Distributed Community Detection for Large Scale Networks Using Stochastic Block Model

Shihao Wu, Zhe Li, Xuening Zhu

School of Data Science, Fudan University

October 12, 2020

Outline

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 ith user knows the jth user; otherwise $a_{ij} = 0$.

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

Stochastic Block Model

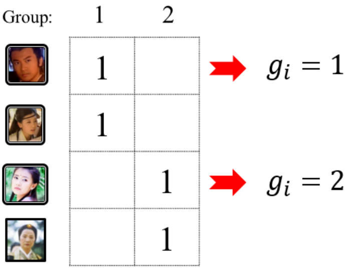
- Adjacency matrix

				
		1	1	1
	1		1	
	1	1		1
	1		1	

- Adjacency matrix $A = (a_{ij})$
- $a_{ij} = 1$ indicates the ith user knows the jth user; otherwise $a_{ij} = 0$.

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

- Membership matrix

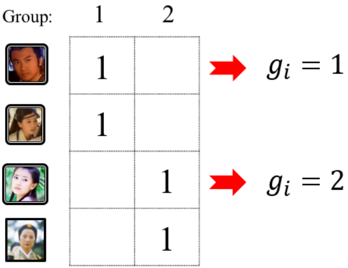


- $\Theta = (\Theta_1, \cdots, \Theta_N)^\top \in \mathbb{R}^{N \times K}$
- For the ith row of Θ, only the g_ith element takes 1 and the others are 0.
- The membership matrix of the left figure is:

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

Stochastic Block Model

- Connectivity Matrix



- $B \in \mathbb{R}^{K \times K}$ with full rank
- The connection probability between the kth and lth 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

$$\text{Given} \begin{cases} \text{the } \textbf{membership matrix } \Theta \\ \text{the } \textbf{connective matrix } B \end{cases},$$

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

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

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

Literature Review

- **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)*;

- **Spectral Clustering under Stochastic Block Models**:

  ★One of the most widely used methods for community detection (*Rohe et al., 2011; Lei and Rinaldo, 2015; Jin et al., 2015; Sarkar et al., 2015; Lei et al., 2020)*

  ★The consistency of spectral clustering (*Rohe et al. 2011 and Lei and Rinaldo 2015)*

Spectral Clustering for SBM

Input: Adjacency matrix A, numbers of communities K

Output: Membership matrix $\widehat{\Theta}$

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

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

Spectral Clustering for SBM

Input: Adjacency matrix A, numbers of communities K

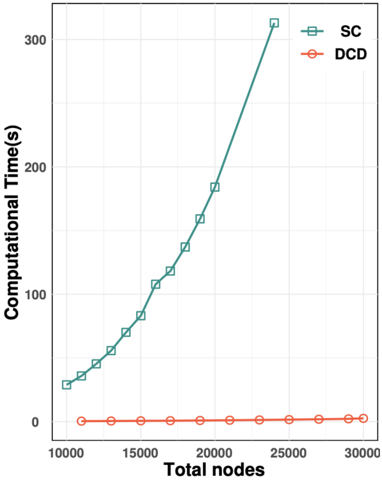
Output: Membership matrix $\widehat{\Theta}$

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

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

Spectral Clustering for SBM

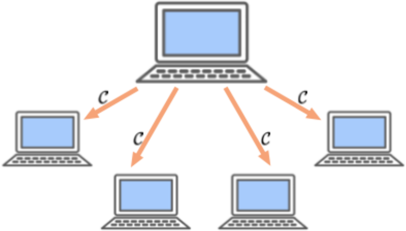
- Comparison of computational time



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

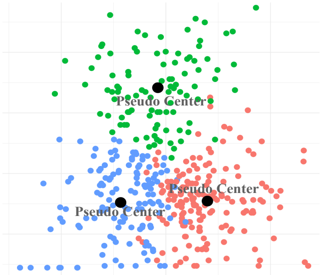
A Distributed Algorithm for SBM

Distributed System



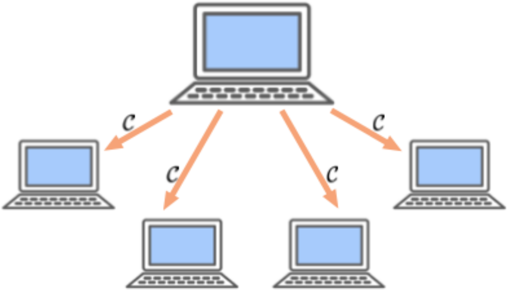
- A distributed system typically consists of a master server and multiple worker servers.
- We can distribute l network nodes on master, who are referred to as pilot nodes.
- On the mth worker, we distribute n_m network nodes together with l pilot nodes.

