Achieving Optimal Misclassification Proportion in Stochastic Block Model - Review

Author: Chao Gao, Zongming Ma, Anderson Y.Zhang, Harrison H.Zhou

Jiaxin Hu

04/20/2020

ABSTRACT

This paper proposes a polynomial time two-stage method to detect the community in network matrix data with stochastic block model(SBM). One stage is to initialize the partition with the provided new greedy clustering method or any other methods with weak convergence condition. The other stage is to refining the node-wise partition via penalized local maximum likelihood estimation. Statistical guarantees for refinement scheme and initialization are given. The paper reviews related literature closely, which provides a good view of network community detection.

1 PROBLEM AND METHODOLOGY

1.1 Problem Formulation

Let $A \in \{0,1\}^{n\times n}$ be the symmetric adjacency matrix of an undirected random graph generated according to an SBM with k communities, where $A_{ii} = 0, i \in [n]$ and $A_{uv} = A_{vu} \sim_{i.i.d} Ber(P_{uv}), \forall u > v \in [n]$. Assume $\mathbb{E}(A_{uv}) = P_{uv} = B_{\sigma(u)\sigma(v)}$, where $B \in [0,1]^{k\times k}$ is the connectivity matrix and $\sigma : [n] \to [k]$ is the label function. If the m-th and n-th node belongs to i-th and j-th community respectively, then $\sigma(m) = i, \sigma(v) = j$ and there is a edge connectivity between u and v with probability B_{ij}. Define the error measure as misclassification proportion:

How does this I compare to our earlier paper?
Goal: find hat \sigma

$$l(\hat{\sigma}, \sigma) = \min_{\pi\in S_k} \frac{1}{n} \sum_{u\in[n]} \mathbb{I}\{\hat{\sigma}(u) \neq \pi(\sigma(u))\},$$

where S_k stands for the symmetric group on $[k]$ consisting of all permutations of $[k]$.

The goal of the paper is to find an estimate of label function $\hat{\sigma}$ that achieves the optimal misclassification proportion $l(\hat{\sigma}, \sigma)$ under weak regularity conditions. Obtaining a good estimate for B is not the main goal of the paper unlike previous TBM and TBSM paper.

1.2 Parameter Space & Assumptions

One widely studied parameter space for (B, σ) of SBM is:

$$\Theta_0(n,k,a,b,\beta) = \Big\{(B,\sigma)|\ \sigma : [n] \to [k], n_i = \sum_{u\in[n]} \mathbb{I}\{\sigma(u) = i\} \in [\frac{n}{\beta k} - 1, \frac{\beta n}{k} + 1], \forall i \in [k],$$
$$B = [\![B_{ij}]\!] \in [0,1]^{k\times k}, B_{ii} = \frac{a}{n} \text{ for } \forall i, B_{ij} = \frac{b}{n} \text{ for } \forall i \neq j\Big\},$$

where $\beta \geq 1$ is an absolute constant which controls the range of community size. This parameter space can be re-written in following assumptions:

- **A1.** The sizes of communities are comparable. For arbitrary community i, its size n_i is not smaller than $\frac{n}{\beta k} - 1$ and not larger than $\frac{\beta n}{k} + 1$. When the special case $\beta = 1$, all communities are of nearly equal size.
- **A2.** Signal difference only exists between within-community connection and between-community connection. All the within-community connection probabilities are $\frac{a}{n}$ and all the between-community connection probabilities are $\frac{b}{n}$, which is a special kind of irreducible setting.

Relaxing the equal within and equal between connection, introduce the following larger parameter space:

$$\Theta(n,k,a,b,\lambda,\beta,\alpha) = \Big\{(B,\sigma)|\ \sigma:[n]\to[k], n_i = \sum_{u\in[n]} \mathbb{I}\{\sigma(u)=i\} \in [\frac{n}{\beta k}-1, \frac{\beta n}{k}+1], \forall i\in[k],$$
$$B = B^T = [\![B_{ij}]\!] \in [0,1]^{k\times k}, \frac{b}{\alpha n} \le \frac{1}{k(k-1)}\sum_{i\neq j} B_{ij} \le \max_{i\neq j} B_{ij} = \frac{b}{n},$$
$$\frac{a}{n} = \min_i B_{ii} \le \max_i B_{ii} \le \frac{\alpha a}{n}, \lambda_k(P) \le \lambda \text{ with } P = [\![P_{uv}]\!] = [\![B_{\sigma(u)\sigma(v)}]\!]\Big\},$$

where β, α are absolute constants which controls the range of community size and the range of signal. This parameter can also be re-written in following assumptions:

