Achieving Optimal Misclassification Proportion in Stochastic Block Model - Review

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

Jiaxin Hu

First edition: 04/19/2020; Final edition: 04/21/2020

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 FORMULATION & MODEL

1.1 *Model*

Let $A \in \{0,1\}^{n\times n}$ be the symmetric adjacency matrix where $A_{ii} = 0, i \in [n]$ and $A_{uv} = A_{vu}, \forall u > v \in [n]$. Assume there are k communities among n nodes. The stochastic block model can be written as below:

$$A_{uv} \sim_{i.i.d} Ber(P_{uv}), \forall u \neq v; \quad \mathbb{E}(A_{uv}) = P_{uv} = B_{\sigma(u)\sigma(v)},$$

where $B \in [0,1]^{k\times k}$ is the connectivity matrix and $\sigma : [n] \to [k]$ is the label function. In matrix form, the model can also be presented as below:

$$\mathbb{E}(A) = P = H^T BH; \quad A = H^T BH + \mathcal{E},$$

where $H \in \{0,1\}^{k\times n}$ is the membership matrix corresponding to the partition σ and $\mathcal{E}$ is a sub-Gaussian noise matrix. This model can be considered as the order-2 case of TBM(M.Wang 2019) and the single layer case of TSBM(J.Lei 2020).

Two parameter spaces are discussed in the paper:

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

where $\beta \geq 1$ is an absolute constant which controls the range of community size. Relaxing the equal within and equal between connection, we have a 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 α is absolute constant which controls the range of signal.

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.

This main goal for this paper is to find an estimate of partition $\hat{\sigma}$ to achieve the optimal minimax misclassification proportion established in Zhang and Zhou(2015).

1.2 *Discussion of Loss measure*

First, the loss measure in Gao's and Zhang's paper is stated below:

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

where S_k stands for the symmetric group on $[k]$ consisting of all permutations of $[k]$ and $\pi(\sigma(\cdot))$ refers a permuted σ. Recall the misclassification proportion in TBM and TBSM:

$$MCR(\hat{M}_k, M_{k,true}) = \max_{r\in[R_k], a\neq a'\in[R_k]} \min\{D^{(k)}_{a,r}, D^{(k)}_{a',r}\}, \text{ where } D^{(k)}_{r,r'} = \frac{1}{d_k}\sum_{i=1}^{d_k} \mathbb{I}[m^{(k)}_{i,r} = \hat{m}^{(k)}_{i,r'} = 1], r,r' \in [R_k].$$

To show the relationship between $l(\hat{\sigma},\sigma)$ and $MCR(\hat{M}_k, M_{k,true})$, define:

$$D_{r,r'} = \sum_{u\in[n]} \mathbb{I}\{\hat{H}_{ur} = H_{ur'} = 1\};\quad D^{\pi}_{r,r'} = \sum_{u\in[n]} \mathbb{I}\{\hat{H}_{ur} = \pi(H_{ur'}) = 1\}$$

For here is an order-2 symmetric case, there is only one confusion matrix D for TBM. Therefore, we can show

$$MCR(\hat{H}, H) = \frac{1}{n}\max_{a\neq a'\in[k], r\in[k]} \min\{D_{ar}, D_{a'r}\}.$$

Let $n_r = \sum_{u\in[n]} \mathbb{I}\{\pi(\sigma(u)) = r\}, r \in [k]$ and $\pi(H)$ refers to the membership matrix corresponds to $\pi(\sigma)$. Then we have,

$$
\begin{aligned}
\sum_{u\in[n]} \mathbb{I}\{\hat{\sigma}(u) \neq \pi(\sigma(u))\} &= \sum_{n\in[n]} \sum_{r=1}^{k} \mathbb{I}\{\hat{H}_{ur} = 0, \pi(H_{ur}) = 1\} \\
&= \sum_{r=1}^{k} \sum_{n\in[n]} \mathbb{I}\{\hat{H}_{ur} = 0, \pi(H_{ur}) = 1\} \\
\text{Because } \sum_{u\in[n]} \mathbb{I}\{\pi(H_{ur}) = 1\} = n_r, \quad &= \sum_{r=1}^{k} \left(n_r - \sum_{u\in[n]} \mathbb{I}\{\hat{H}_{ur} = 1 = \pi(H_{ur})\} \right) \\
&= \sum_{r=1}^{k} (n_r - D^{\pi}_{r,r}) \\
&\geq^* \sum_{r=1}^{k} (n_r - \max_{r'\in[k]} D_{r',r}) \\
&\geq \sum_{r=1}^{k} (\max_{a,a'\in[k]} \min\{D_{ar}, D_{a'r}\}) \\
&\geq \max_{a,a'\in[k], r\in[k]} \min\{D_{ar}, D_{a'r}\},
\end{aligned}
$$

