1 : 1 : L-1$ **do**
  $\boldsymbol{C}_r \leftarrow [\boldsymbol{A}_{\ell_r,m_r}^\top\ ,\ \boldsymbol{A}_{\ell_{r+1},m_r}^\top]^\top$;
  $[\widetilde{\boldsymbol{U}}_{\ell_r}^\top, \widetilde{\boldsymbol{U}}_{\ell_{r+1}}^\top]^\top \boldsymbol{\Sigma}_r \boldsymbol{V}_{m_r}{}^\top \leftarrow \text{svd}_K(\boldsymbol{C}_r)$;
  $\boldsymbol{U}_{\ell_{r+1}} \leftarrow \widetilde{\boldsymbol{U}}_{\ell_{r+1}} \widetilde{\boldsymbol{U}}_{\ell_r}^\dagger \boldsymbol{U}_{\text{ref}}$;
  $\boldsymbol{U}_{\text{ref}} \leftarrow \boldsymbol{U}_{\ell_{r+1}}$;
end
$\boldsymbol{U}_{\text{ref}} \leftarrow \boldsymbol{U}_{\ell_T}$;
for $r = T : -1 : 2$ **do**
  $\boldsymbol{C}_r \leftarrow [\boldsymbol{A}_{\ell_r,m_r}^\top\ ,\ \boldsymbol{A}_{\ell_{r-1},m_r}^\top]^\top$;
  $[\widetilde{\boldsymbol{U}}_{\ell_r}^\top, \widetilde{\boldsymbol{U}}_{\ell_{r-1}}^\top]^\top \boldsymbol{\Sigma}_r \boldsymbol{V}_{m_r}{}^\top \leftarrow \text{svd}_K(\boldsymbol{C}_r)$;
  $\boldsymbol{U}_{\ell_{r-1}} \leftarrow \widetilde{\boldsymbol{U}}_{\ell_{r-1}} \widetilde{\boldsymbol{U}}_{\ell_r}^\dagger \boldsymbol{U}_{\text{ref}}$;
  $\boldsymbol{U}_{\text{ref}} \leftarrow \boldsymbol{U}_{\ell_{r-1}}$;
end
$\widehat{\boldsymbol{U}} \leftarrow [\boldsymbol{U}_1^\top, \dots, \boldsymbol{U}_L^\top]^\top$;
apply SPA on $\widehat{\boldsymbol{U}}$ to estimate $\widehat{\boldsymbol{M}}$.
output: Estimated membership matrix $\widehat{\boldsymbol{M}}$.

inputting edges observed via EQP
top-K SVD of the blocks queried for $r = L/2$
subspace stitching for blocks at $r = L/2, \dots, L$
subspace stitching for blocks at $r = 1, \dots, L/2$
Estimation of U
Estimation of M from U

Identifiability Results

Proposition 1: (Subspace Identifiability - 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 $\text{rank}(\boldsymbol{M}) = \text{rank}(\boldsymbol{B}) = K$. Suppose that the $\boldsymbol{A}_{\ell,m}$'s are queried according to the proposed EQP. Then, the output $\boldsymbol{U}$ by Algorithm 1 satisfies range$(\widehat{\boldsymbol{U}})$ = range$(\boldsymbol{M}^\top)$.

Proposition 2: (Subspace Identifiability - Binary Observation Case)

Let $\rho := \max_{i,j} \boldsymbol{P}(i,j)$ be the maximal entry of $\boldsymbol{P}$. Suppose that $\rho = \Omega(L\log(N/L)/N)$ and $L = O(\rho N/d)$ where d is the maximal degree of all the nodes. Also assume that $N = \Omega\left(\max\left(L^2, \frac{(K\gamma^2)^L \rho\kappa^2(\boldsymbol{B})}{\sigma_{\min}^2(\boldsymbol{B})}\right)\right)$. Then, the output $\widehat{\boldsymbol{U}}$ satisfies the following with probability of at least $1 - O(L^2/N)$:

$$\|\widehat{\boldsymbol{U}} - \boldsymbol{U}\boldsymbol{O}\|_{\text{F}} = O\left(\frac{(K\gamma^2)^{L/2}\kappa(\boldsymbol{B})\sqrt{\rho}}{\sigma_{\min}(\boldsymbol{B})\sqrt{N/L}}\right),$$

where $\boldsymbol{O} \in \mathbb{R}^{K\times K}$ is an orthogonal matrix.

Larger L makes the error bound looser, but larger L means that only fewer queries need to be made, and thus less resource consuming.

Experiment Results

- **Synthetic Data Experiment:**
  - Baselines: GeoNMF [Mao et al., 2017], CD-MVSI [Mao et al., 2017]
  - Parameters: $K = 5$, $L = 10$ with diagonal query pattern

Graph Size	Ideal Case ($\boldsymbol{A} = \boldsymbol{P}$)	Binary Observation Case			
Size	Proposed	Proposed		GeoNMF	CD-MVSI
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

- **Real Data Experiment - Microsoft Academic Graph:**
  - MAG1 ($N = 37680, K = 3$); MAG2 ($N = 19457, K = 3$); $L = 10$

Datasets	Proposed		GeoNMF		CD-MVSI	
	Avg. SRC	Time(s.)	Avg. SRC	Time(s.)	Avg. 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

References

- Xueyu Mao, Purnamrita Sarkar, and Deepayan Chakrabarti. On mixed memberships and symmetric nonnegative matrix factorizations. In International Conference on Machine Learning, pages 2324-2333, 2017.
- Kejun Huang and Xiao Fu. Detecting overlapping and correlated communities without pure nodes: Identifiability and algorithm. In International Conference on Machine Learning, pages 2859-2868, 2019.