Robust and Computational Feasible Community Detection in the Presence of Arbitrary Outlier Nodes

Author: T. Tony Cai and Xiaodong Li

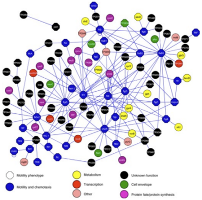
Reporter: Jiaqi Zhang

Nov 3, 2021

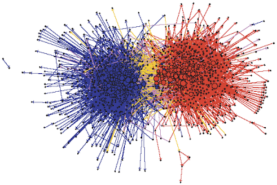
Outline

Overview

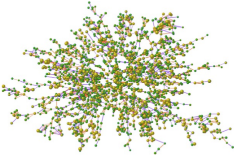
Network data



Protein-Interaction network

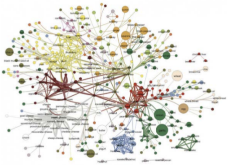


Political Blog network

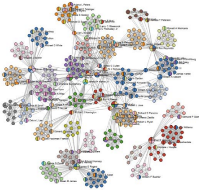


Social network

LinkedIn network



Food flavor network



Professional Network

Figure 1: Network data

A network $G(V, E)$

- vertex/node set $V = [n] = \{1, 2, \dots, n\}$;
- edge set $E \subseteq \{(u, v) : u, v \in V\}$;
- adjacency matrix $\mathbf{A} \in \mathbb{R}^{n \times n}$;
- node degree $d_i = \sum_{j=1}^{n} A_{ij}$;
- undirected, and with no self-loops.

$$\mathbf{A} = \begin{bmatrix} 0 & 1 & 1 & 0 \\ 1 & 0 & 1 & 1 \\ 1 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{bmatrix}$$

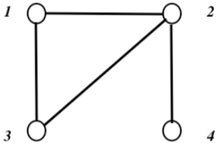


Figure 2: A simple example of a network and the corresponding adjacency matrix $\mathbf{A}$.

Community structure in social network

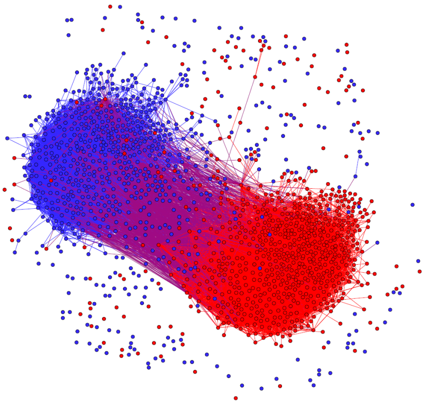


Figure 3: A social network example with community structure.

Stochastic block model

Each of the nodes belongs and only belongs to one and only one of the r nonoverlapping groups.

- labeling function $\phi(j) \in 1, \ldots, r$;
- connectivity matrix $\mathbf{B} \in [0,1]^{r \times r}$;
- $A_{ij} \sim Be\left(B_{\phi(i)\phi(j)}\right)$, independently.

A common assumption:

$$p^- - q^+ := \delta > 0,$$

where $p^- := \min_{1 \le i \le r} B_{ii}$, and $q^+ := \min_{1 \le i < j \le r} B_{ij}$.

Denote the minimum community size by

$$n_{\min} := \min_{1 \le l \le r} \left| \phi^{-1}(l) \right| .$$

The difficulty of the community detection problem is determined by the tuple $(n, r, q^+, p^-, n_{\min})$.

An example of SBM

- $n = 1000$ nodes;
- the first 500 nodes belongs to the same the cluster and the remaining the other;
- connectivity matrix $\mathbf{B} = \begin{bmatrix} 0.17 & 0.11 \\ 0.11 & 0.17 \end{bmatrix}$;
- spectral clustering method applied to both the graph Laplacian and adjacency matrix.