Step 1: Pilot-based Network Spectral Clustering on Master



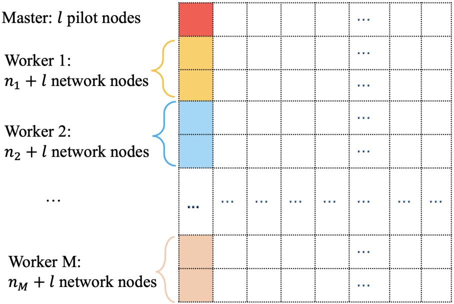
- Conduct eigen-decomposition of L_0 and extract the top K eigenvectors (denoted in matrix $\widehat{U}_0$)
- Conduct k-means algorithm and obtain clustering centers

Step 2: Broadcast Pseudo Centers to Workers



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

Step 3: Community Detection on Workers



- Perform singular value decomposition using $L^{(\mathcal{S}_m)}$ and denote the top K left singular vector matrix as $\widehat{U}^{(\mathcal{S}_m)}$
- Use

$$\widehat{g}_i = \arg \min_{1 \le k \le K, i_k \in \mathcal{C}} \left\| \widehat{U}_i^{(\mathcal{S}_m)} - \widehat{U}_{i_k}^{(\mathcal{S}_m)} \right\|^2$$

to obtain the estimated community labels.

Theoretical Properties

Theoretical properties on population level

Proposition 1.

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

$$\boxed{\mu^\top \Theta_i^{(\mathcal{S}_m)} = \mu^\top \Theta_j^{(\mathcal{S}_m)} \Leftrightarrow \Theta_i^{(\mathcal{S}_m)} = \Theta_j^{(\mathcal{S}_m)}}$$

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 $U^{(\mathcal{S}_m)}$ from $r_m^{-1/2}U_m$.

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

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

Remarks:

- ★ $\alpha^{(\mathcal{S}_m)} = \max_k |\bar{n}_{mk}/\bar{n}_m - m_k/N|$ is the unbalanced effect
- ★ $d_0 \le \min_k n_{0k}/l \le \max_k n_{0k}/l \le u_0$ and
  $d_m \le \min_k n_{mk}/\bar{n}_m \le \max_k n_{mk}/\bar{n}_m \le u_m$
- ★ The upper bound illustrates the relationship between the error bounds and the unbalanced effect

Convergence of Singular Vectors

Theorem 1. Singular vector convergence

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

$$\boxed{\left\| \widehat{U}^{(\mathcal{S}_m)} - U^{(\mathcal{S}_m)} Q^{(\mathcal{S}_m)} \right\|_F \leq \frac{8\sqrt{6}}{\lambda_{K,m}} \sqrt{\frac{K \log(4(n_m+2l)/\varepsilon_m)}{\delta_m}}}$$

Comments:

★ 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 δ_m is higher.

Clustering Accuracy Analysis

Proposition 3.

Node i will be correctly clustered (i.e $\widehat{g}_i = g_i$) as long as

$$\boxed{\left\| \widehat{U}_i^{(\mathcal{S}_m)} - \widehat{C}_{g_i}^{(\mathcal{S}_m)} \right\|_2 < \frac{P_m}{2}}$$

Remarks:

★ $P_m = (2/D_m)^{1/2} - 2\zeta_m$ with $D_m = \max_{1 \le k \le K} \bar{n}_{mk}$ and $\zeta_m = \max_{k \in \{1,\ldots,K\}} \left\| Q^{(\mathcal{S}_m)\top} U_{i_k}^{(\mathcal{S}_m)} - \widehat{C}_k^{(\mathcal{S}_m)} \right\|_2$

★ 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 $\varepsilon_l \to 0$ as $l \to \infty$

(C2) (PILOT NODES) Assume $K^2 \log(l/\varepsilon_l) / \left(b_{\min}\lambda_{K,0}^2\right) \ll l$ with $\varepsilon_l \to 0$ as $l \to \infty$

(C3) (UNBALANCED EFFECT) Let d_0, d_m, u_0, u_m be finite constants and assume $\alpha^{(\mathcal{S}_m)} = o\left(\sigma_{\min}(B)^2/K^4\right)$

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

Then with probability $1-\varepsilon_l$, we have $P_m \geq c_1/\sqrt{\bar{n}_m}$ as $\min\{l, n_m\} \to \infty$ 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 $\mathcal{R}^{(\mathcal{S}_m)}$ as the ratio of misclustered nodes on worker m, then we have

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

with probability at least $1-\varepsilon_l-\varepsilon_m$

- ★ 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 $\alpha^{(\mathcal{S}_m)}$ among the workers.

