

# Distributed Community Detection for Large Scale Networks Using Stochastic Block Model



**Shihao Wu, Zhe Li, Xuening Zhu**

School of Data Science, Fudan University

November 21, 2021








# Outline

- 1 Introduction
- 2 A Distributed Spectral Clustering Algorithm
- 3 Theoretical Properties
- 4 Numerical Study
- 5 Empirical Study
- 6 Summary

# Introduction

# Stochastic Block Model

- Adjacency matrix







				
		1	1	1
	1		1	
	1	1		1
	1		1	

- Adjacency matrix  $A = (a_{ij})$
- $a_{ij} = 1$  indicates the  $i$ th user follows the  $j$ th user; otherwise  $a_{ij} = 0$ .

Q: How to construct an adjacency matrix with SBM model?

# Stochastic Block Model

- Adjacency matrix

				
		1	1	1
	1		1	
	1	1		1
	1		1	

- Adjacency matrix  $A = (a_{ij})$
- $a_{ij} = 1$  indicates the  $i$ th user follows the  $j$ th user; otherwise  $a_{ij} = 0$ .

**Q: How to construct an adjacency matrix with SBM model?**

# Stochastic Block Model

- Membership matrix

Group:      1      2

	1		→ $g_i = 1$
	1		
		1	→ $g_i = 2$
		1	

- $\Theta = (\Theta_1, \dots, \Theta_N)^\top \in \mathbb{R}^{N \times K}$
- For the  $i$ th row of  $\Theta$ , only the  $g_i$ th element takes 1 and the others are 0.
- The membership matrix of the left figure is:

$$\begin{pmatrix} 1 & 0 \\ 1 & 0 \\ 0 & 1 \\ 0 & 1 \end{pmatrix}$$

# Stochastic Block Model

## Connectivity Matrix

Group:	1	2	
	1		→ $g_i = 1$
	1		
		1	→ $g_i = 2$
		1	

- $B \in \mathbb{R}^{K \times K}$  with full rank
- The connection probability between the  $k$ th and  $l$ th community is  $B_{kl}$
- The element  $A_{ij}$  in the adjacency matrix is generated independently from Bernoulli( $B_{g_i, g_j}$ ) distribution.

# Stochastic Block Model

- Construct an adjacency matrix with SBM model

Given  $\left\{ \begin{array}{l} \text{the **membership matrix** } \Theta \\ \text{the **connective matrix** } B \end{array} \right.$ ,

the adjacency matrix  $A = (a_{ij})_{1 \leq i, j \leq n}$  is generated as

$$a_{ij} = \begin{cases} \text{independent Bernoulli } (B_{g_i g_j}), & \text{if } i < j \\ 10, & \text{if } i = j \\ a_{ji}, & \text{if } i > j \end{cases}$$

**Q: How to detect the latent communities in the SBM?**



## ● Recover Community memberships:

★ **Likelihood based methods** (*Zhao et al. 2012*);

★ **Convex Optimization** (*Chen et al., 2012*);

★ **Methods of moments** (*Anandkumar et al., 2014*);

★ **Spectral clustering** (*Lei and Rinaldo, 2015; Rohe et al., 2011; Lei and Rinaldo, 2015*);

# Spectral Clustering for SBM

**Input:** Adjacency matrix  $A$ , numbers of communities  $K$

**Output:** Membership matrix  $\hat{\Theta}$

- 1 Compute Laplacian matrix  $L$  based on  $A$   
( $L = D^{-1/2}AD^{-1/2}$ ,  $D_{ii} = \sum_j A_{ij}$ ).
- 2 Conduct eigen-decomposition of  $L$  and extract the top  $K$  eigenvectors (i.e.,  $\hat{U}$  and the computational time is  $O(n^3)$ )
- 3 Conduct k-means algorithm using  $\hat{U}$  and then output the estimated membership matrix  $\hat{\Theta}$

Q: Does it work for large-scale network data?