where the inequality $\geq^*$ is not necessarily become equality even though $\pi = \pi^0 = \arg\min_{\pi\in S_k} \sum_{u\in[n]} \mathbb{I}\{\hat{\sigma}(u) \neq \pi(\sigma(u))\}$. Take a toy example. Let $k = 3$ and suppose below confusion matrix:

$$
D = \begin{bmatrix} 10 & 9 & 8 \\ 5 & 6 & 6 \\ 5 & 5 & 6 \end{bmatrix}.
$$

Here $D^{\pi_0} = D$ while $D^{\pi_0}_{r,r} < \max_{r'\in[3]} D_{r',r}$ for $r = 2, 3$. We can also calculate $MCR(\hat{H}, H) = \frac{6}{60}$ while $l(\hat{\sigma}, \sigma) = \frac{5+5+9+5+8+6}{60}$. We can conclude that $l(\hat{\sigma}, \sigma) \geq MCR(\hat{H}, H)$. And when the clusters k is fixed, there are $l(\hat{\sigma}, \sigma) \asymp MCR(\hat{H}, H)$.When $k \to \infty$, for $l(\hat{\sigma}, \sigma)$:

$$
l(\hat{\sigma}, \sigma) = \frac{\sum_{r\neq r', r, r'\in[k]} D^{\pi}_{r,r'}}{n} = \frac{k(k-1)\bar{D}^{\pi}_{r,r'}}{n},
$$

where $\bar{D}^{\pi}_{r,r'}$ is the average of $D^{\pi}_{r,r'}, \forall r \neq r' \in [k]$, which has value at the same level of the entry in D. Here I am wonder how does the entry of D change along with k, n. Consider the worst case of random guess when the community sizes are nearly equal, $D_{r,r'} \asymp \frac{n}{k^2}$. From this perspective, $MCR \asymp \frac{D_{r,r'}}{n} \asymp \frac{1}{k^2}$ will goes to infinity as long as $k \to \infty$ while $l(\hat{\sigma}, \sigma) = O(1)$ when $k \to \infty, n \to \infty$ for the worst case.

Similarly, for the $\eta = MCR = \frac{1}{n_{min}} \max_{a\neq a'\in[k], r\in[k]} \min\{D_{ar}, D_{a'r}\}$ in Lei's paper, consider the nearly equal community sizes case, $n_{min} \asymp \frac{n}{k}$. We still have $\eta \asymp \frac{D_{r,r'}k}{n} \asymp \frac{1}{k}$. That also implies η in Lei's paper will vanish when $k \to \infty$ even though the worst case.

1.3 *Discussion of Optimal Misclassification proportion*

The optimal minimax misclassification proportion is established by Zhang and Zhou(2015).

To show the optimal rate, define:

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

which is exactly the Renyi Divergence $D_{1/2}(Ber(a/n)\|Ber(b/n))$. The definition of order 1/2 Renyi Divergence is $D_{1/2}(P\|Q) = -2\log(\sum_i^n \sqrt{p_i q_i})$ where $i \in \mathbb{Z}$ is the outcome of discrete random variable X and $P(X = i) = p_i$. According to the Lemma B.1 in the supplement of Zhang and Zhou(2015), $I^* \asymp \frac{(a-b)^2}{na}$. That makes sense for the distance between the two Bernoulli distributions decreases when $n \to \infty$ and $a = O(1)$.

Therefore the optimal minimax misclassification rate is showed in below theorem.

Theorem 1 (Optimal rate for $\hat{\sigma}$, Zhang and Zhou(2015))**.** *Assume* $\frac{nI^*}{k\log k} = \frac{(a-b)^2}{ak\log k} \to \infty$*, then*

$$\inf_{\hat{\sigma}}\sup_{\Theta}\mathbb{E}l(\sigma,\hat{\sigma}) = \begin{cases} \exp\left(-(1+o(1))\frac{nI^*}{2}\right), & k=2 \\ \exp\left(-(1+o(1))\frac{nI^*}{\beta k}\right), & k\geq 3 \end{cases},$$

for both $\Theta = \Theta_0(n,k,a,b,\beta)$ *and* $\Theta = \Theta(n,k,a,b,\lambda,\beta,\alpha)$ *with* $\lambda \leq \frac{a-b}{2\beta k}$*, where* $\beta \in [1,\sqrt{5/3}]$.

