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

Shahana Ibrahim, Xiao Fu

School of EECS
Oregon State University, Corvallis

IEEE ICASSP 2021
June 6-11, 2021

Graph Clustering (GC)

- *Graph Clustering (GC)* is a core analysis technique frequently applied in various network data:

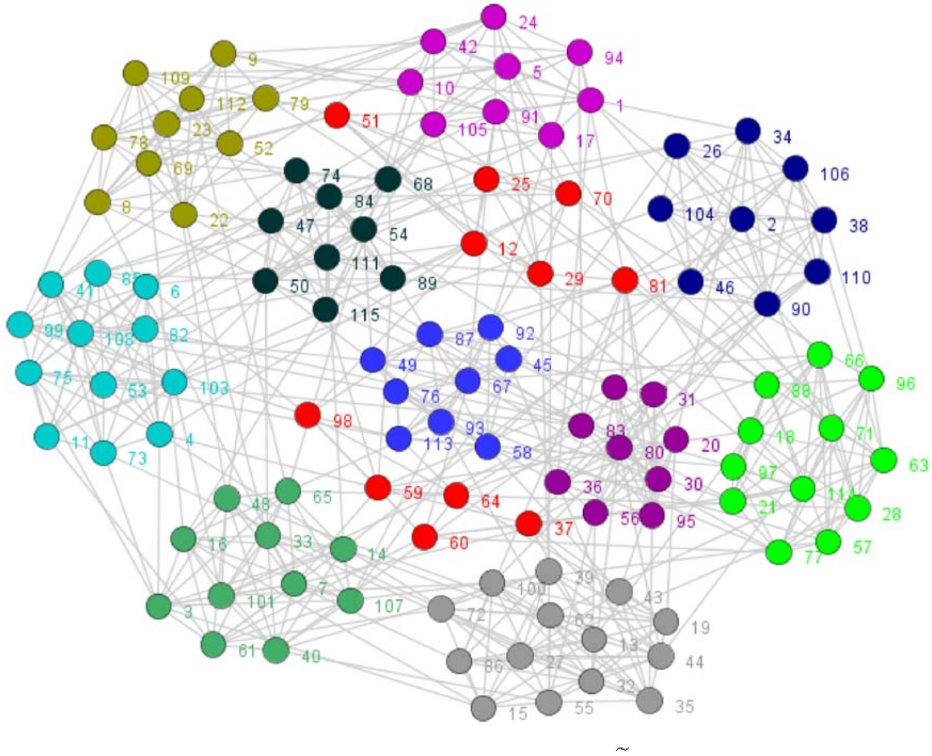
☐ **Social Networks**

☐ **Ecological Networks**

☐ **Transportation Networks**

☐ **Protein-protein Interaction Networks**

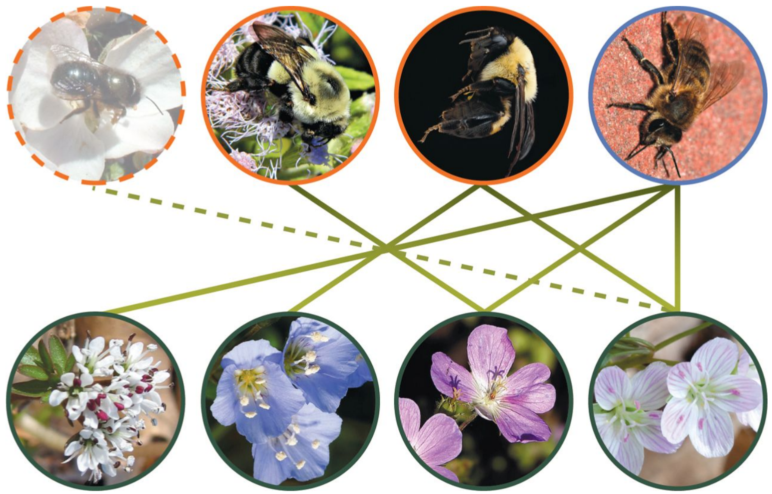
☐ **Brain Networks**



[Source : [Zhang et al., 2007]]

GC under Partial Observation

- Real networks are often available with **partial observation of its edges** due to:
  - **[Massive Data]** e.g., billions of edges in Facebook or Twitter follower-followee network.
  - **[Cost]** e.g., high cost for ecological/biological network data acquisition.
  - **[Security/Privacy]** e.g., intentionally removed or hidden edges in terrorist networks/radical group networks.



