UNIVERSITY OF OXFORD

COMPLEX NETWORKS

Maximum Likelihood Estimation on Stochastic Blockmodels for Directed Graph Clustering

Authors: Mihai Cucuringu[*], Xiaowen Dong[†], and Ning Zhang[*]

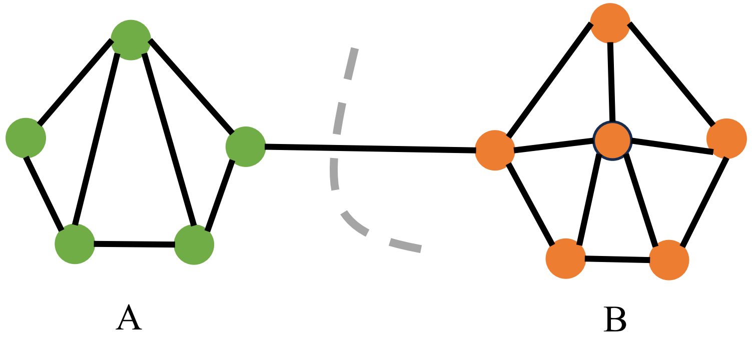
Presenter: Ning Zhang

[*] Department of Statistics, University of Oxford; [†] Department of Engineering Science, University of Oxford

The (undirected) graph clustering problem

▪ Optimization methods

e.g., Ratio Cut [Hagen and Kahng (1992)]

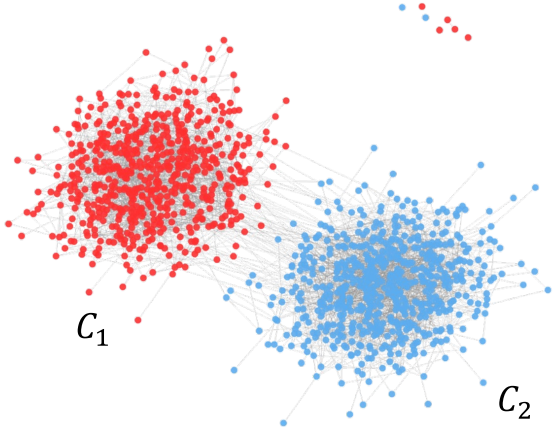


Cut(A,B) = 1

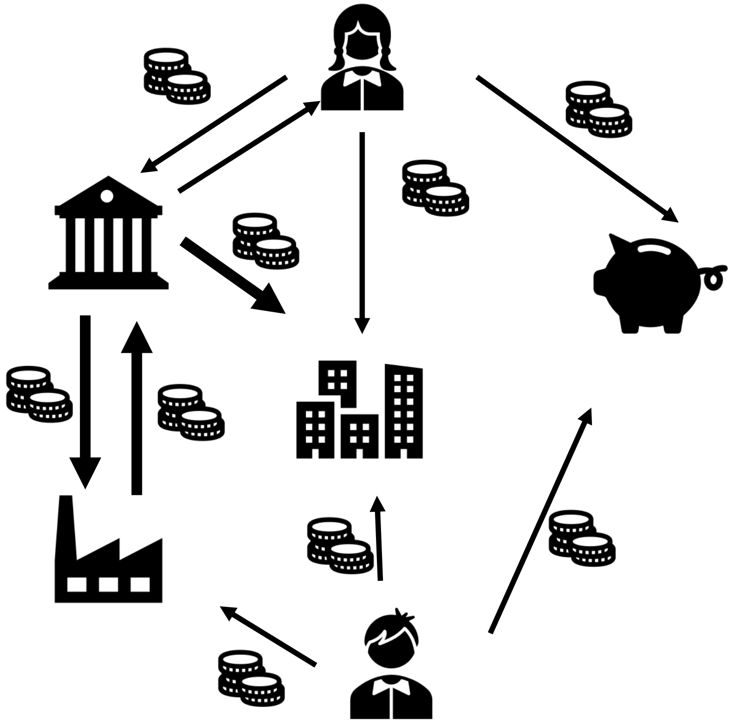
▪Statistical methods

e.g., Community detection in Stochastic Block Models (SBM) [Abbe et al.(2015)]

