A Distributed Community Detection Algorithm for Large Scale Networks Under Stochastic Block Models

Zhe Li

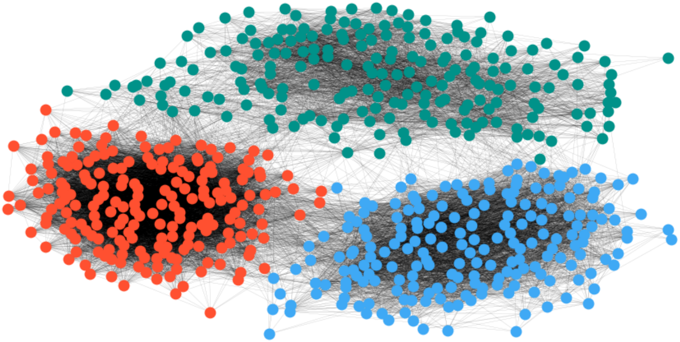
Joint work with Shihao Wu and Xuening Zhu

December 1, 2023

Outline

Introduction

- Communities in Yelp dataset (https://www.yelp.com/dataset)



Introduction

- Community detection is a fundamental task within network analysis
- Numerous methodologies exist for this task.:
  - Likelihood based methods (Zhao et al. 2012)
  - Convex Optimization (Chen et al., 2012)
  - Methods of moments (Anandkumar et al., 2014)
  - Spectral clustering (Rohe et al., 2011; Lei and Rinaldo, 2015);

Algorithm 1: Spectral Clustering for SBM (SC)

Input: Adjacency matrix A; number of communities K.

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

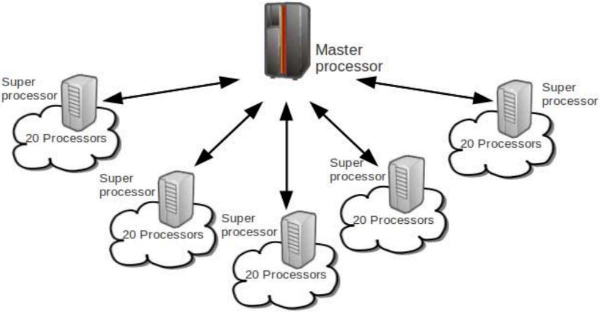
1: Compute Laplacian matrix L based on A.

2: Conduct eigen-decomposition of L and extract the top K eigenvectors (i.e., $\widehat{U}$).

3: Conduct k-means algorithm using $\widehat{U}$ and then output the estimated membership matrix $\widehat{\Theta}$.

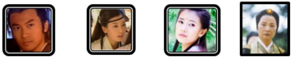
Introduction

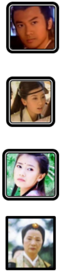
- What if the network is of large scale? $\Rightarrow$ great computational power
- privacy? $\Rightarrow$ stored in a distributed manner across various data centers.



Can we consider a distributed algorithm for the spectral clustering?

Distributed Community Detection under SBM





	1	1	1
1		1	
1	1		1
1		1	

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

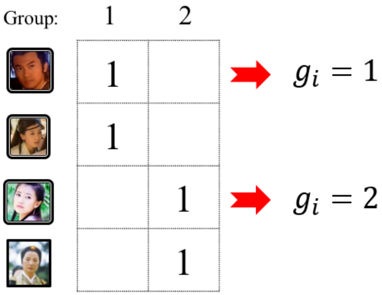
Stochastic Block Model: Membership matrix

- $\Theta = (\Theta_1, \cdots, \Theta_N)^\top \in \mathbb{R}^{N \times K}$
- 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.