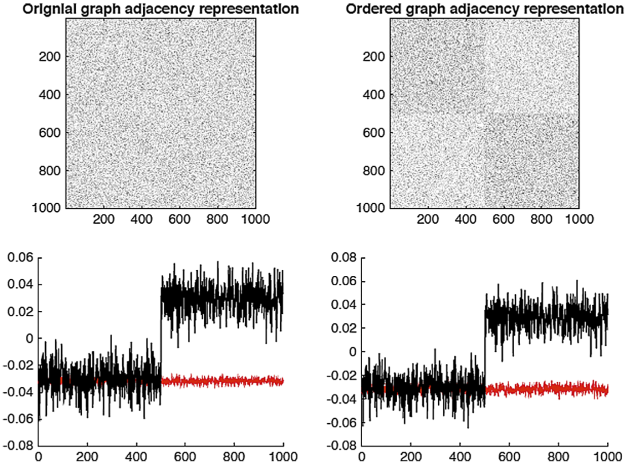


Figure 4: An example of SBM.

Some types of outliers

- Mixed membership;
- Hubs;
- Small clusters;
- Independent neutral nodes;
- ...

Add $m = 30$ outliers to the previous SBM example. Within the outliers, the connectivity is 0.7, and that between each outlier and inlier is from U^2.

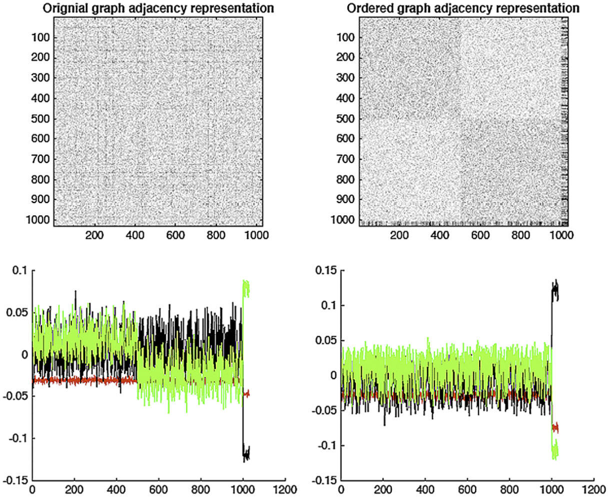


Figure 5: Add outliers to the SBM example.

Overview

Generalized stochastic block model

- totally $N = n + m$ nodes, including n inliers and m outliers;
- labeling function $\phi(i) \in \{1, \dots, r\}$ if $i \in I$, the set of inliers; $\phi(i) = r + 1$ if $i \in O$, the set of outliers;
- the inliers follow a SBM while the connectivity between outliers and inliers and among outliers is arbitrary.

The adjacency matrix of a GSBM can be expressed as

$$\mathbf{A} = \mathbf{P} \begin{bmatrix} \mathbf{K} & \mathbf{Z} \\ \mathbf{Z}^\top & \mathbf{W} \end{bmatrix} \mathbf{P}^\top = \mathbf{P} \begin{bmatrix} \mathbf{K}_{11} & \cdots & \mathbf{K}_{1r} & \mathbf{Z}_1 \\ \vdots & \ddots & \vdots & \vdots \\ \mathbf{K}_{1r}^\top & \cdots & \mathbf{K}_{rr} & \mathbf{Z}_r \\ \mathbf{Z}_1^\top & \cdots & \mathbf{Z}_r^\top & \mathbf{W} \end{bmatrix} \mathbf{P}^\top$$

Semidefinite programming (SDP) of SBM

We derive the convex optimization first from an ordinary SBM model.

- Define a symmetric matrix $\mathbf{X}$ with diagonal entries equal to 1. Let $X_{ij} = 0$, if $\phi(i) \neq \phi(j)$, while $X_{ij} = 1$, if $\phi(i) = \phi(j)$;
- Let $\mathbb{P}(A_{ij} = 1) = q$, if $X_{ij} = 0$; otherwise, let $\mathbb{P}(A_{ij} = 1) = p$.