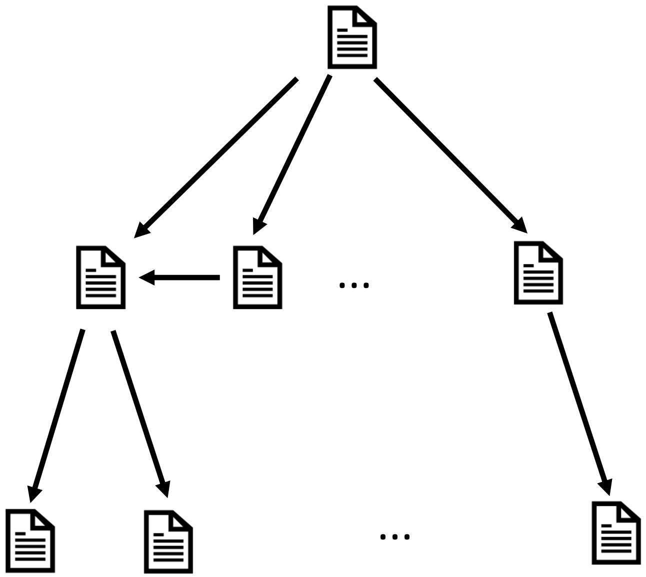
$$\begin{array}{c|cc} & C_1 & C_2 \\ \hline C_1 & p & q \\ C_2 & p & p \end{array} \longrightarrow$$



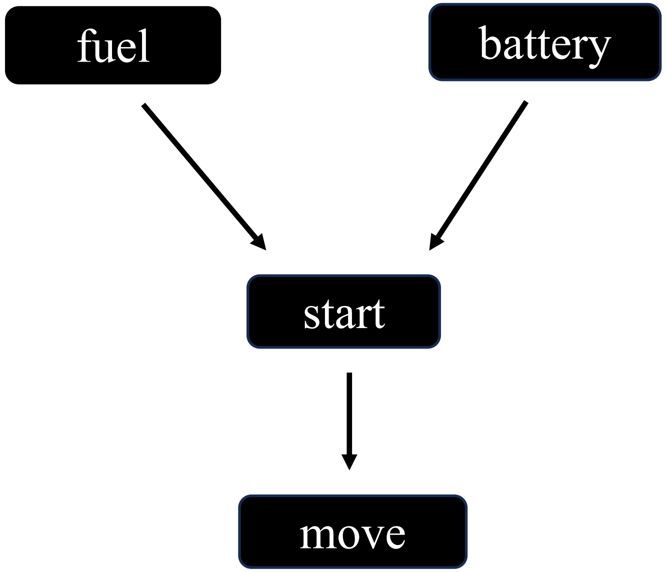
Cluster directed graphs



Financial transition network



Citation network

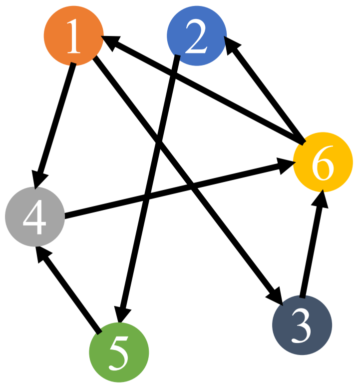


Causal network

Challenges in directed graph clustering

❖ cannot naively apply undirected clustering algorithms

asymmetric edge connection → asymmetric matrix representation



directed graph

0	0	1	1	0	0
0	0	0	0	1	0
0	0	0	0	0	1
0	0	0	0	0	1
0	0	0	1	0	0
1	1	0	0	0	0

graph adjacency matrix A

Existing directed clustering algorithms

- **Symmetrization**

$A + A^T$, $AA^T \& A^T A$ [Kessler (1963), Small (1973), Satuluri and Parthasarathy (2011)]