# Spectral Clustering for SBM

**Input:** Adjacency matrix  $A$ , numbers of communities  $K$

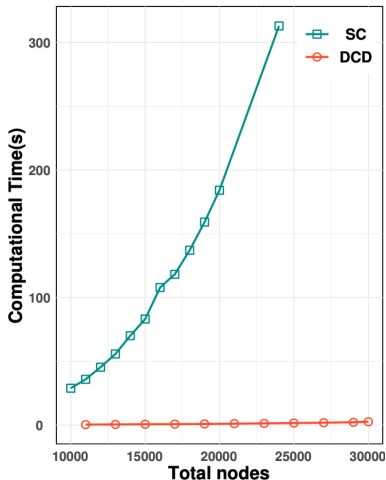
**Output:** Membership matrix  $\hat{\Theta}$

- 1 Compute Laplacian matrix  $L$  based on  $A$   
( $L = D^{-1/2}AD^{-1/2}$ ,  $D_{ii} = \sum_j A_{ij}$ ).
- 2 Conduct eigen-decomposition of  $L$  and extract the top  $K$  eigenvectors (i.e.,  $\hat{U}$  and the computational time is  $O(n^3)$ )
- 3 Conduct k-means algorithm using  $\hat{U}$  and then output the estimated membership matrix  $\hat{\Theta}$

**Q: Does it work for large-scale network data?**

# Spectral Clustering for SBM

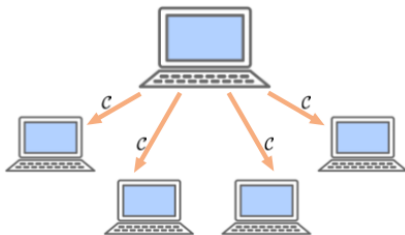
- Comparison of computational time



- The computational time of spectral clustering algorithm increases rapidly as the number of nodes grows
- The algorithm we proposed has the computational advantage

# A Distributed Algorithm for SBM

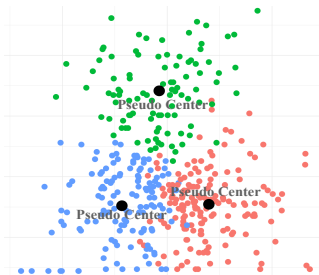
# Distributed System



- A distributed system typically consists of a master server and multiple worker servers.
- We can distribute  $l$  network nodes on master, who are referred to as pilot nodes.
- On the  $m$ th worker, we distribute  $n_m$  network nodes together with  $l$  pilot nodes.

# Distributed Community Detection

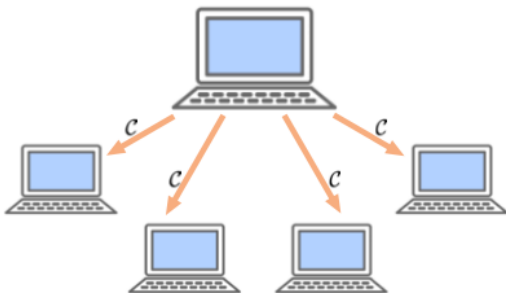
## Step 1: Pilot-based Network Spectral Clustering on Master



- Conduct eigen-decomposition of  $L_0$  and extract the top  $K$  eigenvectors (denoted in matrix  $\hat{U}_0$ )
- Conduct  $k$ -means algorithm and obtain clustering centers

# Distributed Community Detection

## Step 2: Broadcast Pseudo Centers to Workers

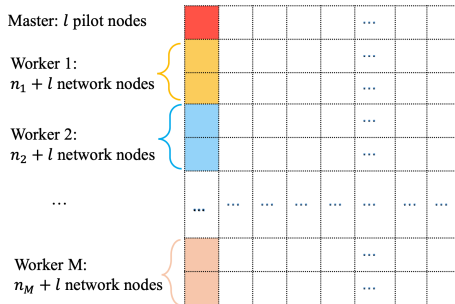


- Determine the indexes of the  $k$ th pseudo centers as
$$i_k = \arg \min_i \left\| \hat{U}_{0i} - \hat{C}_{0k} \right\|_2^2$$
- Broadcast the index set of pseudo centers  $\mathcal{C} = \{i_1, \dots, i_K\}$  to workers



# Distributed Community Detection

## Step 3: Community Detection on Workers



- Perform singular value decomposition using  $L^{(S_m)}$  and denote the top  $K$  left singular vector matrix as  $\hat{U}^{(S_m)}$

- Use

$$\hat{g}_i = \arg \min_{1 \leq k \leq K, i_k \in C} \left\| \hat{U}_i^{(S_m)} - \hat{U}_{i_k}^{(S_m)} \right\|^2$$

to obtain the estimated community labels.

