Clustering in Block Markov Chains

Fundamental limits and algorithms

Se-Young Yun

KAIST (Graduate School of AI)
with Alexandre Proutiere (KTH), Jaron Sanders (TU/e)

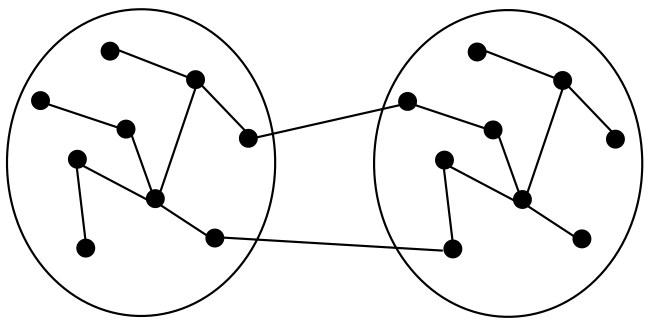
Table of contents

Introduction

Community Detection

Social networks

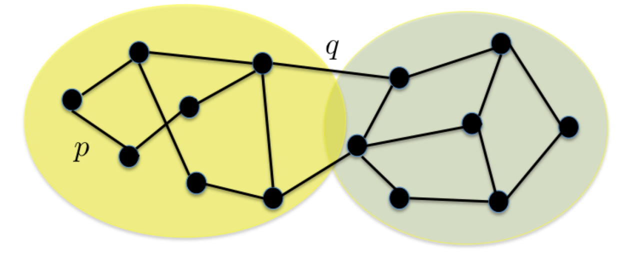
- facebook, twitter, ...
- Social graph



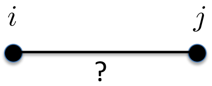
- How to find hidden community from the given graph

Objective
What is the minimum number of misclassified nodes when detecting communities from a graph? [Holland et al, 83],...,[Abbe 18]

Random Graph Model
All edges are independent.



Stochastic Block Model with Sampling [COLT2014]

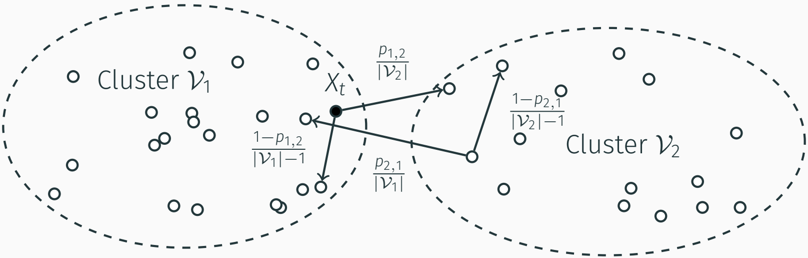


$$X_{ij}(t) \in \{0, 1\}$$

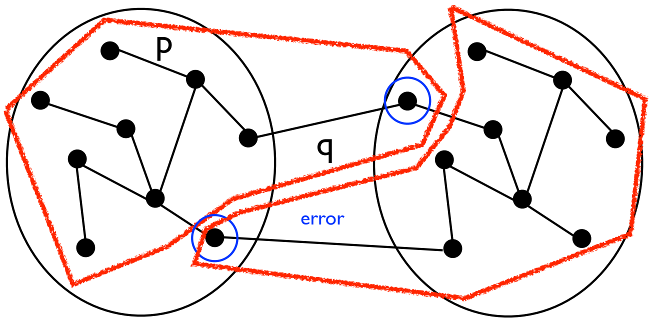
- At round t, sample node pair (i, j), and observe a random independent outcome
- Outcome follows Bernoulli with mean p if nodes are in the same cluster, q otherwise with $q < p$
- Budget: T observations
- Different sampling strategies: Random sampling and Adaptive sampling
- Remark: talks so far are for random sampling w/o replacement, and $T = \frac{n(n-1)}{2}$

Block Markov Chain [AoS2020]

Objective
What is the minimum number of misclassified states when detecting communities from a sample path?



Problem



SBM Analysis

Assumptions:

- $\bar{p} = o(1)$ and $\bar{p}n = \omega(1)$
- Homogeneity: $\exists \eta > 0 : \forall i, j, k, \frac{p_{ij}}{p_{ik}} \leq \eta$
- Separation: $\exists \epsilon > 0 : \forall i \neq j, \sum_k (p_{ik} - p_{jk})^2 \geq \epsilon \bar{p}^2$

Notations:

- Divergence between $p(i)$ and $p(j)$:

$$D_{L^+}(\alpha, p(i), p(j)) = \min_{y \in \mathcal{P}^K} \max_{a \in \{i,j\}} \sum_k \alpha_k KL(y_k, p_{ak})$$