- **Hermitian**

$\mathrm{H} = i\,(A - A^T)$ [Cucuringu et al. (2020)], magnet Laplacian [Fanuel et al. (2017)]

- **SVD**

SVD on $A - A^T$ [Hayashi et al. (2022)] , DI-SIM [Rohe et al. (2016)]

- **Heuristic**

Weighted Cut optimization [Meilă and Pentney (2007)]

…

Limitations:
lack of theoretical justification on the **clustering objective** or **graph matrix representation**

Existing directed clustering algorithms

- **Symmetrization**

$A + A^T$, $AA^T \& A^T A$ [Kessler (1963), Small (1973), Satuluri and Parthasarathy (2011)]

- **Hermitian**

$\mathrm{H} = i\,(A - A^T)$ [Cucuringu et al. (2020)], magnet Laplacian [Fanuel et al. (2017)]

- **SVD**

SVD on $A - A^T$ [Hayashi et al. (2022)] , DI-SIM [Rohe et al. (2016)]

- **Heuristic**

Weighted Cut optimization [Meilă and Pentney (2007)]

…

Our work:

- propose a novel directed clustering objective
- combined views from **statistics** and **optimization**
- introduce **spectral** and **SDP algorithms** for directed graph clustering

Key idea

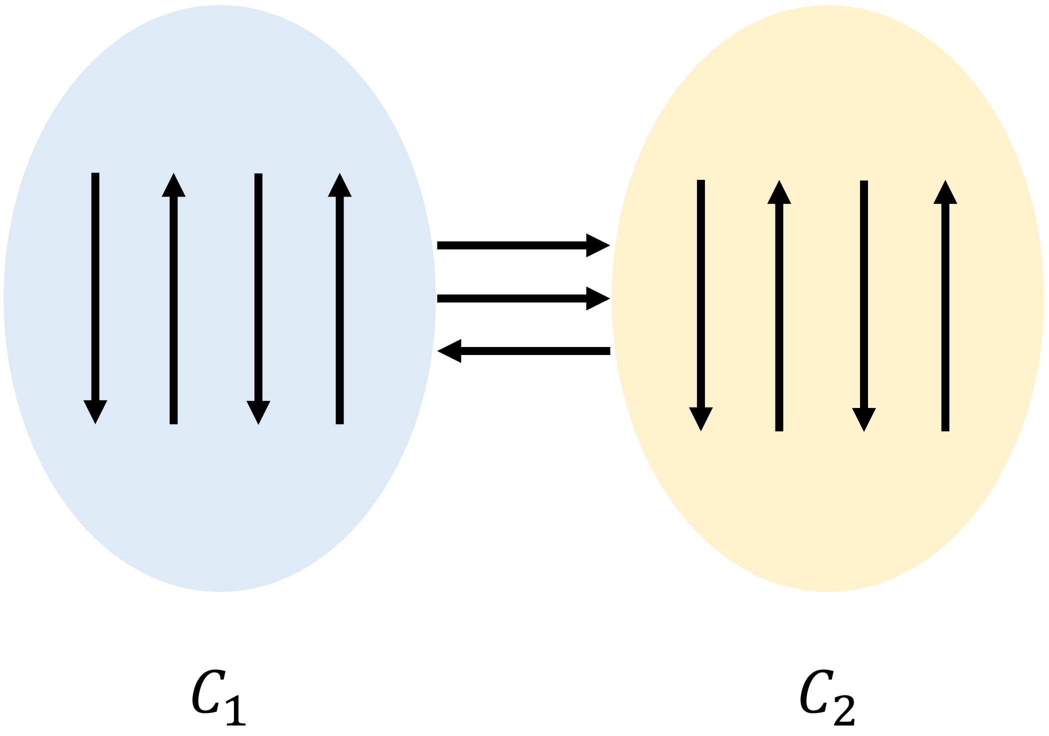
Apply maximum likelihood estimation on Directed-SBM (N, p, q, η)