[Sources : https://associationsnow.com, https://science.sciencemag.org]

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 with provable guarantees, features

- ☐ **single membership identification**
  - the entities often admit mixed membership in real-world networks

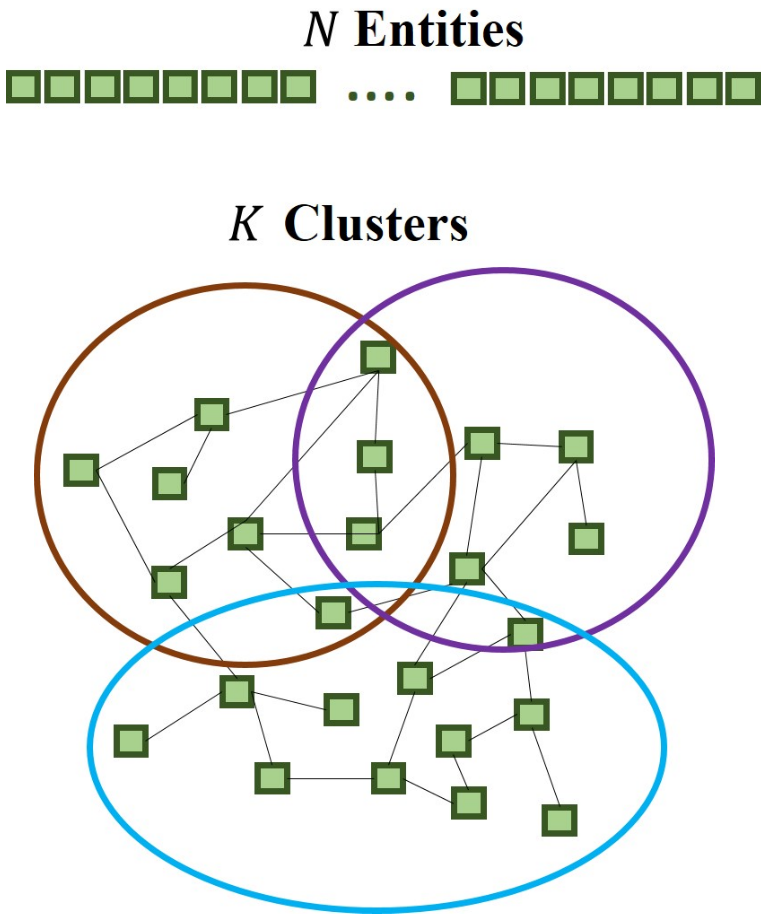
- ☐ **random query based edge acquisition scheme**
  - may not be easy to implement in some applications; e.g., in field surveys and in networks with hidden or intentionally removed edges

- ☐ **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}
  - $\sum_{k=1}^{K} m_{k,n} = 1,\ m_{k,n} \geq 0.$
- $\boldsymbol{m}_n = [m_{1,n}, \ldots, m_{K,n}]^\top$ is called as the **membership vector** of n.
- $\boldsymbol{M} = [\boldsymbol{m}_1, \ldots, \boldsymbol{m}_N] \in \mathbb{R}^{K \times N}$ is called as the **membership matrix**.
- $\boldsymbol{B} \in \mathbb{R}^{K \times K}$ is **cluster-cluster interaction matrix**.
  - $\boldsymbol{B}(p, q)$ denotes the prob. that cluster p connects with cluster q.