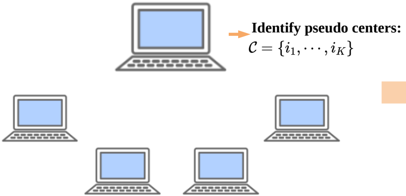
Spectral Clustering under SBM

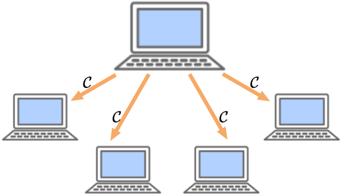
Lemma 1. (Lemma 3.1 in Rohe et al. (2011)).

The eigen-decomposition of $\mathcal{L}$ takes the form $\mathcal{L} = U\Sigma U^\top$, where $U = (U_1, \cdots, U_N)^\top \in \mathbb{R}^{N\times K}$ collects the eigen-vectors and $\Sigma \in \mathbb{R}^{K\times K}$ is a diagonal matrix. Further we have $U = \Theta\mu$, where μ is a $K \times K$ orthogonal matrix and $\Theta_i = \Theta_j$ if and only if $U_i = U_j$.

- $\mathcal{L} = \mathcal{D}^{-1/2}\mathcal{A}\mathcal{D}^{-1/2}$, where $\mathcal{A} = \mathbb{E}(A)$ and $\mathcal{D} = \mathbb{E}(D)$
- U only has K distinct rows and the ith row is equal to the jth row if the corresponding two nodes belong to the same community

A Distributed Algorithm





Step 1:

- Conduct spectral clustering on master server to identify pseudo centers.

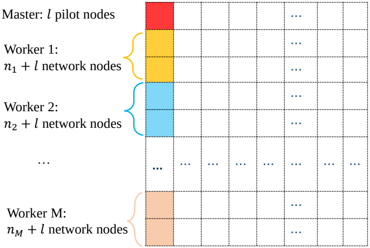
Step 2:

- Broadcast pseudo centers to workers
- Complete distributed community detection task using a SVD type algorithm.

Pilot Network Spectral Clustering on Master Server

- Suppose we have l network nodes on the master $\Rightarrow$ **pilot nodes**.
- In addition we distribute the pilot nodes both on master and workers.

- Conduct the spectral clustering on the pilot network $A_0 \in \mathbb{R}^{l \times l}$ and obtain the clustering centers $\widehat{C}_0 = \left(\widehat{C}_{0k} : 1 \le k \le K\right)^\top$.
- 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.



Community Detection on Workers

- Suppose we distribute n_m network nodes as well as the pilot nodes on the m th worker $\Rightarrow \bar{n}_m = l + n_m$.
- Denote the corresponding sub-adjacency matrix as $A^{(\mathcal{S}_m)} \in \mathbb{R}^{\bar{n}_m \times l}$.
- Permute the row indexes of $A^{(\mathcal{S}_m)}$ to ensure that $A^{\left(\mathcal{S}_m\right)} = \left(A_1^{(\mathcal{S}_m)\top}, A_2^{(\mathcal{S}_m)^\top}\right)^\top$ with $A_1^{(\mathcal{S}_m)} = A_0$.
- Let $D_{ii}^{(\mathcal{S}_m)} = \sum_j A_{ij}^{(\mathcal{S}_m)}$ and $F_{jj}^{(\mathcal{S}_m)} = \sum_i A_{ij}^{(\mathcal{S}_m)}$ be the out- and in-degrees of node i and j in the subnetwork on worker m.
- Define

$$D^{(\mathcal{S}_m)} = \mathrm{diag}\left\{D_{ii}^{(\mathcal{S}_m)} : 1 \le i \le \bar{n}_m\right\} \in \mathbb{R}^{\bar{n}_m \times \bar{n}_m}$$

$$F^{(\mathcal{S}_m)} = \mathrm{diag}\left\{F_{jj}^{(\mathcal{S}_m)} : 1 \le j \le l\right\} \in \mathbb{R}^{l \times l}$$

Community Detection on Workers

- The Laplacian version of $A^{(\mathcal{S}_m)}$ is given by

$$L^{(\mathcal{S}_m)} = \left(D^{(\mathcal{S}_m)}\right)^{-1/2} A^{(\mathcal{S}_m)} \left(F^{(\mathcal{S}_m)}\right)^{-1/2} \in \mathbb{R}^{\bar{n}_m \times l}$$