- For u, v in same cluster:

$$P(u \to v) = p/2$$
$$P(v \to u) = p/2$$

- For $u \in C_1, v \in C_2$

$$P(u \to v) = (1 - \eta)\, q$$
$$P(v \to u) = \eta q$$



Key idea

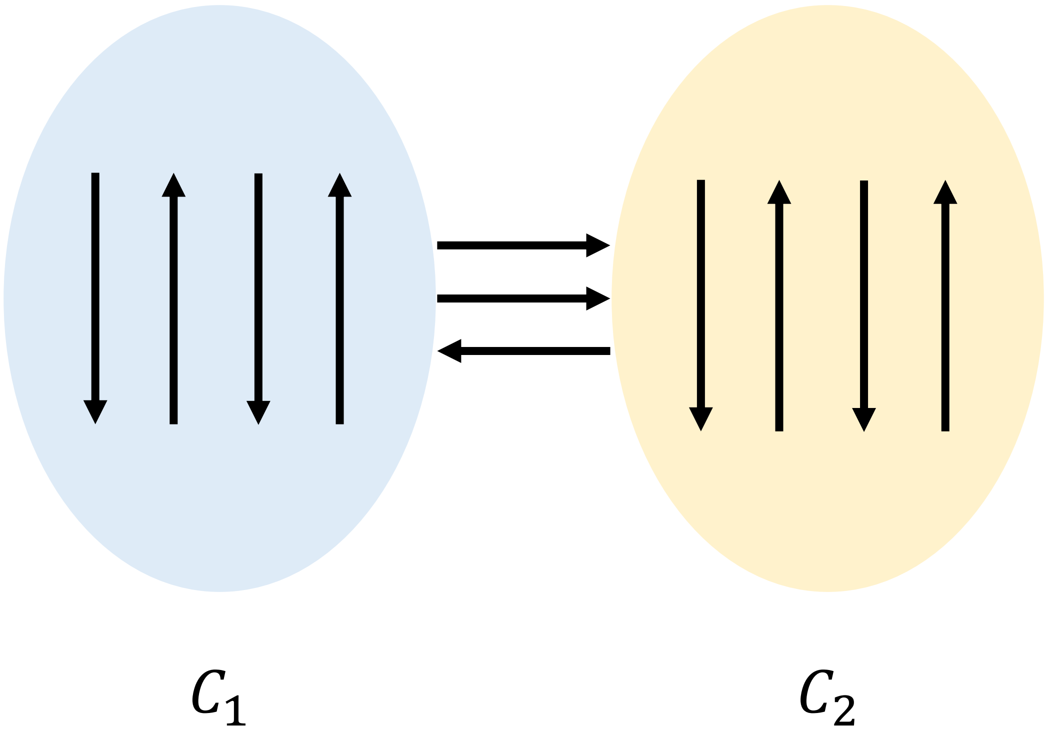
Apply maximum likelihood estimation on Directed-SBM $(N, p, q, \eta\ )$

MLE optimization goal (simplified illustration version)

$$\max \text{Net Flow} \ - \lambda\, \text{Total Flow}\,,$$

- **Total Flow**: $|C_1 \rightarrow C_2| + |C_2 \rightarrow C_1|$

- **Net Flow**: $|C_1 \rightarrow C_2| - |C_2 \rightarrow C_1|$



Key idea

Apply maximum likelihood estimation on Directed-SBM $(N, p, q, \eta\ )$

MLE optimization goal

$$\max_{x \in \{i,1\}^N} x^* H x \quad (\text{Herm}-\text{MLE})$$

where $H = i(A - A^T) + \lambda_1\ (A + A^T) + \lambda_2\ J$ (1) where $\lambda_1, \lambda_2 \in \mathbb{R}$ is function of p, q, η.