Then we have

$$\log \mathbb{P}(A_{ij} = 1|X_{ij}) = X_{ij} \log p + (1 - X_{ij}) \log q,$$

and

$$\log \mathbb{P}(A_{ij} = 0|X_{ij}) = X_{ij} \log(1 - p) + (1 - X_{ij}) \log(1 - q),$$

The log-likelihood function

$$\ell(\mathbf{A}|\mathbf{X}) = \sum_{1\le i<j\le n} \Big\{ A_{ij}\big[X_{ij}\log p + (1-X_{ij})\log q\big] + (1-A_{ij})\big[X_{ij}\log(1-p) + (1-X_{ij})\log(1-q)\big]\Big\}$$

Maximization of the log-likelihood function is equivalent

$$\max_{\mathbf{X}} \big\langle \mathbf{X}, (1-\lambda)\mathbf{A} - \lambda(\mathbf{J}_N - \mathbf{I}_N - \mathbf{A})\big\rangle,$$

the constraint of $\mathbf{X}$ is that it must have the following form:

$$\mathbf{X} = \mathbf{P}\begin{bmatrix} \mathbf{J}_{l_1} & & \\ & \ddots & \\ & & \mathbf{J}_{l_r} \end{bmatrix}\mathbf{P}^\top$$

Relaxed form of the constraint:

- $\mathbf{X}$ is positive semidefinite;
- all its entries are between 0 and 1;
- it is of rank r, far from full-rank.

The relaxed maximum likelihood method becomes

$$\max_{\widetilde{\mathbf{X}}} \langle \widetilde{\mathbf{X}}, (1-\lambda)\mathbf{A} - \lambda(\mathbf{J}_N - \mathbf{I}_N - \mathbf{A}) \rangle$$

$$\text{subject to} \qquad \widetilde{\mathbf{X}} \succeq 0,$$

$$0 \leq \widetilde{X}_{ij} \leq 1, \text{ for } 1 \leq i,j \leq N.$$

SDP of GSBM

We add an additional term in the objective function to penalized the trace

$$\begin{aligned} &\min_{\widetilde{\mathbf{X}}} \langle \widetilde{\mathbf{X}}, \mathbf{E} \rangle \\ &\text{subject to} \qquad \widetilde{\mathbf{X}} \succeq 0, \\ &\qquad\qquad\qquad 0 \leq \widetilde{X}_{ij} \leq 1, \ \text{for } 1 \leq i, j \leq N. \end{aligned} \tag{2.3}$$

where $\mathbf{E} := \alpha \mathbf{I}_N - (1-\lambda)\mathbf{A} + \lambda(\mathbf{J}_N - \mathbf{I}_N - \mathbf{A})$

Recall that $\mathbf{X}$ is a symmetric matrix where $X_{ij} = 0$, if $\phi(i) \neq \phi(j)$, while $X_{ij} = 1$, if $\phi(i) = \phi(j)$, which reveals the clustering structure of the nodes.

- The relaxed form $\widetilde{\mathbf{X}}$ cannot directly show us the clustering structure;
- the second step is conducting k-means clustering algorithm to solve for assigning function $\hat{\phi}$.

Computation

The optimization problem (2.3) can be rewritten as

$$\begin{aligned} \min_{\mathbf{Y},\mathbf{Z}} \quad & \iota(\mathbf{Y} \succeq \mathbf{0}) + \iota\,(\mathbf{0} \leq \mathbf{Z} \leq \mathbf{J}_N) + \langle \mathbf{Y}, \mathbf{E} \rangle, \\ \text{subject to} \quad & \mathbf{Y} = \mathbf{Z}. \end{aligned}$$

Note that the objective function is convex. Define the scaled augmented Lagrangian of this optimization problem as

$$L_\rho(\mathbf{Y}, \mathbf{Z}; \mathbf{\Lambda}) := \iota(\mathbf{Y} \succeq \mathbf{0}) + \iota\,(\mathbf{0} \leq \mathbf{Z} \leq \mathbf{J}_N) + \langle \mathbf{Y}, \mathbf{E} \rangle + \frac{\rho}{2} \|\mathbf{Y} - \mathbf{Z} + \mathbf{\Lambda}\|_F^2$$

To minimize $L_\rho(\mathbf{Y}, \mathbf{Z}; \mathbf{\Lambda})$, the ADMM algorithm tells us to alternately update $\mathbf{Y}$, $\mathbf{Z}$, and $\mathbf{\Lambda}$, with the other two fixed.