- Perform SVD using $L^{(\mathcal{S}_m)}$ and denote the top K left singular vector matrix as $\widehat{U}^{(\mathcal{S}_m)}$.
- For the i th $(l+1 \leq i \leq \bar{n}_m)$ node in $\mathcal{S}_m$, the cluster label g_i is estimated by

$$\widehat{g}_i = \mathrm{argmin}_{1 \leq k \leq K, i_k \in \mathcal{C}} \left\| \widehat{U}_i^{(\mathcal{S}_m)} - \widehat{U}_{i_k}^{(\mathcal{S}_m)} \right\|_2 .$$

复旦大學 Fudan University

Extend to Degree-corrected SBM

Let $\Gamma = \text{diag}\{\Gamma_i, 1 \le i \le N\} \in \mathbb{R}^{N\times N} \Rightarrow \mathbb{E}(A) = \Gamma\Theta B\Theta^\top\Gamma$

Algorithm 4: Regularized Distributed Community Detection (r-DCD)

Input: Adjacency matrix A_0; sub-adjacency matrices $\{A^{(\mathcal{S}_m)}\}_{m=1,\cdots,M}$; regularization parameter τ; number of communities K.

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

STEP 1 PILOT-BASED NETWORK SPECTRAL CLUSTERING ON MASTER SERVER

- STEP 1.1 Let $L_{0\tau} = D_{0\tau}^{-1/2} A_0 D_{0\tau}^{-1/2}$, where $D_{0\tau} = D_0 + \tau I$. Conduct eigen-decomposition of $L_{0\tau}$ and extract the top K eigenvectors (denoted in matrix $\widehat{U}_0$).
- STEP 1.2 Normalize each row of $\widehat{U}_0$ with unit L_2-norm and obtain $\widehat{U}_{0\tau}$.
- STEP 1.3 Conduct k-means algorithm on $\widehat{U}_{0\tau}$ and obtain clustering centers $\widehat{C}_0 = \left(\widehat{C}_{0k} : 1 \le k \le K\right)^\top$.

STEP 2 BROADCAST PSEUDO CENTERS TO WORKERS

- STEP 2.1 Determine the indexes of the kth pseudo centers as $i_k = \arg\min_i \|\widehat{U}_{0\tau,i} - \widehat{C}_{0k}\|_2^2$, where $\widehat{U}_{0\tau,i}$ is the ith row vector of $\widehat{U}_{0\tau}$.
- STEP 2.2 Broadcast the index set of pseudo centers $\mathcal{C} = \{i_1, \cdots, i_K\}$ to workers.

STEP 3 COMMUNITY DETECTION ON WORKERS

- STEP 3.1 Let $L_\tau^{(\mathcal{S}_m)} = (D^{(\mathcal{S}_m)} + \tau I)^{-1/2} A^{(\mathcal{S}_m)} (F^{(\mathcal{S}_m)} + \tau I)^{-1/2}$. Perform singular value decomposition using $L_\tau^{(\mathcal{S}_m)}$ and denote the top K left singular vector matrix as $\widehat{U}^{(\mathcal{S}_m)}$.
- STEP 3.2 Normalize each row of $\widehat{U}^{(\mathcal{S}_m)}$ with unit L_2-norm and obtain $\widehat{U}_\tau^{(\mathcal{S}_m)}$.
- STEP 3.3 Use (3) to obtain the estimated community labels.

復旦大學 Fudan University

Theoretical Properties

Theorem 3.1. (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 $\delta_m = \min_i \mathcal{D}_{\mathrm{ii}}^{(\mathcal{S}_m)}$. 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

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

where $Q^{(\mathcal{S}_m)} \in \mathbb{R}^{K \times K}$ is a $K \times K$ orthogonal matrix.