Key idea

Apply maximum likelihood estimation on Directed-SBM $(N, p, q, \eta\ )$

MLE optimization goal (complex version)

$$\max_{x \in \{i,1\}^N} x^* H x \quad (\text{Herm}-\text{MLE})$$

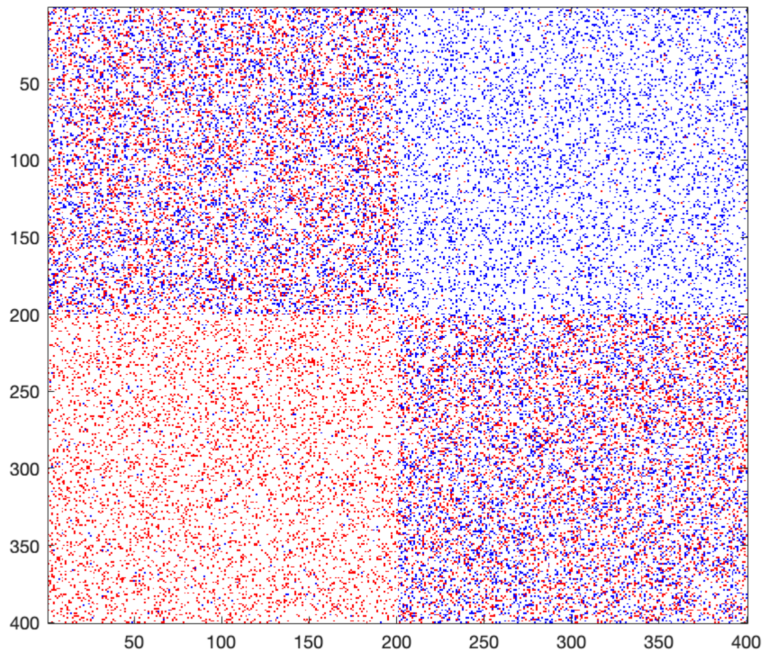
$$\text{where } H = i(A - A^T) + \lambda_1\,(A + A^T) + \lambda_2\, J \qquad (1)$$

✓Relax (Herm-MLE) to the PSD cone → Algorithm MLE-SDP

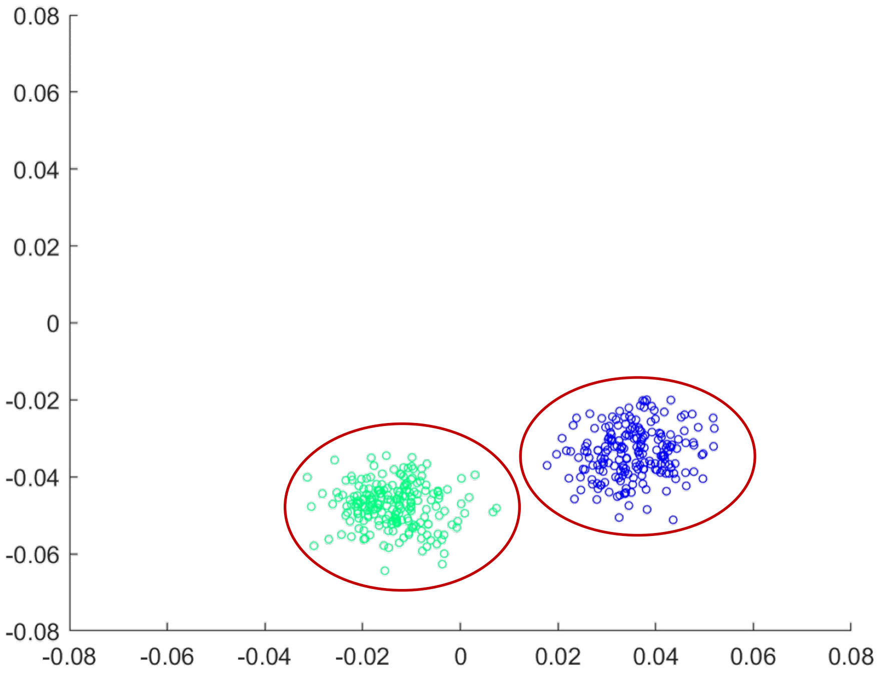
✓Relax (Herm-MLE) to C^N → Algorithm MLE-SC

Algorithms (MLE-SC)