- Divergence of the model: $D(\alpha, p) = \min_{i,j:i \neq j} D_{L^+}(\alpha, p(i), p(j))$

Random Sampling

Spectral Algorithm+ [NeurIPS 2016]

Step 1. Input: matrices A ($A_{vw} = 1$ iff (v, w) is connected)

1. Trimming + Spectral method (PI+SV thresholding)
2. Output $S_1, \ldots, S_{\hat{K}}$

Step 2. Input: A, and $S_1, \ldots, S_{\hat{K}}$

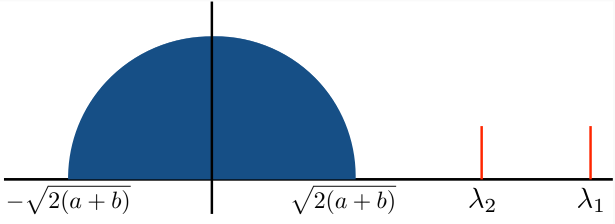
1. Estimate p: $\hat{p}_{ij} \leftarrow \frac{\sum_{u \in S_i} \sum_{v \in S_j} A_{uv}}{|S_i||S_j|}$
2. $\lceil \log(n) \rceil$ improvement iterations: in each iteration, for all v, assign v to

$$\arg\max_k \sum_i \sum_{w \in S_i} A_{vw} \log \hat{p}_{ki}$$

Eigen values of adjacent matrix A

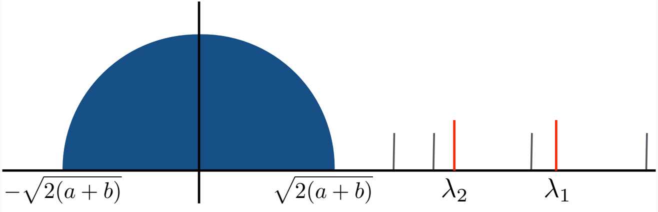
Example: 2 clusters with equal sizes, $np = a$, $nq = b$.

- Eigen values of A and two separated eigen values
  - $z_1 = \frac{1}{2}(a-b) + \frac{a+b}{a-b}$ and $z_2 = \frac{1}{2}(a+b) + 1$
  - z_1 appears only if $(a-b)^2 > 2(a+b)$
  - Eigen vector of z_1 : $\frac{\theta}{\sqrt{n}} u + \frac{1-\theta^2}{\sqrt{n}} \mathcal{N}(0,1)$ where $\theta^2 = \frac{(a-b)^2 - 2(a+b)}{(a-b)^2}$



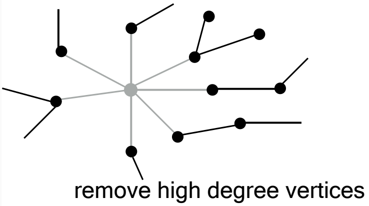
Noise on eigenvalues

- The case of $d < \log n$ is more challenging
  - The graph becomes less regular
  - Even if a and b are constants, there exists $\Omega\left(\frac{\log n}{\log\log n}\right)$ degree vertex
  - The largest eigenvalue becomes $\Omega\left(\sqrt{\frac{\log n}{\log\log n}}\right)$ (the largest eigenvalue of n-star graph $\approx \sqrt{n}$)
  - Impossible to pick λ_1 and λ_2



Trimming

- Pre-process (trimming)



- Trimming of Erdös-Rényi graph
  - Consider a ER graph with average degree d
  - Trimming : remove all the vertices of degree greater than $(1+\varepsilon)d$
  - Let A' be the adjacent matrix after trimming

Theorem (Feige, Ofek '05)
The second largest singular value of A' is less than $c\sqrt{d}$

Achievability (cont'd)

Theorem 1 *Let $s = o(n)$. If $\liminf_{n\to\infty} \frac{nD(\alpha,p)}{\log(n/s)} \geq 1$, then the SP algorithm misclassifies at most s nodes with high probability.*

Remark 1: This result covers exact detection ($s = 1$).
Remark 2: SP runs in $O(\bar{p}n^2 \log(n))$, and does not need to know K nor (α, p)

Proof: Key ingredients

- The spectral norm of the random observation matrix
- Chernoff-Hoeffding's inequality to understand the number of connections to each cluster
- *i.i.d observations*

Proofs

Spectral Analysis

It is very important to understand the spectral norm of $A - \mathbb{E}[A]$

Some works study $\sup_{x \in \mathbb{R}^n : \|x\|_2 = 1} \|(A - \mathbb{E}[A])x\|_\infty$

Greedy Improvement

Concentration inequalities are very important!