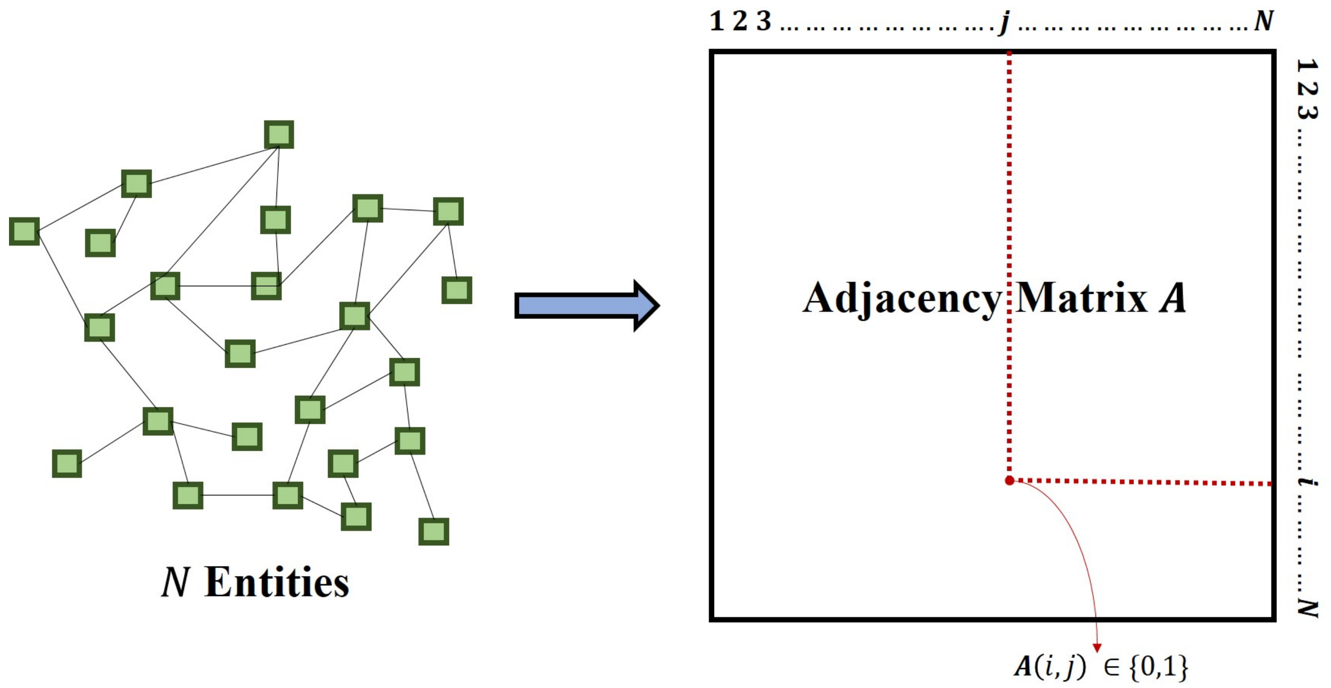


If all $\boldsymbol{m}_n$'s are unit vectors (single cluster membership), it is the so-called the *stochastic block model* (SBM) [Snijders and Nowicki, 1997].

Mixed Membership Model

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

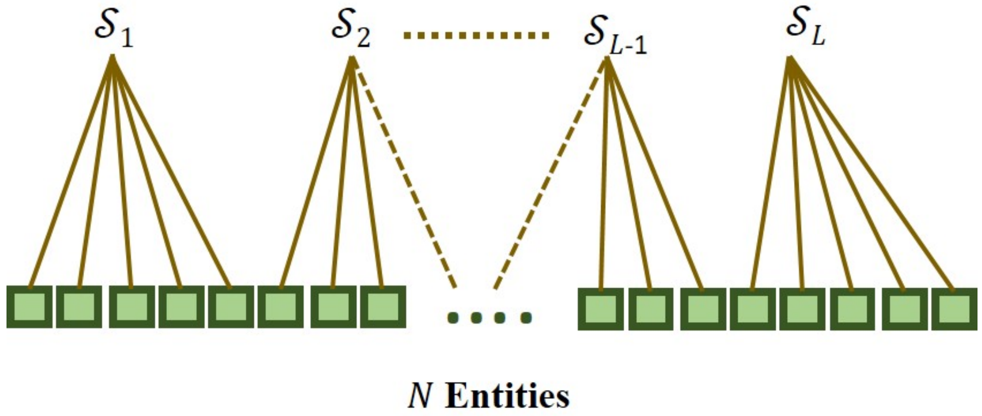
$$\boldsymbol{A}(i,j) \sim \text{Bernoulli}\left(\boldsymbol{P}(i,j)\right), \quad \boldsymbol{P} = \boldsymbol{M}^\top \boldsymbol{B}\boldsymbol{M}, \quad \boldsymbol{1}^\top\boldsymbol{M} = \boldsymbol{1}^\top, \quad \boldsymbol{M} \geq \boldsymbol{0}.$$



- The model is reminiscent of the *mixed membership stochastic block* (MMSB) model in overlapped community detection [Airoldi et al., 2008; Mao et al., 2017].