# Theoretical Properties

# Theoretical properties on population level

## Proposition 1.

Let  $\Theta^{(S_m)} \in \mathbb{R}^{\bar{n}_m \times K}$  be the membership matrix on the  $m$ th worker. Then we have  $U^{(S_m)} = \Theta^{(S_m)} \mu$ , where  $\mu \in \mathbb{R}^{K \times K}$  is a rotation matrix, and

$$\mu^\top \Theta_i^{(S_m)} = \mu^\top \Theta_j^{(S_m)} \Leftrightarrow \Theta_i^{(S_m)} = \Theta_j^{(S_m)}$$

## Remarks:

- ★ The singular vectors could play the same role as the eigenvectors of the adjacency matrix in the community detection.

# Theoretical properties on population level

**Proposition 2.** An upper bound for the deviation of  $U^{(S_m)}$  from  $r_m^{-1/2} U_m$ .

Let  $b_{\min} = \min_{1 \leq i, j \leq K} B_{ij}$ . It holds

$$\left\| U^{(S_m)} - r_m^{-1/2} U_m Q_m \right\|_F \leq \frac{14\sqrt{2}K^2 u_m \max\{u_0^{1/2}, u_m^{1/2}\} \alpha^{(S_m)1/2}}{\sigma_{\min}(B) b_{\min}^3 d_0^2 d_m^3 (d_0 + d_m)} + \frac{\alpha^{(S_m)}}{d_0}$$

## Remarks:

- ★  $U_m = (U_i : i \in S_m)^\top \in \mathbb{R}^{\bar{n}_m \times K}$  is the sub-matrix of  $U$  (the eigenvector matrix of  $\mathcal{L}$ , i.e.,  $\mathcal{L} = U\Lambda U^\top$ )
- ★  $\alpha^{(S_m)} = \max_k |\bar{n}_{mk}/\bar{n}_m - m_k/N|$  is the unbalanced effect ( $\bar{n}_m = l + n_m$  with  $|S_m| = \bar{n}_m$ )
- ★  $d_0 \leq \min_k n_{0k}/l \leq \max_k n_{0k}/l \leq u_0$  and  $d_m \leq \min_k n_{mk}/\bar{n}_m \leq \max_k n_{mk}/\bar{n}_m \leq u_m$
- ★ The upper bound illustrates the relationship between the error bounds and the unbalanced effect

# Convergence of Singular Vectors

## Theorem 1. Singular vector convergence

Let  $\lambda_{1,m} \geq \lambda_{2,m} \geq \dots \geq \lambda_{K,m} > 0$  be the top  $K$  singular values of  $\mathcal{L}^{(S_m)}$ . Define  $\delta_m = \min_i \mathcal{D}_{ii}^{(S_m)}$ . Then for any  $\epsilon_m > 0$  and  $\delta_m > 3 \log(n_m + 2l) + 3 \log(4/\epsilon_m)$ , with probability at least  $1 - \epsilon_m$  it holds

$$\left\| \hat{U}^{(S_m)} - U^{(S_m)} Q^{(S_m)} \right\|_F \leq \frac{8\sqrt{6}}{\lambda_{K,m}} \sqrt{\frac{K \log(4(n_m + 2l)/\epsilon_m)}{\delta_m}}$$

### Comments:

- ★  $\mathcal{D}^{(S_m)} = E \left[ D^{(S_m)} \right] \in \mathbb{R}^{\bar{n}_m \times \bar{n}_m}$  ( $D_{ii}^{(S_m)} = \sum_j A_{ij}^{(S_m)}$  is the out-degrees of node  $i$ )
- ★ The error bound is related to  $\lambda_{K,m}$ . If  $\lambda_{K,m}$  is larger, the eigengap between the eigenvalues of interest and the rest will be higher. This enables us to detect communities with higher accuracy level.
- ★ The upper bound is lower if the minimum out-degree  $\delta_m$  is higher.