- **Remarks:**
  - The error bound is related to the eigen-gap $\lambda_{K,m}$
  - The upper bound is lower if the minimum out-degree δ_m is higher

Theoretical Properties

Theorem 3.2. (Bound if Mis-clustering Rates)

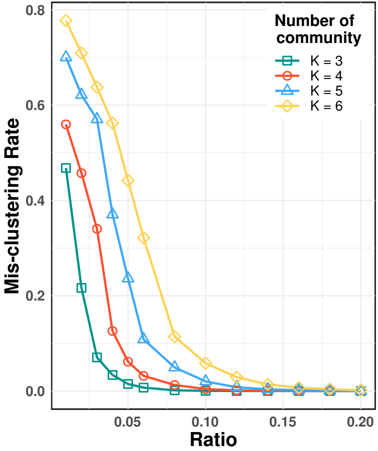
Assume some conditions hold. Let $\mathcal{R}^{(\mathcal{S}_m)}$ denote the ratio of misclustered nodes on worker m, then we have

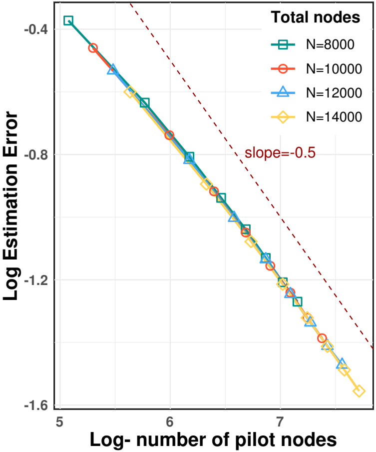
$$\mathcal{R}^{(\mathcal{S}_m)} = O\left(\frac{u_m K^2 \log\left(l/\epsilon_l\right)}{d_0 b_{\min} l \lambda_{K,0}^2} + \frac{K \log\left(4\left(n_m + 2l\right)/\epsilon_m\right)}{\lambda_{K,m}\delta_m} + \frac{u_m \alpha_0^2 K + d_0 \alpha_m^2 K}{d_0 d^2}\right)$$

with probability at least $1 - \epsilon_l - \epsilon_m$, where $u_m = \max_k \pi_k^{(\mathcal{S}_m)}$.

- **Remarks:**
  - The first term is related to the convergence of eigenvectors on the maste
  - The second term is determined by convergence of singular vectors on the mth worker.
  - the third term is mainly related to the unbalanced effect α_m among the workers and α_0 on the master.