- **A'1.** The sizes of communities are comparable. Same as **A1**.
- **A'2.** The signals of within and between community connections are comparable. For within community, $\frac{a}{n} = \min_i B_{ii} \ge \frac{1}{\alpha}\max_i B_{ii} \ge \frac{1}{\alpha}\min_i B_{ii}$. For between community, $\frac{1}{k(k-1)}\sum_{i\neq j} B_{ij} = \bar{B}_{ij} \ge \frac{b}{\alpha n} = \frac{1}{\alpha}\max_{i\neq j} B_{ij}$. The constant α can control the range of signal in these two kinds of connections.
- **A'3.** The k-th eigenvalue of probability matrix P is lower bounded by λ. Since there are k communities, the rank of P should be at least k.

Additionally, authors assume $0 < \frac{b}{n} < \frac{a}{n} \le 1-\epsilon$ for some constant $\epsilon \in (0,1)$ throughout the paper. That means $0 < \max_{i\neq j} B_{ij} < \min_i B_{ii} < 1-\epsilon$. To ensure $\Theta_0(n,k,a,b,\beta) \subset \Theta(n,k,a,b,\lambda,\beta,\alpha)$, the paper also requires $\lambda \le \frac{a-b}{2\beta k}$ throughout the paper. The theoretical results have slightly different for these two parameter sets.

1.3 Methodology

For reading convenience, this section is written in the implementation order.

1.3.1 Initialization

Though the theoretical results show that refinement stage only require the weak consistency condition of initialization method, authors here introduce a greedy spectral clustering algorithm.

First, consider the trimming unnormalized spectral clustering (USC) with adjacency matrix A. Let $d_u = \sum_{v\in[n]} A_{uv}$ be the degree of u-th node. Obtain the trimmed adjacency matrix $T_\tau(A)$ be replacing $A_{u\cdot}$ and $A_{\cdot u}$ to 0 if $d_u \ge \tau$.
Second, consider the trimming normalized spectral clustering (NSC) with graph Laplacian $L(A)$, where $[\![L(A)_{uv}]\!] = d_u^{-1/2} d_v^{-1/2} A_{uv}$. Obtain the trimmed Laplacian $L(A_\tau)$ by replacing A to $A_\tau = A + \frac{\tau}{n}\mathbf{1}\mathbf{1}^T$.

The sketch of the greedy clustering method (Algorithm 1) is below:

Algorithm 1 Initialization

Input: Data matrix $\hat{U} = T_\tau(A)$ or $L(A_\tau)$, # of communities k, critical radius $r = \mu\sqrt{\frac{k}{n}}$ with some constant μ.

Output: Community assignment $\hat{\sigma}$.

1: $S = [n]$;
2: **for** $i = 1$ to k **do**
3: $t_i = \arg\max_{u\in S} \sum_{v\in S} \mathbb{I}\{\|\hat{U}_{v\cdot} - \hat{U}_{u\cdot}\| \le r\}$;
4: $\hat{\mathcal{C}}_i = \{v \in S : \|\hat{U}_{v\cdot} - \hat{U}_{t_i\cdot}\| \le r\}$;
5: Label $\hat{\sigma}(u) = i$ for all $u \in \hat{\mathcal{C}}_i$;
6: $S \leftarrow S \setminus \hat{\mathcal{C}}_i$.
7: **end for**
8: If $S \neq \varnothing$, then for any $u \in S$, $\hat{\sigma}(u) = \arg\min_{i\in[k]} \frac{1}{|\hat{\mathcal{C}}_i|} \sum_{v\in\hat{\mathcal{C}}_i} \|\hat{U}_{u\cdot} - \hat{U}_{v\cdot}\|$.

1.4 Refinement

1.4.1 Heuristics

To minimized the misclassification proportion $l(\hat{\sigma}, \sigma)$, MLE is always a nature choice. Under the parameter space $\Theta_0(n, k, a, b, 1)$ with equal community size, the MLE for σ is:

$$\hat{\sigma} = \arg\max_{\sigma} \sum_{u<v} A_{uv}\mathbb{I}\{\sigma(u) = \sigma(v)\},$$

which is a combinatorial optimization problem and is computationally intractable.

However, we can easily write the close form solution for the node-wise optimization. If we know the values of $\{\sigma(u)\}_{u=2}^n$, then for $\sigma(1)$:

$$\hat{\sigma}(1) = \arg\max_{i\in[k]} \sum_{v\neq 1:\sigma(v)=i} A_{1v}.$$

The estimate can be interpreted as the first node should belong to the community where the first node has the most neighbours. In practice, we do know the labels in advance. However, we can estimate the labels for $(n-1)$ nodes with initialization algorithm. We note σ^0 as initialization algorithm for the $A_{-u} \in \{0,1\}^{(n-1)\times(n-1)}$, where A_{-u} is a submatrix of A with its u-th row and columns are removed. Repeat this node-wise optimization among every node, we get the refinement scheme for community detection (Algorithm 2).