- ➢Step 1. Compute the Hermitian matrix H according to (1) ;
- ➢Step 2. Compute the top eigenvector $\hat{v}$ of H;
- ➢Step 3. Apply k-means on the matrix $[Re(\hat{v}); Im(\hat{v})]$



H



Algorithms (MLE-SDP)

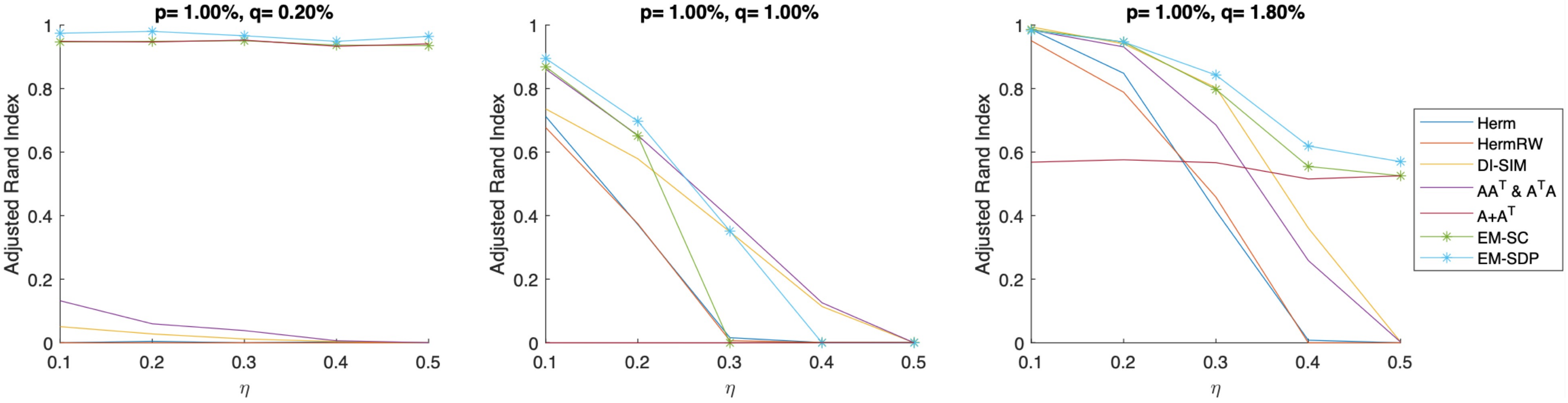
- Step 1. Compute the Hermitian matrix H according to (1) ;
- Step 2. Solve the following SDP

$$\max_{\substack{s.t.\, X \in \mathcal{H} \\ X \succcurlyeq 0 \\ \operatorname{diag}(X) = I}} \langle H, X \rangle \quad \text{(SDP-MLE)}$$

and compute the top eigenvector of $\hat{v}$;