Proposed Systematic Edge Query

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



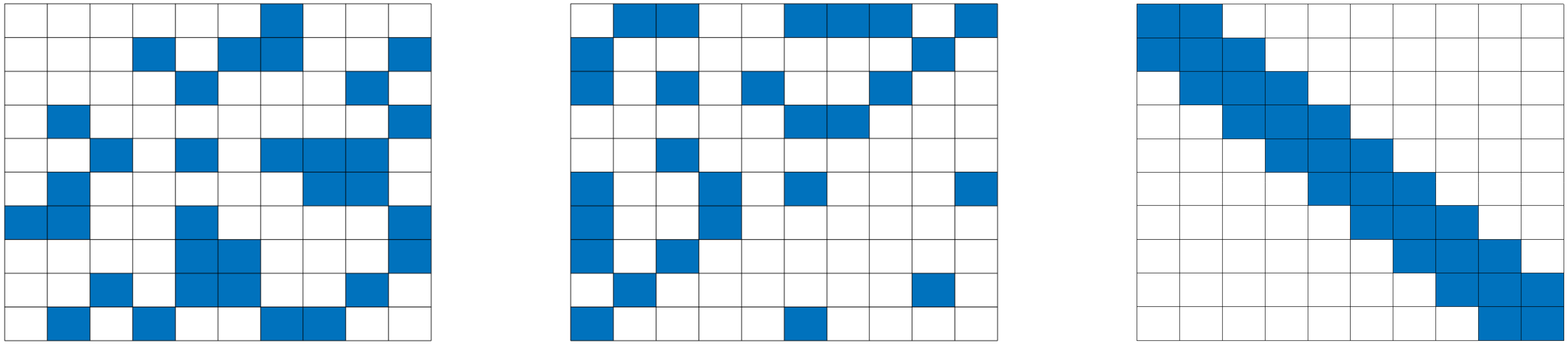
Adjacency Submatrix between $\mathcal{S}_\ell$ and $\mathcal{S}_m \Longrightarrow \boldsymbol{A}_{\ell,m} \in \mathbb{R}^{|\mathcal{S}_\ell| \times |\mathcal{S}_m|}$

Edge Query Principle (EQP)

- For every $\ell \in [L]$, $\boxed{K \leq |\mathcal{S}_\ell|}$ holds. Let $\boxed{m_r \in [L]}$ and $\boxed{\{\ell_r\}_{r=1}^{L} = [L]}$.
- For every ℓ_r, there exists a pair of indices m_r and ℓ_{r+1} where $\boxed{\ell_{r+1} \neq \ell_r}$ such that the edges from the blocks $\boxed{\boldsymbol{A}_{\ell_r,m_r} \text{ and } \boldsymbol{A}_{\ell_{r+1},m_r} \text{ are queried}}$.

EQP Patterns

Some patterns for $\boldsymbol{A}$ following EQP with $N = 1000, K = 5$ and $L = 10$.



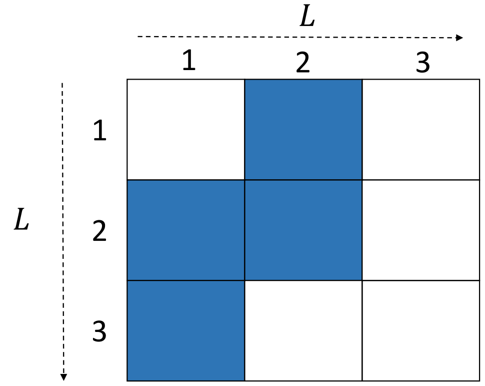
Goal : Learn $\boldsymbol{M}$ by observing $\boldsymbol{A}$ via EQP

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

Subspace Estimation via Block Subspace Stitching

A toy example with $L = 3$ and the ideal case $\boldsymbol{A}_{\ell,m} = \boldsymbol{P}_{\ell,m} = \boldsymbol{M}_\ell^\top \boldsymbol{B} \boldsymbol{M}_m$:

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



- 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:

