Econometrics **September 14, 2023**

Topic 19: Community Detection

by Sai Zhang

Key points: .

Disclaimer: *The note is built on Prof. Jinchi Lv's lectures of the course at USC, DSO 607, High-Dimensional Statistics and Big Data Problems.*

19.1 Stochastic Block Model (Abbe et al., 2015)

Consider an undirected graph G, with nodes V and edges E. Let

- n be a positive integer: the number of **vertices**
- k be a positive integer: the number of **communities**
- $p = (p_1, \cdots, p_k)$ be a probability vector on $\{1, \cdots, k\} := [k]$: the **prior** on the k communities
- $\mathbf{W}$ be a $k \times k$ symmetric matrix with entries $W_{ij} \in [0, 1]$: the matrix of **connectivity probabilities**

then we have

Definition 19.1.1: Stochastic Block Model

The pair $(\mathbf{X}, G)$ is drawn under $SBM(n, p, \mathbf{W})$ if $\mathbf{X}$ is an n dimensional random vector with i.i.d. components distributed under p, and G is an $n-$vertex simple graph where vertices i and j are connected with probability W_{X_i,X_j}, **independently** of other pairs of vertices. And the **community** sets can be defined by

$$\Omega_i = \Omega_i(\mathbf{X}) := \{v \in [n] : X_v = i\}, i \in [k]$$

Immediately, we can define the symmetry of SBM as:

Definition 19.1.2: Symmetric SBM

An SBM is called symmetric if

- p is **uniform**
- $\mathbf{W}$ takes the same value **on the diagonal** and the same value **off the diagonal**

$(\mathbf{X}, G)$ is drawn under $SSBM(n, k, A, B)$ if $p = \{1/k\}^k$ and $\mathbf{W}$ takes avlue A on the diagonal and B off the diagonal.

19.1.1 Recovery

The goal of community detection is to recover the labels $\mathbf{X}$ by observing G, up to some level of accuracy. First, define **agreement** as

Definition 19.1.3: Agreement of Communities

The agreement between two community vectors $\mathbf{x}, \mathbf{y} \in [k]^n$ is obtained by maximizing the common components between $\mathbf{x}$ and any relabelling of $\mathbf{y}$, that is

$$A(\mathbf{x}, \mathbf{y}) = \max_{\pi \in S_k} \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}\left[x_i = \pi(y_i)\right]$$

where S_k is the group of permutations on $[k]$.

The **relabelling** permutation is used to handle symmetric communities such as in SSBM, as it is impossible to recover the actual labels in this case. But it's possible to recover the **partition**. There are 2 types of partition recovery we consider

Exact Recovery First, consider the case of **exact recovery**:

Definition 19.1.4: Exact Recovery

Let $(\mathbf{X}, G) \sim SBM(n, p, W)$, the exact recovery is solved if there exists an algorithm that takes G as an input and outpus $\hat{\mathbf{X}} = \hat{\mathbf{X}}(G)$ such that $\mathbb{P}\left\{A(\mathbf{X}, \hat{\mathbf{X}}) = 1\right\} = 1 - o_p(1)$

In the SSBM case, algorithms that guarantee

$$A(\mathbf{X}, \hat{\mathbf{X}}) \to \frac{1}{k}$$

would be trivial.

Weak Recovery On the other hand, we the case of **weak recovery** defined as

Definition 19.1.5: Weak Recovery

Weak recovery or detection is solved $SSBM(n, k, A, B)$ if for $(\mathbf{X}, G) \sim SSBM(n, k, A, B)$, then $\exists \epsilon > 0$ and an algorithm that takes G as an input and outputs $\hat{\mathbf{X}}$ such that

$$\mathbb{P}\left\{A(\mathbf{X}, \hat{\mathbf{X}}) \geq \frac{1}{k} + \epsilon\right\} = 1 - o(1)$$

19.1.2 Example: SSBM(n,2)

Let's look at the example of $SSBM(n, 2, \alpha\frac{\log n}{n}, \beta\frac{\log n}{n})$, where

- n: number of vertices (assumed to be even for simplicity)
- for each $v \in [n]$, a binary label X_v is attached s.t.