Simulation: Pilot Nodes

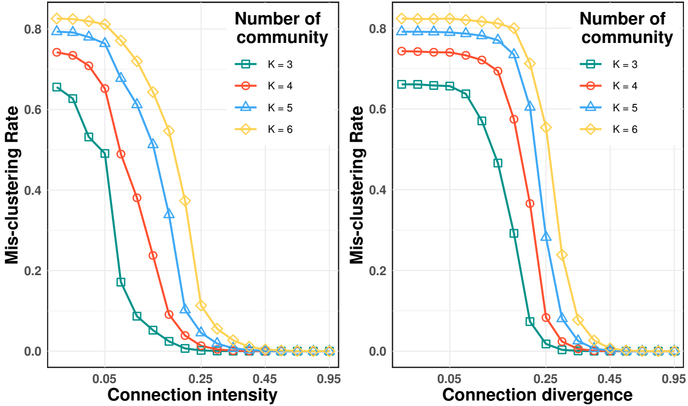




Fudan University

Simulation: Signal Strength

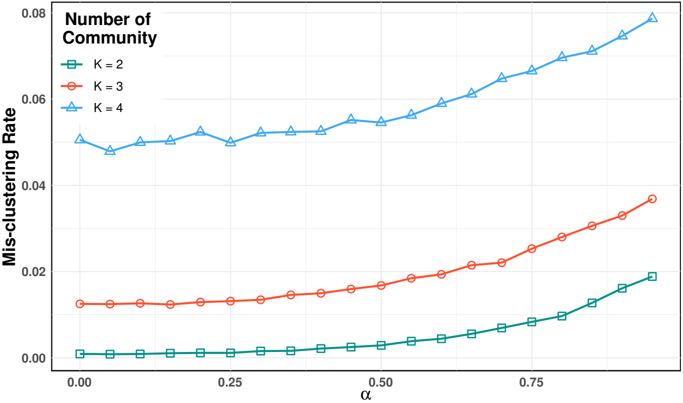
$$B = \nu \left\{ \lambda \boldsymbol{I}_K + (1-\lambda) \mathbf{1}_K \mathbf{1}_K^\top \right\}$$