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

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

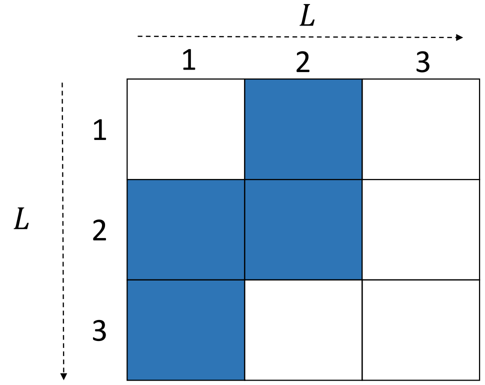
$$\boldsymbol{U}_1 = \boldsymbol{M}_1^\top \boldsymbol{B} \boldsymbol{\Theta}, \quad \boldsymbol{U}_2 = \boldsymbol{M}_2^\top \boldsymbol{B} \boldsymbol{\Theta}, \quad \widetilde{\boldsymbol{U}}_3 = \boldsymbol{M}_3^\top \boldsymbol{B} \boldsymbol{\Phi},$$

$\boldsymbol{\Phi} \neq \boldsymbol{\Theta}$ in general.

Subspace Estimation via Block Subspace Stitching

- Our goal is to get a certain $\boldsymbol{U}_3$ such that the bases can be "stitch"ed to have

$$\text{range}(\underbrace{[\boldsymbol{U}_1^\top, \boldsymbol{U}_2^\top, \boldsymbol{U}_3^\top]^\top}_{\boldsymbol{U}}) = \text{range}(\underbrace{[\boldsymbol{M}_1, \boldsymbol{M}_2, \boldsymbol{M}_3]^\top}_{\boldsymbol{M}^\top}).$$



- We can obtain such $\boldsymbol{U}_3$ as below:

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

- 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 $r = 1, \ldots, L-1$.

Proposed Algorithm

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

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

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 $r = L/2, \ldots, L$

subspace stitching for blocks at $r = 1, \ldots, L/2$

Estimation of U

Estimation of M from U

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 $\mathrm{rank}(\boldsymbol{M}) = \mathrm{rank}(\boldsymbol{B}) = K$. 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

$$\text{range}(\widehat{\boldsymbol{U}}) = \text{range}(\boldsymbol{M}^\top).$$

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

- Algorithm 1 employs successive projection algorithm (SPA) [Gillis and Vavasis, 2014] to identify $\boldsymbol{M}$ from $\boldsymbol{U}$.

- SPA can provably identify $\boldsymbol{M}$ in K steps, if $\boldsymbol{G}$ is nonsingular and if there exists $\{n_1, \ldots, n_K\}$ such that $\boldsymbol{M}(:, n_k) = \boldsymbol{e}_k$ (**pure nodes**).

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 $\widehat{\boldsymbol{U}}$ by Algorithm 1 satisfies the following with probability of at least $1 - O(L^2/N)$:

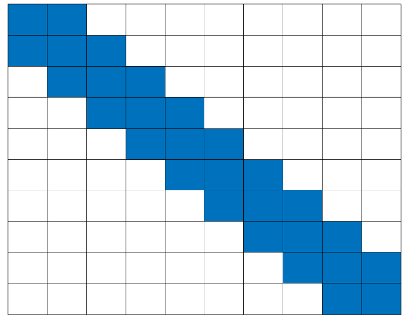
$$\|\widehat{\boldsymbol{U}} - \boldsymbol{U}\boldsymbol{O}\|_{\mathrm{F}} = O\left(\frac{(K\gamma^2)^{L/2}\kappa(\boldsymbol{B})\sqrt{\rho}}{\sigma_{\min}(\boldsymbol{B})\sqrt{N/L}}\right),$$

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

Synthetic Data Experiments

- The membership vectors $\boldsymbol{m}_n$ are drawn from the Dirichlet distribution with parameters being $(1/K)\mathbf{1}$.



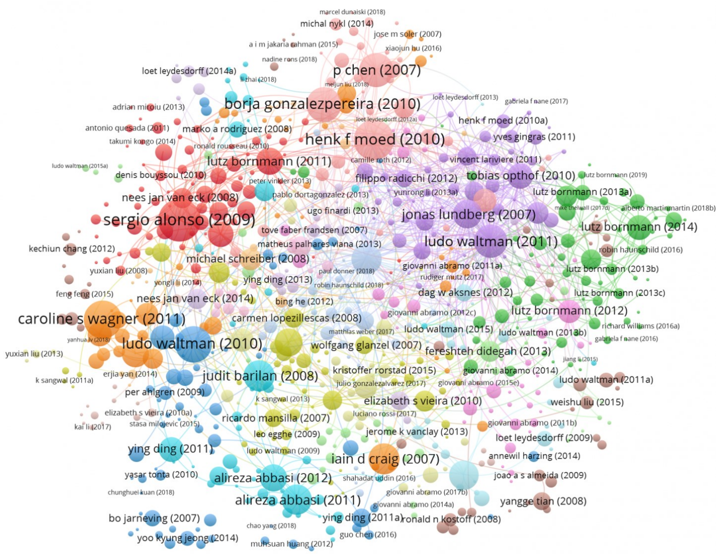
- The entries of matrix $\boldsymbol{B}$ are drawn from $[0, 1]$ uniformly at random and is made diagonally dominant.
- Fixed $L = 10$ and $K = 5$.