$$|\{v \in [n] : X_v = 1\}| = n/2$$

- for each pair of distinct nodes $u, v \in [n]$, an edge is placed with probability

  - $\alpha \frac{\log n}{n}$ if $X_u = X_v$
  - $\beta \frac{\log n}{n}$ if $X_u \neq X_v$

  where edges are placed independently conditionally on the vertex labels
- WLOG, $\alpha > \beta$

then we have the following theorem

Theorem 19.1.6: Exact Recovery in $SSBM(n, 2, \alpha\log(n)/n, \beta\log(n)/n)$

- Exact recovery in $SSBM(n, 2, \alpha\log(n)/n, \beta\log(n)/n)$ is solvable and efficiently so if $\left|\sqrt{\alpha} - \sqrt{\beta}\right| > \sqrt{2}$ nad unsolvable if $\left|\sqrt{\alpha} - \sqrt{\beta}\right| < \sqrt{2}$
- Exact recovery of the ground truth assignment of the partition (A, B) is also achieveable, that is: if

$$\frac{\alpha+\beta}{2} - \sqrt{\alpha\beta} > 1$$

i.e.

$$\alpha+\beta > 2,\ (\alpha-\beta)^2 > 4(\alpha+\beta) - 4$$

the maximum likelihood estimator exactly recovers the communities (up to a global flip), with high probability.

See Abbe (2017) for the proof of this theorem.

In summary, for a graph structure $G = (V, E)$ represented by adjacency matrix $\mathbf{X}_{n\times n}$, Stochastic Block Model (SBM)

- assumes that there is a symmetric matrix $\mathbf{P} = \{p_{ij}\} \in \mathbb{R}^{k\times k}$, for $k \ll n$ and a map $C : \{1, \cdots, n\} \to \{1, \cdots, k\}$, s.t. $\Pr(\mathbf{X}_{ij} = 1) = \mathbf{P}_{C(i),C(i)}$
- Define $\mathbf{\Pi} = (\pi_1, \cdots, \pi_n)' \in \mathbb{R}^{n\times k}$ where $\mathbf{\Pi}_{ij} = 1$ if $C(i) = j$, and $\mathbf{\Pi}_{ij} = 0$ otherwise
- Let $\mathbf{H} = \mathbb{E}(\mathbf{X})$ be the probability matrix, then $\mathbf{H} = \mathbf{\Pi P \Pi}'$
- A variant of SBM is degree corrected SBM which incorporates the degree heterogeneity.
  - each node is assigned a parameter $\theta_i > 0$ such that $\Pr(\mathbf{X}_{ij} = 1) = \theta_i\theta_j\mathbf{P}_{C(i),C(j)}$
  - $\mathbf{H} = \mathbf{\Theta\Pi P\Pi}'\mathbf{\Theta}$, where $\mathbf{\Theta} = \operatorname{diag}(\theta_1, \cdots, \theta_n)$

19.2 SIMPLE Model (Fan et al., 2022)

In SBM, each $\pi_i \in \{e_1, \cdots, e_K\}$ with e_k a one entry vector whose k-th component is one. But what if each node i can belong to K different communities? We generalize π_i to be a compositional vector, and interpret it as community membership profile for node i, then

$$\Pr(\mathbf{X}_{ij} = 1) = \theta_i\theta_j \sum_{k=1}^{K}\sum_{l=1}^{K} \pi_i(k)\pi_j(l)p_{kl}$$

and $\mathbf{H} = \mathbf{\Theta\Pi P\Pi}'\mathbf{\Theta}$. Now, consider a new statistical tests for testing whether any given pair of nodes share the same membership profiles, and providing the associated p-values.