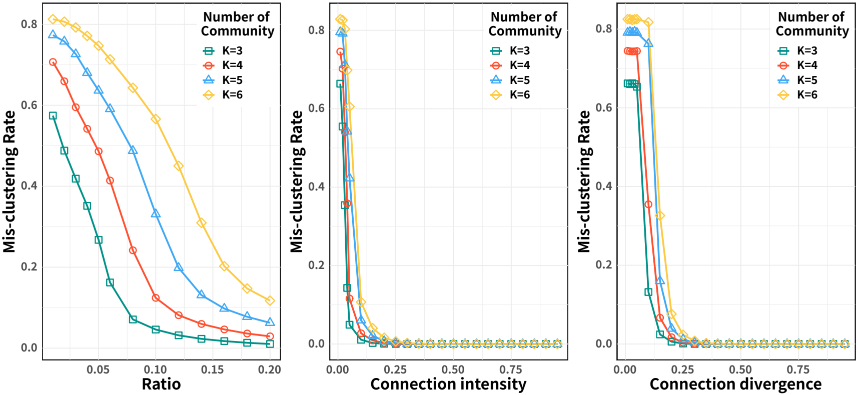
Fudan University

Simulation: Unbalanced Effect

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

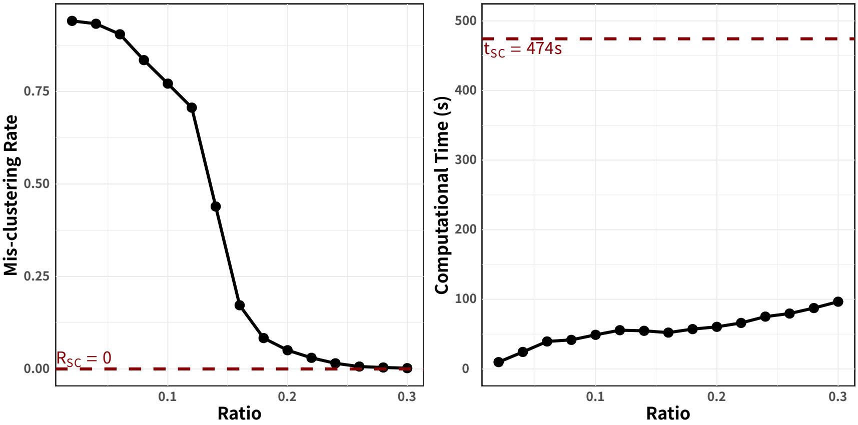


Simulation: DC-SBM

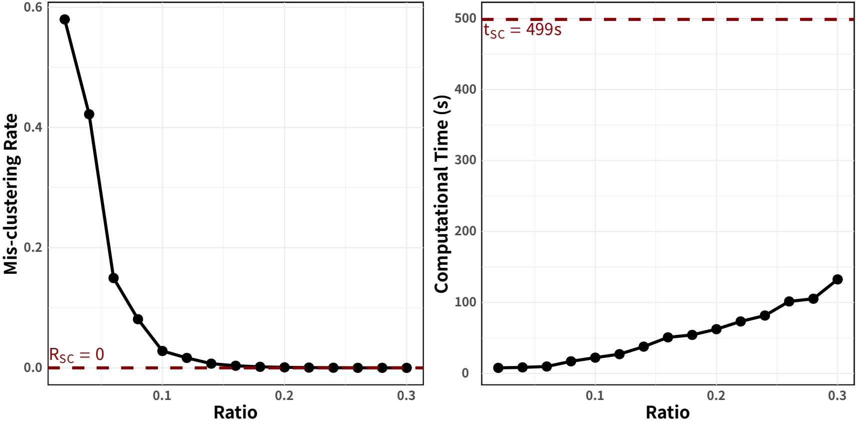


Fudan University

Simulation: Large Scale ($N = 2 \times 10^6$, $K = 20$)

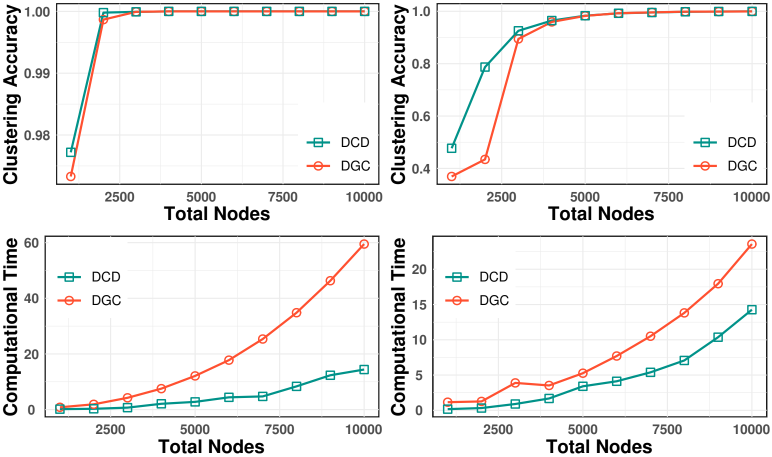


Simulation: Large Scale ($N = 10^7$, $K = 5$)