- We employ two state-of-the-art mixed membership learning algorithms, namely, GeoNMF [Mao et al., 2017] and CD-MVSI [Huang and Fu, 2019] as baselines

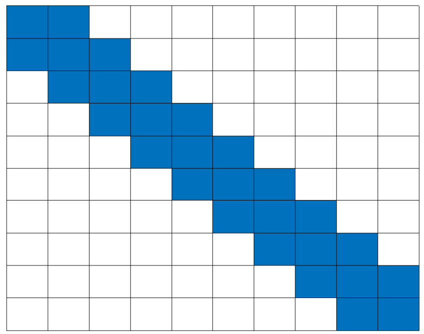
Graph Size	Ideal Case ($\boldsymbol{A} = \boldsymbol{P}$)	Binary Observation Case ($\boldsymbol{A}(i,j) \sim$ Bernoulli ($\boldsymbol{P}(i,j)$))			
	Proposed	Proposed		GeoNMF	CD-MVSI
N	Subspace Distance	Subspace Distance	MSE	MSE	MSE
1×10^4	7.34×10^{-13}	0.342	**0.0475**	0.0554	0.0839
2×10^4	2.80×10^{-13}	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 Experiments - Microsoft Academic Graph (MAG)

- The entities represent the authors of research papers published in 3 different fields.
- The diagonal query pattern is chosen.
- The averaged *Spearman's rank correlation* coefficient (SRC) is used to evaluate the methods:
  - The SRC takes values between -1 and 1.
  - SRC is high if the ranking of the entries in two vectors are similar.



[Illustration of MAG Data, Source : https://www.cwts.nl]



Real Data Experiments

Table 1: Averaged SRC and runtime in seconds for MAG1 ($N = 37680, K = 3$) and MAG2 ($N = 19457, K = 3$) fixing $L = 10$.

Datasets	Proposed		GeoNMF		CD-MVSI	
	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 2: 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
CD-MVSI	53.45	21.82	14.57	13.53	11.71
SC-Norm	64.80	67.29	59.80	52.70	55.90

Summary

- Proposed a **novel framework** that enables **provable graph clustering with partially observed edges**.

- The highlights of the proposed framework are:

  - ☐ **systematic edge query scheme** useful for some important applications
  - ☐ **lightweight algorithm** based on truncated SVD
  - ☐ **mixed membership learning** of the entities with provable guarantees
  - ☐ **promising performance** on synthetic and real data experiments

Thank You!!

*

References

Edoardo M Airoldi, David M Blei, Stephen E Fienberg, and Eric P Xing. Mixed membership stochastic blockmodels. *Journal of Machine Learning Research*, 9 (Sep):1981–2014, 2008.

Yudong Chen, Ali Jalali, Sujay Sanghavi, and Huan Xu. Clustering partially observed graphs via convex optimization. *The Journal of Machine Learning Research*, 15 (1):2213–2238, 2014.

N. Gillis and S.A. Vavasis. Fast and robust recursive algorithms for separable nonnegative matrix factorization. *IEEE Trans. Pattern Anal. Mach. Intell.*, 36(4): 698–714, April 2014.

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*, pages 2859–2868, 2019.

Ramya Korlakai Vinayak and Babak Hassibi. Crowdsourced clustering: Querying edges vs triangles. In *Advances in Neural Information Processing Systems 29*, pages 1316–1324. 2016.

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*, pages 2996–3004, 2014.

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.

Tom Snijders and Krzysztof Nowicki. Estimation and prediction for stochastic blockmodels for graphs with latent block structure. *Journal of Classification*, 14: 75–100, 01 1997.

Shihua Zhang, Rui-Sheng Wang, and X. Zhang. Identification of overlapping community structure in complex networks using fuzzy c-means clustering. *Physica A-statistical Mechanics and Its Applications*, 374:483–490, 2007.