Note that **Thm 1** requires $\frac{(a-b)^2}{ak\log k} \to \infty$. That means if a is a constant, the connectivity signal is $O(n^{-1})$ and $nI^* = O(1)$, where no consistent estimate $\hat{\sigma}$ exists. Therefore, to have a consistent estimate, at least $a > O(1)$.

When k is fixed, the $l(\hat{\sigma},\sigma)$(called MCR_{opt} below) converges to 0 as $exp(-a) \to 0$, which is equal to $exp(a) \to \infty$. Recall the asymptotic result in TBM and TBSM, the MCR_{TBM} and MCR_{TBSM} goes to 0 as $n \to \infty$. If $O(1) < a < O(\log n)$, $exp(a)$ is slower than n to go to the infinity. If $a = O(\log n)$, MCR_{opt} has the same rate as $MCR_{TBM,TBSM}$. If $a > O(\log n)$, then MCR_{opt} is faster than the other two. Notice that in TBM and TBSM, we require the irreducibility of the connectivity, however, in Θ_0 and Θ, if $a < O(n)$, then the signal a/n and b/n will vanish. To also satisfies the irreducibility, we should require $a \geq O(n)$. Therefore, when 3 models all have irreducibility, MCR_{opt} is much faster than $MCR_{TBM,TBSM}$.

When k is diverge, MCR_{TBM} and MCR_{TBSM} are unstable and degenerate, we can not tell any conclusion about them this time. For MCR_{opt}, the model allows $k \to \infty$ if satisfies $\frac{(a-b)^2}{ak\log k} \to \infty$. Therefore, if $k\log k < O(a)$, the misclassification proportion will converge to 0 as $exp(a/k) \to \infty$.

According to Zhang and Zhou(2015), the optimal lower bound is achieved by a novel reduction of the *global* minimax rate in to a *local* testing problem. Different with the regular meaning of *global* in *global optimum*, here *global* refers to that the optimal rate is a *global* property for the **whole** network. In contrast, *local* loss will only focuses on one node, i.e. node-wise loss. Lemma 2.1 in Zhang and Zhou(2015) makes it possible to study the global optimum via local loss when the model is homogeneous($B_{ii} = a/n, B_{i,j} = b/n, \forall i \neq j \in [k]$) and close under permutation.

Lemma 2 (Global to Local)**.** *Let* Γ *be any parameter space that SBM is homogeneous and close under permutation. Let* $S_\sigma(\hat{\sigma}) = \{\sigma' : \sigma' = \pi \circ \hat{\sigma}, \pi = \arg\min_{\pi\in S_k}\sum_{u\in[n]}\mathbb{I}\{\pi(\hat{\sigma}(u)) \neq \sigma(u)\}$ *and* $l(\hat{\sigma}(i),\sigma(i)) =$

$\sum_{\sigma' \in S_\sigma(\hat{\sigma})} \frac{\mathbb{I}\{\sigma'(i) \neq \sigma(i)\}}{|S_\sigma(\hat{\sigma})|}$. *Assume* $L(\hat{\sigma}) = \frac{1}{|\Gamma|} \sum_{\sigma \in \Gamma} \mathbb{E} l(\sigma, \hat{\sigma})$ *and* $L(\hat{\sigma}(1)) = \frac{1}{|\Gamma|} \sum_{\sigma \in \Gamma} \mathbb{E} l(\sigma(1), \hat{\sigma}(1))$. *Then we have:*

$$\inf_{\hat{\sigma}} L(\hat{\sigma}) = \inf_{\hat{\sigma}} L(\hat{\sigma}(1))$$

Intuitively, for heterogeneous SBM, we can add penalty to find an optimal $\hat{\sigma}$ that also reaches the optimal rate. Indeed, according to Zhang and Zhou(2015), a range of penalized likelihood-type estimates can achieve the optimal rate. Inspired by the penalized likelihood method and the thoughts of global to local, this paper proposes a computational feasible algorithm whose estimates can be rate-optimal.

2 ALGORITHM