Fudan University

Simulation: Comparison

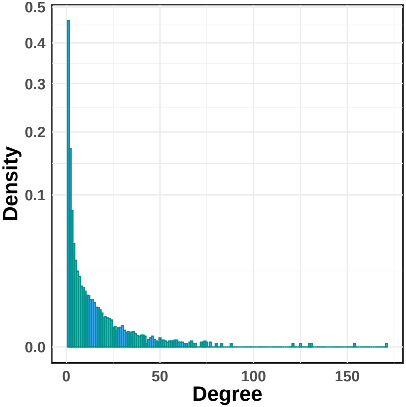


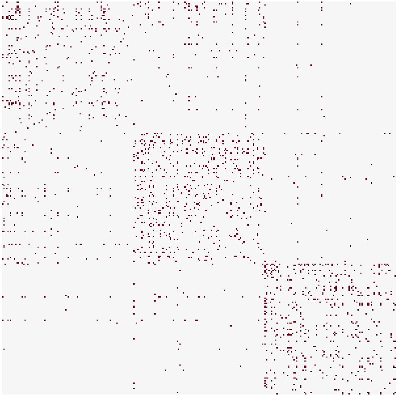
Fudan University

Empirical Study: Pubmed Dataset

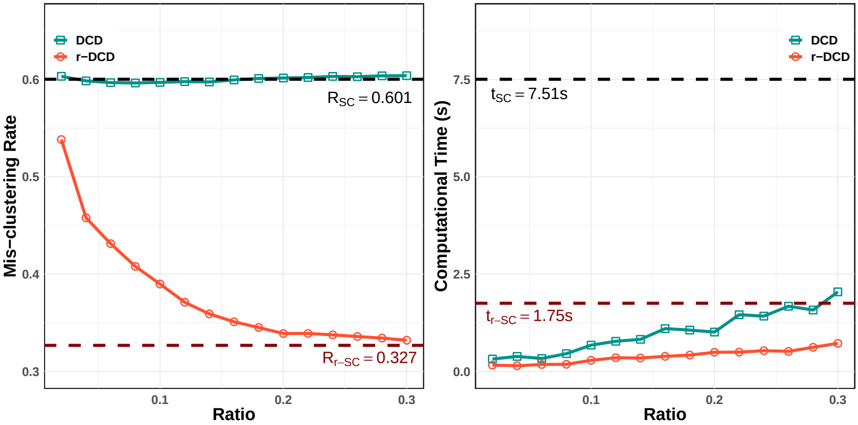
- The Pubmed dataset consists of 19,717 scientific publications
- Each publication is identified as one of the three classes, i.e., Diabetes Mellitus Experimental, Diabetes Mellitus Type 1, Diabetes Mellitus Type 2. $\Rightarrow K = 3$.
- The sizes of the three classes are 4,103, 7,875, and 7,739 respectively.
- The network link is defined using the citation relationships among the publications.
- The resulting network density is 0.028%.

Empirical Study: Pubmed Dataset

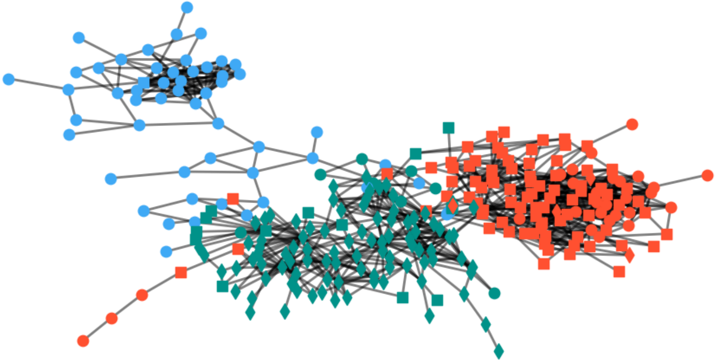




Empirical Study: Pubmed Dataset



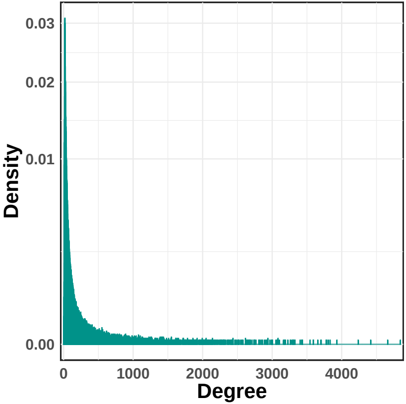
Empirical Study: Pubmed Dataset

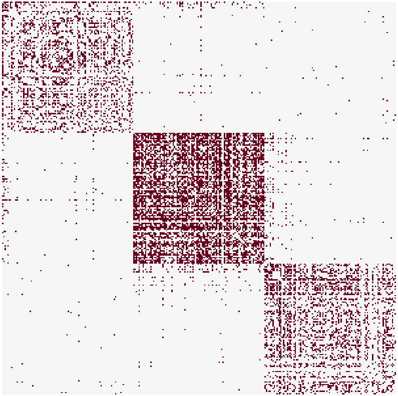


Empirical Study: Pubmed Dataset

- The Yelp is one of the most popular online review platform and the dataset contains 200,193 active users in the network.
- If the ith user is a friend of the jth user, then there is a connection between the two users, i.e., $A_{ij} = 1$
- The resulting network density is 0.031%
- Define the relative density as $\text{RED} = \text{Den}_{\text{between}} \,/\, \text{Den}_{\text{within}}$, where
  - $\text{Den}_{\text{between}} = \sum_{i,j} A_{ij} I(\widehat{g}_i \neq \widehat{g}_j) \,/\, \sum_{i,j} I(\widehat{g}_i \neq \widehat{g}_j)$ is the between-community density
  - $\text{Den}_{\text{within}} = \sum_{i,j} A_{ij} I(\widehat{g}_i = \widehat{g}_j) \,/\, \sum_{i,j} I(\widehat{g}_i = \widehat{g}_j)$ is the within-community density.