# Clustering Accuracy Analysis

## Proposition 3.

The mode  $i$  will be correctly clustered (i.e.  $\hat{g}_i = g_i$ ) as long as

$$\left\| \hat{U}_i^{(S_m)} - \hat{C}_{g_i}^{(S_m)} \right\|_2 < \frac{P_m}{2}$$

## Remarks:

- ★  $\hat{\mathbf{C}}(S_m) = \left( \hat{C}_1^{(S_m)}, \dots, \hat{C}_K^{(S_m)} \right)^\top \in \mathbb{R}^{K \times K}$  be the pseudo centers on the worker  $m$
- ★  $P_m = (2/D_m)^{1/2} - 2\zeta_m$  with  $D_m = \max_{1 \leq k \leq K} \bar{n}_{mk}$  and  $\zeta_m = \max_{k \in \{1, \dots, K\}} \left\| Q^{(S_m)\top} U_{i_k}^{(S_m)} - \hat{C}_k^{(S_m)} \right\|_2$
- ★ If with a high probability that the pseudo nodes are correctly clustered, then  $P_m$  will be higher

# A lower bound for $P_m$

## Technical Conditions

- (C1) (EIGENVALUE AND EIGENGAP ON MASTER) Let  $\delta_0 = \min_i \mathcal{D}_{0,ii}$ . Assume  $\delta_0 > 3 \log(2l) + 3 \log(4/\varepsilon_l)$  and  $\varepsilon_l \rightarrow 0$  as  $l \rightarrow \infty$
- (C2) (PILOT NODES) Assume  $K^2 \log(l/\varepsilon_l) / (b_{\min} \lambda_{K,0}^2) \ll l$  with  $\varepsilon_l \rightarrow 0$  as  $l \rightarrow \infty$
- (C3) (UNBALANCED EFFECT) Let  $d_0, d_m, u_0, u_m$  be finite constants and assume  $\alpha^{(\mathcal{S}_m)} = o(\sigma_{\min}(B)^2/K^4)$

**Proposition 4.** Assume Conditions (C1)-(C3).

Then with probability  $1 - \varepsilon_l$ , we have  $P_m \geq c_1 / \sqrt{n_m}$  as  $\min \{l, n_m\} \rightarrow \infty$  with rotation  $Q_c$ , where  $c_1$  is a positive constant.

# Bound of mis-clustering Rates

**Theorem 2.** Assume conditions in Theorem 1 and Proposition 4

Denote  $\mathcal{R}^{(S_m)}$  as the ratio of misclustered nodes on worker  $m$ , then we have

$$\mathcal{R}^{(S_m)} = o\left(\frac{K^2 \log(l/\epsilon_l)}{b_{\min} \lambda_{K,0}^2} + \frac{K \log(4(n_m + 2l)/\epsilon_m)}{\lambda_{K,m} \delta_m} + \frac{K^4 \alpha^{(S_m)}}{\sigma_{\min}(B)^2 b_{\min}^6}\right)$$

with probability at least  $1 - \epsilon_l - \epsilon_m$

- ★ The first and second terms are related to convergence of spectrum on master and workers.
- ★ the third term is mainly related to the unbalanced effect  $\alpha^{(S_m)}$  among the workers.



# Bound of mis-clustering Rates

**Corollary 1.** Assume the same conditions as in Theorem 2

In addition, assume  $n_1 = n_2 = \dots = n_M \stackrel{\text{def}}{=} n$  and  $\alpha^{(S_m)} = 0$  for  $1 \leq m \leq M$ . Denote  $\mathcal{R}_{all}$  as number of all mis-clustered nodes across all workers. Then with probability  $1 - (M+1)/l$  we have

$$\mathcal{R}_{all} = O\left(\frac{K(\log n + \log l)}{l \lambda_K^2}\right)$$

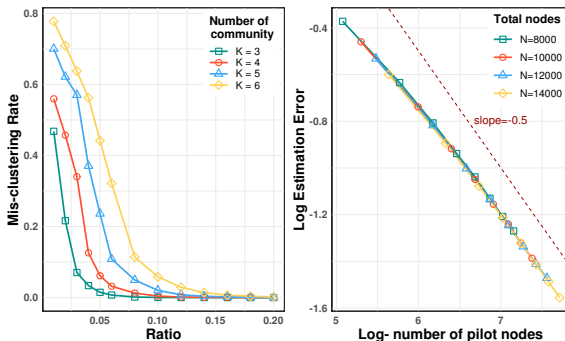
where  $\lambda_K = \min_m \lambda_{K,m}$ .

## Remarks:

- ★ If  $l = rN$  with  $r \in (0, 1)$  being a finite positive constant, then the mis-clustering rate is almost the same as we use the whole adjacency matrix  $A$
- ★ The computational time is roughly  $r^2$  smaller than using the whole adjacency matrix.