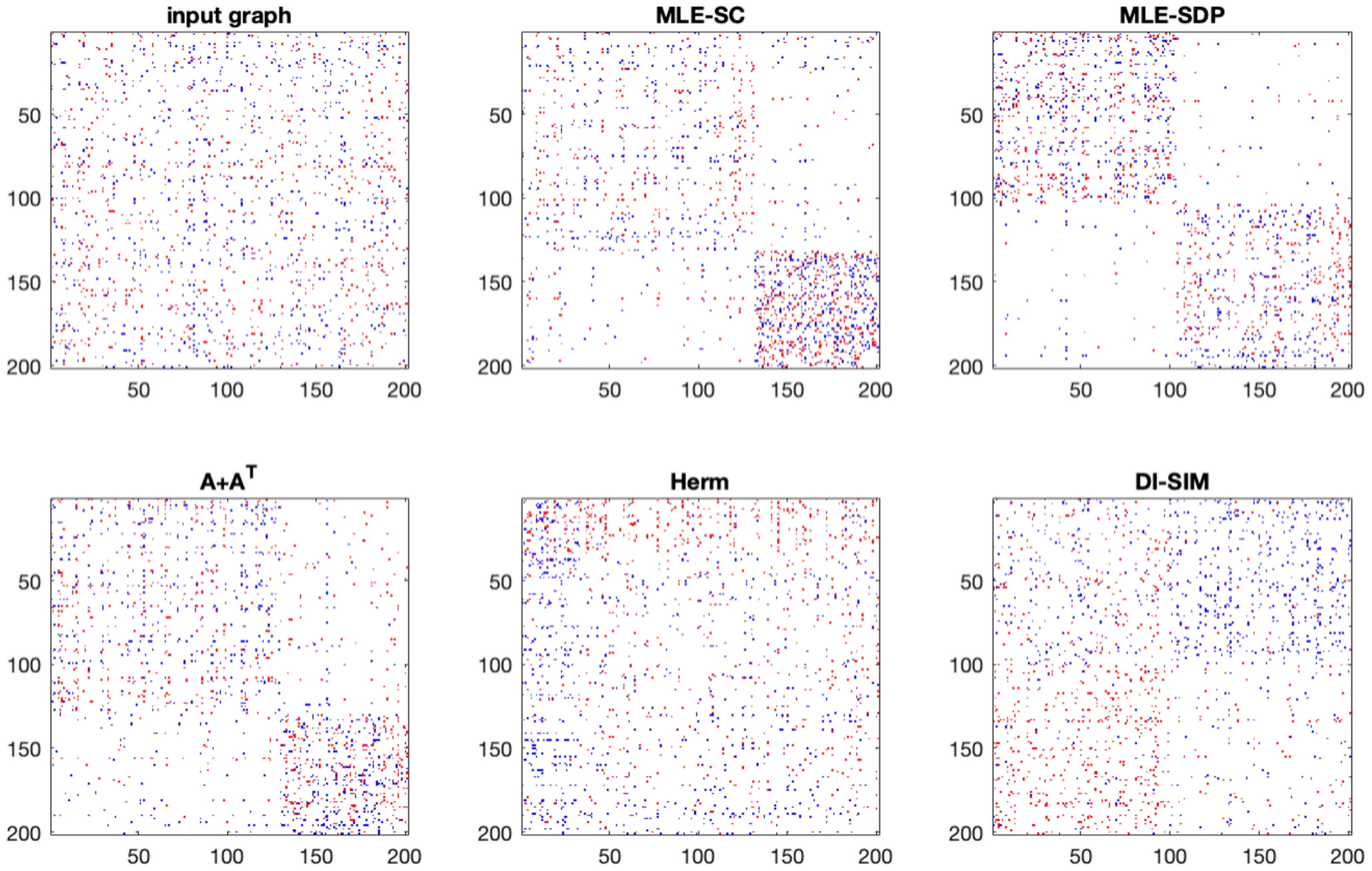
- Step 3. Apply k-means on the matrix $[Re(\hat{v}); Im(\hat{v})]$

Experiment on synthetic data



Experiments on graphs generated from the DSBM(N, p, q, η) ensemble, with different parameters.

Experiment on real-word digraphs



Before & after clustering the Email-Eu-core graph

Experiment on real-word digraphs

Data set	Herm	HermRW	$A^T A \& A A^T$	DI-SIM	$A + A^T$	MLE-SC	MLE-SDP
email-Eu-core [1]	0.045	-0.002	-0.007	-0.005	0.301	**0.608**	**0.757**
PolBlog [2]	0.012	-0.002	-0.001	-0.001	**0.206**	0.030	**0.809**

ARIs from test on real-world data.

[1] Yin, Hao, et al. "Local higher-order graph clustering." *Proceedings of the 23rd ACM SIGKDD international conference on knowledge discovery and data mining*. 2017.
[2] Adamic, Lada A., and Natalie Glance. "The political blogosphere and the 2004 US election: divided they blog." *Proceedings of the 3rd international workshop on Link discovery*. 2005.

Conclusion

- Derive the **MLE on DSBM** and use it as a new **directed clustering objective**
- Propose a **novel Hermitian matrix** representation for directed graphs
- Introduce **two directed clustering algorithms**
- (to appear) Prove a high probability **error bound**

Thanks for your attention!