1.4.2 Detailed algorithm

See below Algorithm 2.

2 THEORETICAL RESULTS

Definition 2.1. In previous work(Zhang and Zhou, 2015), the minimax risk is governed by:

$$I^* = -2\log\left(\sqrt{\frac{a}{n}}\sqrt{\frac{b}{n}} + \sqrt{1-\frac{a}{n}}\sqrt{1-\frac{b}{n}}\right).$$

With **A'3**, we can show $I^* \asymp \frac{(a-b)^2}{na}$. When $\frac{a}{n} = o(1)$, $I^* = (2+o(1))H^2\left(\text{Bern}\left(\frac{a}{n}\right), \text{Bern}\left(\frac{b}{n}\right)\right)$, where $H^2(P,Q) = \frac{1}{2}\int(\sqrt{\mathrm{d}P} - \sqrt{\mathrm{d}Q})^2$ is the squared Hellinger distance between P, Q.

Connection between this signal level vs. our signal level (in previous TBM…)

Algorithm 2 Refinement

Input: Adjacency matrix A, # of communities k, initialization algorithm σ^0.
Output: Community assignment $\hat{\sigma}$.

1: **for** $u = 1$ to n **do**
2: Apply σ^0 to A_{-u}, obtain $\sigma_u^0(v), \forall v \neq u$ and let $\sigma_u^0(u) = 0$
3: Define $\tilde{\mathcal{C}}_i^u = \{v : \sigma_u^0(v) = i\}$, let $\tilde{\mathcal{E}}_i^u = \sum_{x<y, \sigma_u^0(x), \sigma_u^0(y) \in \tilde{\mathcal{C}}_i^u} A_{xy}$ and $\tilde{\mathcal{E}}_{ij}^u = \sum_{\sigma_u^0(x) \in \tilde{\mathcal{C}}_i^u, \sigma_u^0(y) \in \tilde{\mathcal{C}}_j^u} A_{xy}$.
4: Define

$$\widehat{B}_{ii}^u = \frac{|\widetilde{\mathcal{E}}_i^u|}{\frac{1}{2}|\widetilde{\mathcal{C}}_i^u|\left(|\widetilde{\mathcal{C}}_i^u| - 1\right)}, \quad \widehat{B}_{ij}^u = \frac{|\widetilde{\mathcal{E}}_{ij}^u|}{|\widetilde{\mathcal{C}}_i^u||\widetilde{\mathcal{C}}_j^u|}, \quad \forall i \neq j \in [k]$$

5: Define $\hat{\sigma}_u(v) = \sigma_u^0(v)$ for all $i \neq j$, and

$$\widehat{\sigma}_u(u) = \underset{l \in [k]}{\operatorname{argmax}} \sum_{\sigma_u^0(v) = l} A_{uv} - \rho_u \sum_{v \in [n]} \mathbf{1}_{\{\sigma_u^0(v) = l\}},$$

where

$$t_u = \frac{1}{2} \log \frac{\widehat{a}_u\left(1 - \widehat{b}_u/n\right)}{\widehat{b}_u\left(1 - \widehat{a}_u/n\right)}, \rho_u = -\frac{1}{2t_u} \log\left(\frac{\frac{\widehat{a}_u}{n} e^{-t_u} + 1 - \frac{\widehat{a}_u}{n}}{\frac{\widehat{b}_u}{n} e^{t_u} + 1 - \frac{\widehat{b}_u}{n}}\right).$$

6: **end for**
7: **Consensus:** Define $\hat{\sigma}(1) = \hat{\sigma}_1(1)$, for $u = 2, ..., n$, define

$$\widehat{\sigma}(u) = \underset{l \in [k]}{\operatorname{argmax}} \left|\{v : \widehat{\sigma}_1(v) = l\} \cap \{v : \widehat{\sigma}_u(v) = \widehat{\sigma}_u(u)\}\right|.$$

Theorem 1 (Zhang and Zhou(2015)). *When* $\frac{(a-b)^2}{ak \log k} \to \infty$, *we have*

$$\inf_{\widetilde{\sigma}} \sup_{(B,\sigma) \in \Theta} \ell(\widetilde{\sigma}, \sigma) = \begin{cases} \exp\left(-(1+\eta)\frac{nI^*}{2}\right), & k = 2 \\ \exp\left(-(1+\eta)\frac{nI^*}{\beta k}\right), & k \geq 3 \end{cases}$$

for both Θ_0 *and* Θ *with* $\lambda \leq \frac{a-b}{2\beta k}$ *and and* $\beta \in [1, \sqrt{5/3}]$, *where* $\eta = \eta_n \to 0$ *as* $n \to \infty$.