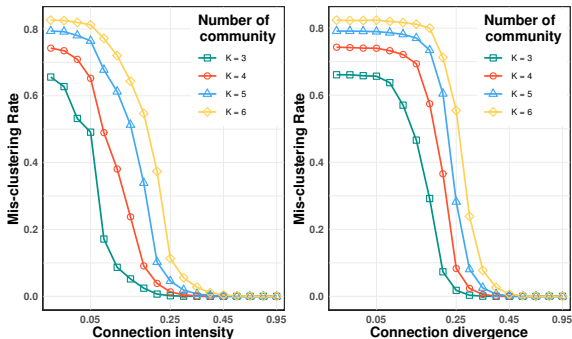
# Numerical Study

# The effect of pilot nodes



- The mis-clustering rate converges to zero as  $l$  grows
- The estimation error of eigenvectors decreases with the slope of LEE (Log Estimation Error) roughly parallel with  $-\frac{1}{2}$  as  $\log(l)$  grows.

# The effect of signal strength

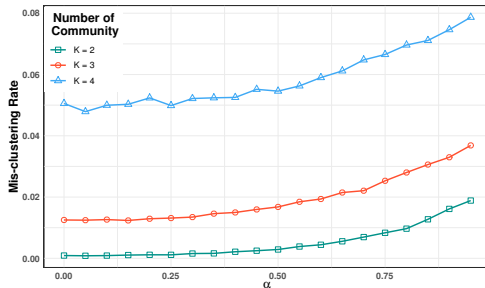


- The connectivity matrix  $B$  is set as  $B = v \{ \lambda I_K + (1 - \lambda) \mathbf{1}_K \mathbf{1}_K^T \}$ , the connection intensity is then parameterized by  $v$  and the connection divergence is characterized by  $\lambda$
- As  $\lambda$  increases, the mis-clustering rate converges to zero.
- As  $v$  increases, the signal strength will increase accordingly, which results on a smaller mis-clustering rate

# Unbalanced effect

Denote  $\pi_{mk}$  as the ratio of nodes in the  $k$ th community on the  $m$ th worker. We set  $\pi_{mk}$  as

$$\pi_{mk} = \frac{1}{K} + \left(k - \frac{K+1}{2}\right) \text{sign}\left(m - \frac{M+1}{2}\right) \frac{\alpha}{K(K-1)}$$



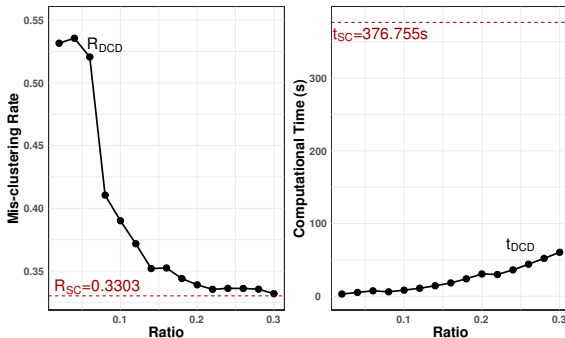
The mis-clustering rates increase, which verifies the result of Theorem 2

# Empirical Study

# Pubmed Dataset: a Citation Network

- The Pubmed dataset consists of **19,717** scientific publications from PubMed database
- Each publication is identified as one of the three classes, i.e., Diabetes Mellitus Experimental, Diabetes Mellitus Type 1, Diabetes Mellitus Type 2. The sizes of the three classes are **4,103**, **7,875**, and **7,739** respectively.
- Specifically, if the  $i$ th publication cites the  $j$ th one (or otherwise), then  $A_{ij} = 1$ , otherwise  $A_{ij} = 0$ . The resulting network density is **0.028%**.
- We use both spectral clustering algorithm(*SC*) and distributed clustering detection algorithm(*DCD*) on Pubmed dataset for comparison.

# Pubmed Dataset: a Citation Network



The mis-clustering rates of the DCD algorithm is comparable to the SC algorithm when

$r = \frac{l}{N} = 0.22$ , while the computational time is much lower.



# Summary

- 1 We propose a distributed community detection (DCD) algorithm to tackle community detection task in large scale networks ( $O(MI^3 + NI^2)$ ).
- 2 We provide rigorous **theoretical analyses** of both parameter estimation and computational complexity.
- 3 As for future studies, better mechanisms can be designed to select pilot nodes on the master server and it is interesting to extend the proposed method to directed network by considering sending and receiving clusters respectively.