Update $\mathbf{Y}$

Minimizing $L_\rho(\mathbf{Y}, \mathbf{Z}; \mathbf{\Lambda})$ with respect to $\mathbf{Y}$ is equivalent to minimizing

$$\iota(\mathbf{Y} \succeq \mathbf{0}) + \frac{\rho}{2}\|\mathbf{Y} - \mathbf{Z} + \mathbf{\Lambda} + \frac{\mathbf{E}}{\rho}\|_F^2.$$

For any symmetric matrix $\mathbf{X}$ with eigendecomppositon $\mathbf{X} = \mathbf{V}\mathbf{\Sigma}\mathbf{V}^\top$, define $\mathbf{X}_+ := \mathbf{V}\mathbf{\Sigma}_+\mathbf{V}^\top$. Then the solution to $\mathbf{Y}$ is

$$\underset{\mathbf{Y}}{\operatorname{argmin}} L_\rho(\mathbf{Y}, \mathbf{Z}; \mathbf{\Lambda}) = \left(\mathbf{Z} - \mathbf{\Lambda} - \frac{\mathbf{E}}{\rho}\right)_+.$$

Update $\mathbf{Z}$

Minimizing $L_\rho(\mathbf{Y}, \mathbf{Z}; \mathbf{\Lambda})$ with respect to $\mathbf{Z}$ is equivalent to minimizing

$$\iota\left(\mathbf{0} \leq \mathbf{Z} \leq \mathbf{J}_N\right) + \frac{\rho}{2}\|\mathbf{Y} - \mathbf{Z} + \mathbf{\Lambda}\|_F^2.$$

There still exist a closed-form solution

$$\underset{\mathbf{Z}}{\text{argmin}} L_\rho(\mathbf{Y}, \mathbf{Z}; \mathbf{\Lambda}) := \min\left(\max(\mathbf{Y} + \mathbf{\Lambda}, \mathbf{0}), \mathbf{J}_N\right)$$

Update $\boldsymbol{\Lambda}$ and the remainings about computation

According to the ADMM algorithm, the dual variable $\boldsymbol{\Lambda}$ is updated to $\boldsymbol{\Lambda} + (\mathbf{Y} - \mathbf{Z})$.

- The parameters are initialized as $\mathbf{Z}_0 = \mathbf{0}$ and $\boldsymbol{\Lambda}_0 = \mathbf{0}$;
- The ‘step size‘ is set to $\rho = 1$;
- The maximum number of iterations is set to 100.

Theoretical results

Theorem 3.1.

Let $\mathbf{A}$ be the adjacency matrix of the semi-random graph under the GSBM. Let $\widehat{\mathbf{X}}$ be a solution to the semidefinite program (2.3). Suppose that $p^- \geq C\frac{\log n}{n_{\min}}$, $\alpha \geq 3m$ and

$$\delta > C\left(\sqrt{\frac{p^-\log n}{n_{\min}}} + \frac{\alpha}{n_{\min}} + \frac{\sqrt{nq^+}}{n_{\min}} + \frac{m\sqrt{r}}{n_{\min}} + \frac{nmp^-}{(\alpha-2m)n_{\min}}\right)$$

for some sufficiently large numerical constant C, and the tuning parameter λ satisfies

$$q^+ + \frac{4}{\delta} < \lambda < p^- - \frac{4}{\delta}.$$

Theorem 3.1. (continued)

Then with probability at least $1 - \frac{1}{n} - \frac{n^2}{2r} - \frac{cr}{n_{\min}^4}$ for some constant c, $\widehat{\mathbf{X}}$ must be of the form

$$\widehat{\mathbf{X}} = \mathbf{P} \begin{bmatrix} \mathbf{J}_{l_1} & & & \widehat{\mathbf{Z}}_1 \\ & \ddots & & \vdots \\ & & \mathbf{J}_{l_r} & \widehat{\mathbf{Z}}_r \\ \widehat{\mathbf{Z}}_1^\top & \cdots & \widehat{\mathbf{Z}}_r^\top & \widehat{\mathbf{W}} \end{bmatrix} \mathbf{P}^\top$$

Theorem 3.2.