Condition 2.1. There exists constants $C_0, \delta > 0$ and positive sequence $\gamma = \gamma_n$, s.t.

$$\inf_{(B,\sigma) \in \Theta} \min_{u \in [n]} \mathbb{P}_{B,\sigma}\left\{\ell\left(\sigma, \sigma_u^0\right) \leq \gamma\right\} \geq 1 - C_0 n^{-(1+\delta)}$$

for some Θ.

Theorem 2 (Performance for refinement). *Suppose as* $n \to \infty, \frac{(a-b)^2}{ak \log k} \to \infty, a \asymp b$ *and Condition 1 satisfies for* ***1:*** $\gamma = o\left(\frac{1}{k \log k}\right)$ *and* $\Theta = \Theta_0(n, k, a, b, \beta)$ *or* ***2:*** $\gamma = o\left(\frac{1}{k \log k}\right)$ *and* $o\left(\frac{a-b}{ak}\right)$ *and* $\Theta = \Theta(n, k, a, b, \lambda, \beta, \alpha)$. *Then there is a sequence* $\eta \to 0$ *s.t.*

How does this rate compare to our TBM?
their model is a special TBM (e.g. order-two, two-blocks equal p within-block…..)

$$\sup_{(B,\sigma) \in \Theta} \mathbb{P}_{B,\sigma}\left\{\ell(\sigma, \widehat{\sigma}) \geq \exp\left(-(1-\eta)\frac{nI^*}{2}\right)\right\} \to 0, \quad \text{if } k = 2$$

$$\sup_{(B,\sigma) \in \Theta} \mathbb{P}_{B,\sigma}\left\{\ell(\sigma, \widehat{\sigma}) \geq \exp\left(-(1-\eta)\frac{nI^*}{\beta k}\right)\right\} \to 0, \quad \text{if } k \geq 3$$

1. Is the error metric comparable? connection between l(sigma,\hat sigma) and our MCR?
2. the actual convergence bound…

assumption… with v.h.p
MCR <…

Theorem 3 (Performance for initialization(USC)). *Assume $e \le a \le C_1 b$ for some constant $C_1 > 0$ and $\frac{ka}{\lambda_k^2} \le c$ for sufficient small $c \in (0,1)$. Consider $USC(\tau)$ with sufficient small μ and $\tau = C_2\bar{d}$, where $\bar{d} = 1/n \sum_{u\in[n]} d_u$ and sufficient large constant $C_2 > 0$. For any C', there exists some $C > 0$ depends on C_1, C_2, C', μ s.t.*

$$\ell(\widehat{\sigma}, \sigma) \le C \frac{a}{\lambda_k^2}$$

with probability at least $1 - n^{-C'}$. If k is fixed, the same conclusion holds without $a \le C_1 b$.

Theorem 4 (Performance for initialization(NSC)). *Assume $e \le a \le C_1 b$ for some constant $C_1 > 0$ and $\frac{ka\log(a)}{\lambda_k^2} \le c$ for sufficient small $c \in (0,1)$. Consider $USC(\tau)$ with sufficient small μ and $\tau = C_2\bar{d}$, where $\bar{d} = 1/n \sum_{u\in[n]} d_u$ and sufficient large constant $C_2 > 0$. For any C', there exists some $C > 0$ depends on C_1, C_2, C', μ s.t.*

$$\ell(\widehat{\sigma}, \sigma) \le C \frac{a\log(a)}{\lambda_k^2}$$

with probability at least $1 - n^{-C'}$. If k is fixed, the same conclusion holds without $a \le C_1 b$.

3 TO DO LIST & QUESTIONS

3.1 To do list

- Go through the proof and technical details.
- Skim the related papers mentioned in the paper.

3.2 Questions to be answered

- The two kinds of parameter spaces both use α to control the signal range of both within-community and between-community. Will this setting lead to some restriction? Why not we set α_1 for within-community and α_2 for between-community? give a conjecture. how the new theorem will read
- What's the meaning of ρ_u in refinement? Why ρ_u can not only deal with different community sizes but also different values on diagonal and non-diagonal of B?
- What's the key point that can allow $k \to \infty$? TBM: K fixed or equivalently, K = O(1). Lei et al: K = O(1) (bounded)
- Where is the α in Theorem 2 when $k \ge 3$?
- Can we extend it to high-order community detection?
- If there is a sparse network, will anything change?