Theorem 2 *Let $s = o(n)$. If there exists a clustering algorithm π such that $\limsup_{n\to\infty} \mathbb{E}[\varepsilon^{\pi}(n)]/s \leq 1$, then:*

$$\liminf_{n\to\infty} \frac{nD(\alpha, p)}{\log(n/s)} \geq 1$$

Hence, the average number of misclassified nodes should scale at least as $n \exp(-nD(\alpha, p)(1 + o(1))$

Theorem 3 *If there exists a clustering algorithm classifying each node correctly with high probability, then:*

$$\liminf_{n\to\infty} \frac{nD(\alpha, p)}{\log(n)} \geq 1$$

Adaptive Sampling

Adaptive Spectral Algorithm [NeurIPS 2019]

Step 1. Randomly sample δT edges with a small $\delta > 0$ and run the spectral clustering algorithm to extract the first guess $\mathcal{S}_1, \ldots, \mathcal{S}_K$.

Step 2. Estimate SBM parameters $\boldsymbol{\alpha}$ and $\boldsymbol{p}$

Step 3. Solve the LP $D^A(\boldsymbol{p}, \boldsymbol{\alpha}) = \max_{\boldsymbol{x} \in \mathcal{X}(\boldsymbol{\alpha})} D(\boldsymbol{x}, \boldsymbol{p})$

Step 4. Sample edges between $v \in \mathcal{S}_i$ and $\mathcal{S}_j$ using $2(1 - 2\delta) x_{ij} \frac{T}{n}$ budget and classify nodes

Step 5. Use the remaining budgets to classify unclear nodes

We can control the sampling sequence.

New Divergence:

$D^A(\boldsymbol{p}, \boldsymbol{\alpha})$ is defined as: $D^A(\boldsymbol{p}, \boldsymbol{\alpha}) = \max_{x \in \mathcal{X}(\boldsymbol{\alpha})} D(\boldsymbol{x}, \boldsymbol{p})$,

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

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

Inference Limits in the SBM with Adaptive Sampling

Theorem 4 *Let $s = o(n)$. If there exists a clustering algorithm π such that $\varepsilon^\pi(n) \leq s$ with high probability, then:*

$$\liminf_{n\to\infty} \frac{2TD^A(\boldsymbol{p},\boldsymbol{\alpha})}{n\log(n/s)} \geq 1.$$

Theorem 5 *Let $s = o(n)$. Adaptive Spectral Algorithm guarantee that $\varepsilon^\pi(n) \leq s$ with high probability, when*

$$\liminf_{n\to\infty} \frac{2TD^A(\boldsymbol{p},\boldsymbol{\alpha})}{n\log(n/s)} \geq 1.$$

Block Markov Chain

Mixing time

Analyzing and bounding the **mixing time** of a BMC is crucial.

Without mixing within T time steps, we would not expect to be able to cluster.

We define $d(t) \triangleq \sup_{x\in\mathcal{V}}\{d_{\mathrm{TV}}(P^t_{x,\cdot},\Pi)\}$ and $t_{\mathrm{mix}}(\varepsilon) \triangleq \min\{t \geq 0 : d(t) \leq \varepsilon\}$, where

$$d_{\mathrm{TV}}(\mu,\nu) \triangleq \frac{1}{2}\sum_{x\in\mathcal{V}} |\mu_x - \nu_x|. \tag{1}$$

Proposition 1 *There exists a strictly positive absolute constant* c_{mix} *such that* $t_{\mathrm{mix}}(\varepsilon) \leq -c_{\mathrm{mix}} \ln \varepsilon$, *for every BMC of finite size* $n \geq K$.

Mixing Time for Block Markov Chain

Let $\alpha_i n$ is the number of nodes in $\mathcal{V}_i$. The transition matrix is

$$P_{x,y} = \frac{p_{\sigma(x),\sigma(y)}}{|V_{\sigma(y)}|} \quad \text{for all} \quad x, y \in \mathcal{V}.$$

Let $\alpha_{\min} = \min_k \alpha_k$ and $\eta = \max_{a,b,c}\{p_{b,a}/p_{c,a}, p_{a,b}/p_{a,c}\}$.

Proposition 1 *For any BMC with $n \geq 4/\alpha_{\min}$, $t_{mix}(\epsilon) \leq -c_{mix} \log \epsilon$, where $c_{mix} = -1/\log(1 - 1/2\eta)$.*

Information theoretical lower bound

For $\alpha \in \Delta^{K-1}$ and $p \in \mathbb{\Delta}^{(K-1)\times K}$, let

$$I(\alpha, p) \triangleq \min_{a \neq b} \left\{ \sum_{k=1}^{K} \frac{1}{\alpha_a} \left(\pi_a p_{a,k} \ln \frac{p_{a,k}}{p_{b,k}} + \pi_k p_{k,a} \ln \frac{p_{k,a}\alpha_b}{p_{k,b}\alpha_a} \right) + \left(\frac{\pi_b}{\alpha_b} - \frac{\pi_a}{\alpha_a} \right) \right\}. \tag{2}$$

Here π denotes the solution to $\pi^{\mathrm{T}} p = \pi^{\mathrm{T}}$.