Suppose the assumption in Theorem 3.1 hold as well as $m < \frac{2r+4}{r_{\min}}$. Then, with high probability, the misclassification rate among the inlier nodes is no more than $\frac{(2r+3)m}{n}$.

Overview

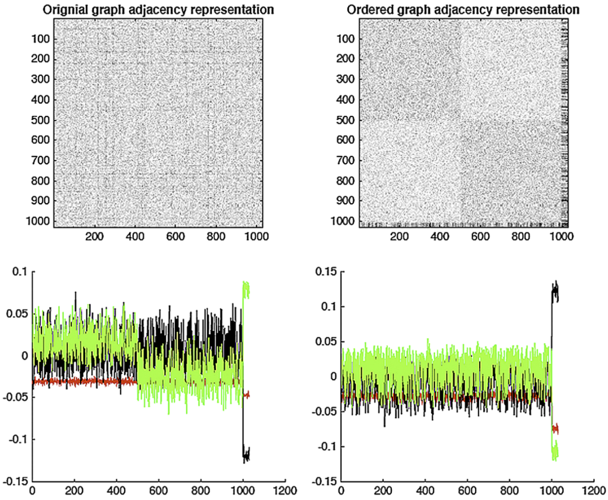
Simulations

- $n = 1000$ nodes;
- the first 500 nodes belongs to the same the cluster and the remaining the other;
- connectivity matrix $\mathbf{B} = \begin{bmatrix} 0.17 & 0.11 \\ 0.11 & 0.17 \end{bmatrix}$;
- spectral clustering method applied to both the graph Laplacian and adjacency matrix.

Add $m = 30$ outliers to the previous SBM example. Within the outliers, the connectivity is 0.7, and that between each outlier and inlier is from U^2.

The nodes with degrees above the 80th percentile or below the 20th percentile are eliminated from the graph, and λ is chosen as the mean density of the subgraph of the remaining nodes.

- 10 independent graphical date sets, the average misclassification rate is 0.0063;
- while those of spectral clustering on the graph Laplacians and adjacency matrices are 0.4792 and 0.5000;
- applying spectral clustering with $k = 3$ gets misclassification rates 0.3083 and 0.4730.


Orignial graph adjacency representation
Ordered graph adjacency representation
100
200
300
400
500
600
700
800
900
1000
200
400
600
800
1000
0.1
0.05
0
-0.05
-0.1
-0.15
0.15
0
200
400
600
800
1000
1200

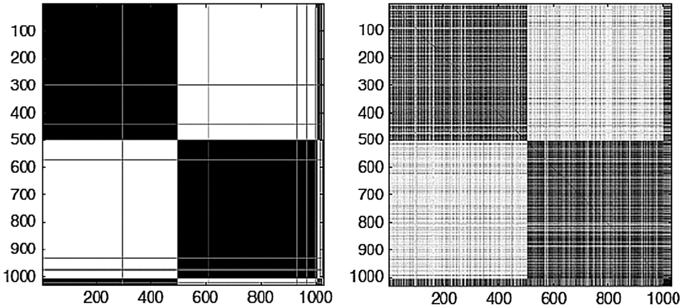


Figure 6: Results of the proposed method in one replicate.