19.2.1 Problem Setting

For an undirected graph $G = (V, E)$ with n nodes, let $\mathbf{X} = \{x_{ij}\} \in \mathbb{R}^{n\times n}$ be the **symmetric** adjacency matrix. Under a probabilistic model, assume x_{ij} is an independent realization from a Bernoulli random variable for all upper triangular entries of random matrix $\mathbf{X}$. Consider the adjacency matrix with the deterministic-random latent structure

$$\mathbf{X} = \mathbf{H} + \mathbf{W}$$

where

- $\mathbf{H} = \{h_{ij}\} \in \mathbb{R}^{n\times n}$ is the deterministic mean matrix of low rank $K \geq 1$
- $\mathbf{W} = \{w_{ij}\} \in \mathbb{R}^{n\times n}$ is a symmetric random matrix with zero mean and independent entries on and above the diagonal

Assume V is decomposed into K disjoint latent communities

$$C_1, \cdots, C_K$$

where each node i is associated with the community membership probability vector

$$\boldsymbol{\pi}_i = (\pi_i(1), \cdots, \pi_i(K))' \in \mathbb{R}^K$$

s.t.

$$\Pr(i \in C_k) = \pi_i(k),\ k = 1, \cdots, K$$

here, K is unknown but bounded away from ∞.

19.2.2 Hypothesis Testing

For any given pair of nodes $i \neq j \in V$, the goal is to infer whether they share the same community identity with quantified uncertainty level based on adjacency matrix $\mathbf{X}$, the hypothesis is

$$H_0 : \pi_i = \pi_j \qquad\qquad H_1 : \pi_i \neq \pi_j$$

More explicitly, consider the DCMM (Degree Corrected Mixed Membership) model as the underlying network model, s.t. the probability of a link between nodes i and j can be written as

$$\Pr(\mathbf{X}_{ij} = 1) = \theta_i\theta_j \sum_{k=1}^{K}\sum_{l=1}^{K} \pi_i(k)\pi_j(l)p_{kl}$$

and

$$\mathbf{H} = \mathbf{\Theta\Pi P \Pi'\Theta}$$

in matrix form, where $\mathbf{\Pi} = (\pi_1, \cdots, \pi_n)' \in \mathbb{R}^{n\times k}$ and $\mathbf{\Theta} = \operatorname{diag}(\theta_1, \cdots, \theta_n)$. Consider

- No degree homogeneity: $\mathbf{\Theta} = \sqrt{\theta}\mathbf{I}_n$, then $\mathbf{H} = \theta\mathbf{\Pi P\Pi'}$. If we eigen-decompose $\mathbf{H} = \mathbf{VDV'}$ where $\mathbf{D} = \operatorname{diag}(d_1, \cdots, d_K)$ with $|d_1| \geq |d_2| \geq \cdots |d_K| > 0$ is the matrix of all K non-zero eigenvalues and $V = (v_1, \cdots, v_K) \in \mathbb{R}^{n\times K}$ is the eigenvectors.
  - the column space spanned by $\mathbf{\Pi}$ is the same as the eigenspace spanned by the top K eigenvectors of matrix $\mathbf{H}$
  - mean matrix $\mathbf{H}$ is **not** observable: replace it with adjacency matrix $\mathbf{X}$ and conduct eigen-decomposition to get eigenvalues $\hat{d}_1, \cdots, \hat{d}_n$ and eigenvectors $\hat{v}_1, \cdots, \hat{v}_n$. We assume that

    $$|\hat{d}_1| \geq |\hat{d}_2| \geq \cdots \geq |\hat{d}_n|$$

    and let $\hat{\mathbf{V}} = (\hat{v}_1, \cdots, \hat{v}_K) \in \mathbf{R}^{n\times K}$.

Without degree heterogeneity first, consider the case where $\mathbf{\Theta} = \sqrt{\theta}\mathbf{I}_n$ and $\mathbb{E}(\mathbf{X}) = \mathbf{H} = \theta\mathbf{\Pi}\mathbf{P}\mathbf{\Pi}'$. If $\pi_i = \pi_j$, then nodes i and j are exchangeable and $\mathbf{V}(i) = \mathbf{V}(j)$. The test statistic for membership information of node i and j is given as

$$T_{ij} = \left[\hat{\mathbf{V}}(i) - \hat{\mathbf{V}}(j)\right]' \mathbf{\Sigma}_1^{-1} \left[\hat{\mathbf{V}}(i) - \hat{\mathbf{V}}(j)\right]$$

where $\mathbf{\Sigma}_1^{-1} = \mathrm{Cov}\left[(e_i - e_j)'\mathbf{W}\mathbf{V}\mathbf{D}^{-1}\right]$ is the asymptotic variance of $\left[\hat{\mathbf{V}}(i) - \hat{\mathbf{V}}(j)\right]$. The regularity conditions are