Bound of mis-clustering Rates

Corollary 1. Assume the same conditions as in Theorem 2

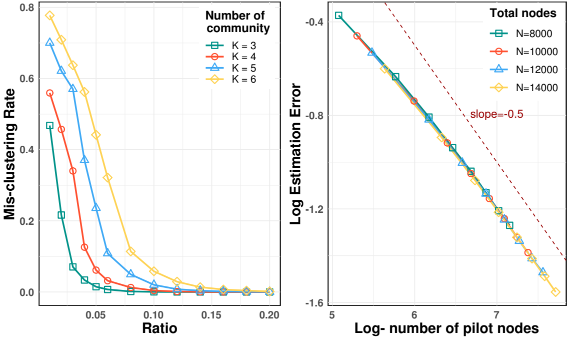
In addition, assume $n_1 = n_2 = \cdots = n_M \overset{\text{def}}{=} n$ and $\alpha^{(\mathcal{S}_m)} = 0$ for $1 \le m \le M$. Denote $\mathcal{R}_{all}$ asnumberofall mis-clustered nodes across all workers. Then with probability $1 - (M+1)/l$ we have

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

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

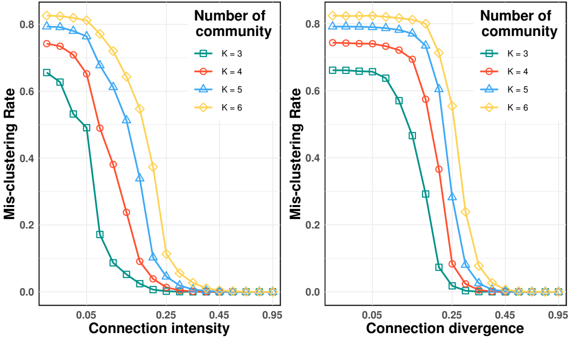
Numerical Study

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

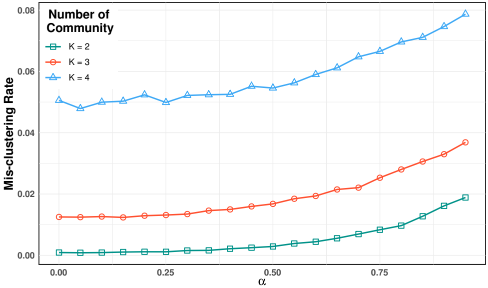


- The connectivity matrix B is set as $B = \nu\left\{\lambda I_K + (1-\lambda)\mathbf{1}_K\mathbf{1}_K^\top\right\}$, the connection intensity is then parameterized by ν and the connection divergence is characterized by λ
- As λ increases, the mis-clustering rate converges to zero.
- As ν increases,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

Unbalanced effect

Denote π_{mk} as the ratio of nodes in the kth community on the mth worker. We set π_{mk} as

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



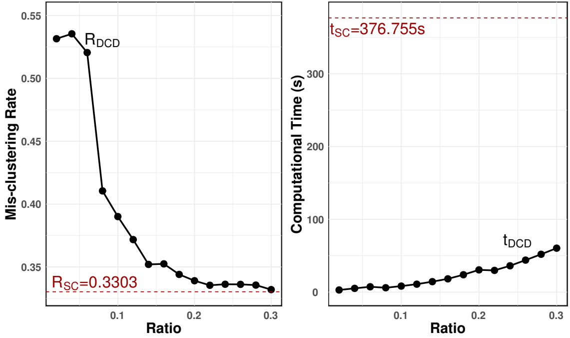
The mis-clustering rates increase, which verifies the result of Theorem 2

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 ith publication cites the jth 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.
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.

The End

- Wu S., Li Z., Zhu X. (2020), "Distributed Community Detection for Large Scale Networks Using Stochastic Block Model", **https://arxiv.org/abs/2009.11747**.
- DCD Package: **https://github.com/lkerlz/dcd**