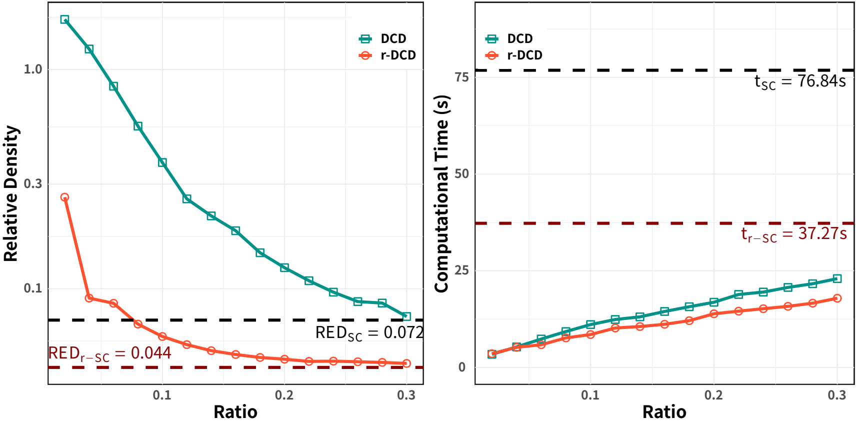
Empirical Study: Yelp Dataset



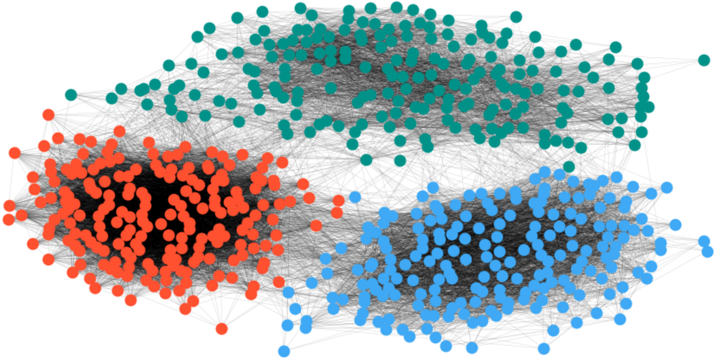


Fudan University
FUDAN UNIVERSITY 1905
復旦大學

Empirical Study: Yelp Dataset



Empirical Study: Yelp Dataset



Conclusion

- We propose a distributed community detection (DCD) algorithm to tackle community detection task in large scale networks.
  - the communication cost is low
  - no further iterative algorithm is used on workers
  - both the computational complexity and the storage requirements are much lower

- **Paper:** https://www.sciencedirect.com/science/article/pii
- **Code:** https://github.com/Ikerlz/dcd
- **Slide:** https://ikerlz.github.io/uploads/DSBM.pdf

Thank
You
Spasibo
welalin
obrigado
Terma Kasih
Grazie
Nirringrazzjak
Multumesc
Kia ora
Matondo
Salamat
Vinaka
Mochchakkeram
Terma Kasih
Merci
Mochchakkeram
Kiitos
Raibh Maith Agat
Chokrane
Asante
Kiitos
Grazie
Maake
Mamana
Kiitos
Dankon
Arigato
obrigado
Dank Je
obrigado
Mochchakkeram
Grazie
Mochchakkeram
Matur Nuwun
Maake
Matondo
Matondo
Dankon
Asante
Dankon
Ua Tsaug Rau Koj
obrigado
Chokrane
Matur Nuwun
Mochchakkeram
Multumesc
Merci
Kiitos
Grazie
Raibh Maith Agat
Cam on ban
Spasibo
Arigato
Merci
Juspaxar
Nirringrazzjak
Salamat
Kia ora
Chokrane
Multumesc
Spasibo
Spasibo