Optimal Sampling and Clustering in the Stochastic Block Model

Seyoung Yun (KAIST) and Alexandre Proutiere (KTH)

NeurIPS 2019

Problem formulation

SBM: Random graph with n nodes and K non-overlapping clusters, $\mathcal{V}_1, \dots, \mathcal{V}_K$, of respective sizes $\alpha_1 n, \dots, \alpha_K n$ with $\alpha_k > 0$ for all k.
If node pair $(v, w) \in \mathcal{V}_i \times \mathcal{V}_j$ is sampled, an edge is observed w.p. p_{ij}.
$\boldsymbol{p} = [p_{ij}]_{1 \leq i,j \leq K}$.

Given a sampling budget T, design a sampling and clustering algorithm recovering, from the edge observations, the clusters as accurately as possible.

Information-theoretical limits

An algorithm π is (s, β)-*locally stable* at $(\boldsymbol{p}, \boldsymbol{\alpha})$, if there exists a sequence $\eta_n \geq 0$ with $\lim_{n\to\infty} \eta_n = 0$ such that for all partition vectors $\tilde{\boldsymbol{\alpha}}$ such that $\|\tilde{\boldsymbol{\alpha}} - \boldsymbol{\alpha}\|_2 \leq \beta$, π mis-classifies at most s nodes with probability greater than $1 - \eta_n$ for any n.

Theorem 1. *Let $s = o(n)$. Assume that there exists a (s, β)-locally stable clustering algorithm at $(\boldsymbol{p}, \boldsymbol{\alpha})$ for $\beta \geq \frac{s}{n} \log\left(\frac{n}{s}\right)$. Then we have:* $\liminf_{n\to\infty} \frac{2TD(\boldsymbol{p},\boldsymbol{\alpha})}{n \log(n/s)} \geq 1$, *where:*

$$D(\boldsymbol{p}, \boldsymbol{\alpha}) = \max_{\boldsymbol{x} \in \mathcal{X}(\boldsymbol{\alpha})} \Delta(\boldsymbol{x}, \boldsymbol{p}),$$

$$\Delta(\boldsymbol{x}, \boldsymbol{p}) = \min_{i,j: i \neq j} \sum_{k=1}^{K} x_{ik} KL(p_{ik}, p_{jk}) \quad \textit{and}$$

$$\mathcal{X}(\boldsymbol{\alpha}) = \left\{ \boldsymbol{x} : \alpha_i x_{ij} = \alpha_j x_{ji},\ \sum_{i=1}^{K} \alpha_i \sum_{j=1}^{K} x_{ij} = 1,\ \textit{and } x_{ij} \geq 0,\ \forall i,j \right\}.$$

Condition for exact recovery

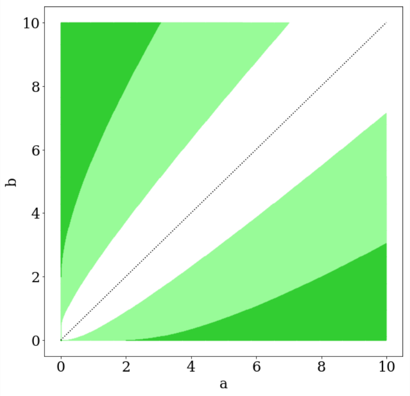
Binary symmetric SBM: $K = 2$, $\boldsymbol{\alpha} = (1/2, 1/2)$, $p_{11} = p_{22} = \frac{af(n)}{n}$, and $p_{12} = p_{21} = \frac{bf(n)}{n}$.
Budget $T = n(n-1)/2$.

Exact recovery is possible if either $f(n) = \omega(\log(n))$ or $f(n) = \log(n)$ and

$$\max\{\sqrt{a} - \sqrt{b}, \sqrt{b} - \sqrt{a}\} > \sqrt{2} \quad \text{(Non-adaptive sampling)}$$
$$\max\{a \log(\frac{a}{b}) + b - a, b \log(\frac{b}{a}) + a - b\} > \frac{1}{2} \quad \text{(Adaptive sampling)}$$

(For non-adaptive sampling, all node pairs are sampled once)

Condition for exact recovery



Light green: exact recovery is possible using adaptive sampling.
Dark green: exact recovery is possible using non-adaptive sampling.

Algorithm

ASP algorithm:

Step 1. Use a small fraction of the observation budget and apply spectral methods to obtain initial cluster estimates;

Step 2. Use these estimates to estimate the SBM parameters, and derive $\hat{\boldsymbol{x}}^\star$ of $\boldsymbol{x}^\star(\boldsymbol{p}, \boldsymbol{\alpha}) = \arg\max_{\boldsymbol{x} \in \mathcal{X}(\alpha)} \Delta(\boldsymbol{x}, \boldsymbol{p})$;

Step 3. $\hat{\boldsymbol{x}}^\star$ dictates the way to sample edges with the remaining budget, and based on these additional observations, the cluster estimates are improved.

Computational complexity: $O(T \log(n))$.

Error rate under the ASP algorithm

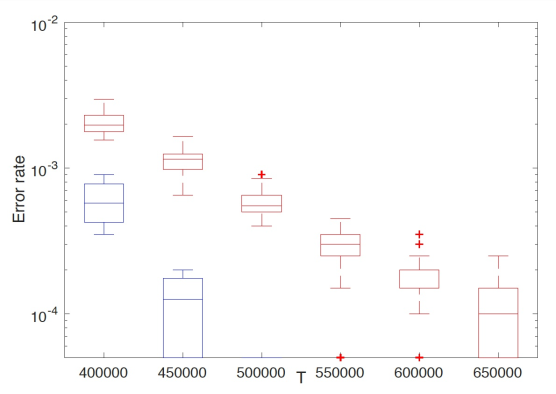
Assume that there exist positive constants κ_L and κ_U such that

$$\text{(A1)} \quad \left|\log\left(\frac{p_{ik}(1-p_{jk})}{p_{jk}(1-p_{ik})}\right)\right| \le \kappa_U \quad \text{for all} \quad i, j, k$$

$$\text{(A2)} \quad \kappa_L \le \left|\log\left(\frac{p_{ik}}{p_{jk}}\right)\right| \quad \text{for all} \quad i, j, k.$$

Theorem 2. *Let $s = o(n)$. The ASP algorithm mis-classifies less than s nodes with high probability, if* $\liminf_{n\to\infty} \frac{2TD(\boldsymbol{p},\boldsymbol{\alpha})}{n\log(n/s)} \ge 1$.

Numerical experiments



Blue: ASP algorithm.

Red: Optimal clustering algorithm with non-adaptive sampling.