Sensitivity to the choice of λ

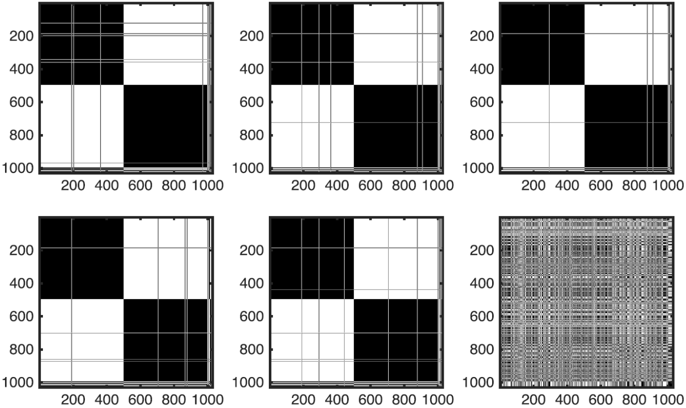


Figure 7: Sensitivity to λ.

Sensitivity to within connectivity p

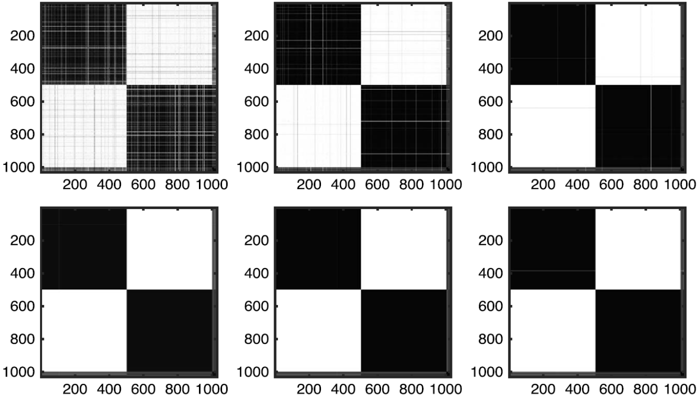


Figure 8: Sensitivity to p.

Real data analysis

Political blogs network data:

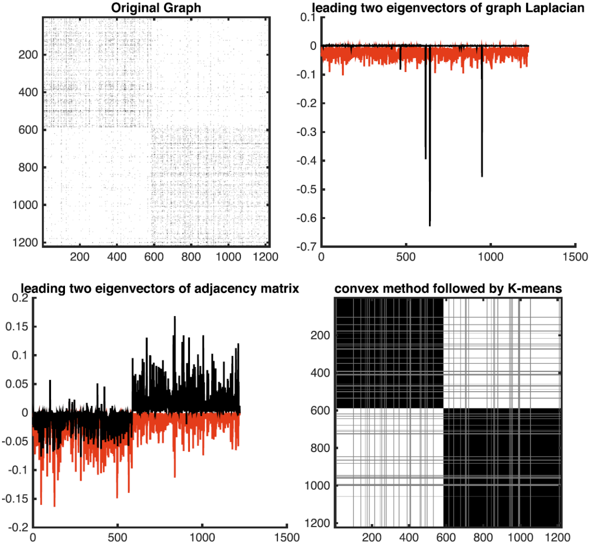
- political blogs connected with hyperlinks;
- 1222 nodes and 16,714 edges;
- manually labeled in previous study.

Use a modified version of (2.3) by letting

$$\mathbf{E} := -\left(\mathbf{I}_N - \mathbf{D}\right)^{1/2}\mathbf{A}\left(\mathbf{I}_N - \mathbf{D}\right)^{1/2} + \mathbf{D}^{1/2}\left(\mathbf{J}_N - \mathbf{I}_N - \mathbf{A}\right)\mathbf{D}^{1/2}.$$

The misclassification rate is 63/1222. While ordinary spectral clustering fails on this dataset and the misclassification rate of different modified versions of spectral clustering is at least 0.2.

Sensitivity to within connectivity p



Overview

Discussion

- The GSBM for robust community detection is proposed with strong theoretical guarantees in the performance in finding the clustering structure;
- the assumption $\delta = p^{-} - q^{+}$ is too strong for some real-world applications;
- degree-corrected SBM;
- choice of α.

Thank you!