Theorem 6 *Let $s = o(n)$. If there exists a clustering algorithm π such that $\limsup_{n\to\infty} \mathbb{E}[\varepsilon^{\pi}(n)]/s \geq 1$, then:*

$$\liminf_{n\to\infty} \frac{(T/n)I(\alpha, p)}{\log(n/s)} \geq 1$$

Achievability

The error lower bounds are tight! Spectral Algorithm+ [AOS 2020]

Step 1. Input: matrices A

1. Trimming + Spectral method (PI+SV thresholding)
2. Output $S_1, \ldots, S_{\hat{K}}$

Step 2. Input: A, and $S_1, \ldots, S_{\hat{K}}$

1. Estimate p: $\hat{p}(i,j) \leftarrow \frac{\sum_{u \in S_i} \sum_{v \in S_j} A_{uv}}{|S_j|}$
2. $\lceil \log(n) \rceil$ improvement iterations: in each iteration, for all v, assign v to

$$\arg\max_c \Big\{ \sum_{k=1}^{K} \big(\hat{N}_{x,\hat{\mathcal{V}}_k} \ln \hat{p}(c,k) + \hat{N}_{\hat{\mathcal{V}}_k,x} \ln \frac{\hat{p}(k,c)}{\hat{\alpha}_c} \big) - \frac{T}{n} \cdot \frac{\hat{\pi}_c}{\hat{\alpha}_c} \Big\}$$

Theorem 7 *Let $s = o(n)$. If $\liminf_{n\to\infty} \frac{(T/n)I(\alpha,p)}{\log(n/s)} \geq C$ with a constant $C > 0$, then the SP algorithm misclassifies at most s nodes with high probability.*

Remark 1: This result is not tight. Here, $C < 1$.
Remark 2: We utilize concentration inequalities for Markov chains, but they are not enough to make the tight result.

$$\sum_{k=1}^{K}\left(\hat{N}_{x,\mathcal{V}_k} \ln p_{c,k} + \hat{N}_{\mathcal{V}_k,x} \ln \frac{p_{k,c}}{\alpha_c}\right) - \frac{T}{n} \cdot \frac{\pi_c}{\alpha_c}$$

[D. Paulin, 2015] *Let $X_1, \ldots,$ be a Markov chain with transition matrix P. Let Π be the stationary distribution. Let $f \in L^2(\Pi)$ with $|f(x) - \mathbb{E}_\Pi(f)| < C$ for every $x \in \Omega$ and some constant $C > 0$. Let V_f be the variance of $f(X)$ when X follows the stationary distribution Π. Then, for any $z > 0$,*

$$\mathbb{P}_\Pi\left(|\sum_{t=1}^{T} f(X_t) - \mathbb{E}_\Pi[\sum_{t=1}^{T} f(X_t)]| \geq z\right) \leq 2\exp\left(-\frac{z^2\gamma_{ps}}{8(T + 1/\gamma_{ps})V_f + 20zC}\right),$$

where

$$\gamma_{ps} = \max_{i \geq 1} \frac{1 - \lambda((P^*)^i P^i)}{i} \geq \frac{1-\epsilon}{t_{mix}(\epsilon/2)} \quad \text{with} \quad P^*(x,y) = \frac{P(x,y)}{\Pi(x)}\Pi(y).$$

Concentrations for BMC

The block Markov chain has

$$\gamma_{ps} \geq \frac{1}{2(t_{mix}(1/4)+1)} \geq \frac{1}{2(4\eta+1)}.$$

Therefore, from the concentration inequality for Markov chains by Paulin,

$$\mathbb{P}\left(|\hat{N}_{\mathcal{A},\mathcal{B}} - N_{\mathcal{A},\mathcal{B}}| \geq c\sqrt{nT}\right) \leq 2\exp\left(-\frac{c^2}{16(4\eta+1)}n(1+o(1))\right),$$

which can analyze the accuracy of parameter estimations.

Conclusion

Summary

- The stochastic block model (SBM) is a natural performance benchmark for community detection.
- We address the finer and more challenging question of determining, under the general LSBM, the minimal number of misclassified items given the parameters of the model.
- We extend our results to the block Markov chain model.
- The results for the block Markov chain model is not tight. To obtain the tightness, it is necessary to derive a much better concentration inequality for the Markov chain sample path.

Thank you!

Questions?