C1 $\exists c_0 > 0$ s.t.

$$\min\left\{\frac{|d_i|}{|d_j|} : 1 \le i \le j \le K, d_i \ne -d_j\right\} \ge 1 + c_0$$

C2 $\exists c_0 \in (0,1), c_2 \in [0, 1/2), c_1 \in (0, 1-2c_2)$ s.t. $\lambda_k(\mathbf{\Pi}'\mathbf{\Pi}) \ge c_0 n$, $\lambda_K(\mathbf{P}) \ge n^{-c_2}$ and $\theta \ge n^{-c_1}$

C3 as $n \to \infty$, all the eigenvalues of $\theta^{-1}\mathbf{D}\mathbf{\Sigma}_1\mathbf{D}$ are bounded away from 0 and ∞

and the test statistics follow the theorem

Theorem 19.2.1: Test Statistics Distribution

Under Condition **C1** and **C2**, and $\mathbf{\Theta} = \sqrt{\theta}\mathbf{I}_n$,

- If **C3** holds too, then under the null

$$H_0 : T_{ij} \xrightarrow{\mathcal{D}} \chi^2_K$$

  as $n \to \infty$, where χ^2_K is the chi-square distribution with K degrees of freedom
- under the alternative,
  - if $n^{1/2-c_2}\sqrt{\theta}\left\|\pi_i - \pi_j\right\| \to \infty$, then for arbitrarily large constant $C > 0$, we have

$$\Pr\left(T_{ij} > C\right) \xrightarrow{n\to\infty} 1$$

  - in addition, if Condition **C3** holds, $c_2 = 0$, $\left\|\pi_i - \pi_j\right\| \sim \frac{1}{\sqrt{n\theta}}$, and

$$\left[\mathbf{V}(i) - \mathbf{V}(j)\right]' \mathbf{\Sigma}_1^{-1} \left[\mathbf{V}(i) - \mathbf{V}(j)\right] \to \mu$$

    , then

$$T_{ij} \xrightarrow{\mathcal{D}} \chi^2_K(\mu)$$

    as $n \to \infty$, where $\chi^2_K(\mu)$ is a noncentral chi-square distribution with mean μ and K degrees of freedom.

Under the joint null $H_{0,ij} : \pi_i = \pi j, \forall 1 \le i \ne j \le n$, a uniform version of Thm.19.2.1 is

$$\lim_{n\to\infty} \sup_{1\le i \ne j \le n} \left|\Pr\left(T_{ij} \le x\right) - \Pr\left(X \le x\right)\right| = 0, \forall x \in \mathbf{R}$$

where $X \sim \chi^2_K$. But the test statistic T_{ij} is not directly applicable since the population parameters K and $\mathbf{\Sigma}_1$. For consistent estimators satisfying the following condition

$$\Pr(\hat{K} = K) = 1 - o(1)$$
$$\theta^{-1}\left\|\mathbf{D}\left(\hat{\mathbf{S}}_1 - \mathbf{\Sigma}_1\right)\mathbf{D}\right\|_2 = o(1)$$

then the asymptotic results in Thm. 19.2.1 holds.

With degree heterogeneity Define componentwise ratio

$$Y(i,k) = \frac{\hat{v}_k(i)}{\hat{v}_1(i)}, \qquad 1 \le i < n, 2 \le k \le K$$

where $\hat{v}_k(i)$ is the i-th component of k-th eigenvector of $\mathbf{X}$. Due to the **exchangeability** of nodes i and j, under the null it holds that

$$\frac{v_k(j)}{v_1(j)} = \frac{v_k(j)}{v_1(j)}, \qquad 2 \le k \le K$$

Denote $\mathbf{Y}_i = (Y(i,2), \cdots, Y(i,K))'$, the new test statistics is proposed as

$$G_{ij} = \left(\mathbf{Y}_i - \mathbf{Y}_j\right)' \mathbf{\Sigma}_2^{-1} \left(\mathbf{Y}_i - \mathbf{Y}_j\right)$$

where $\mathbf{\Sigma}_2$ is the asymptotic variance of $\mathbf{Y}_i - \mathbf{Y}_j$, which is much harder to derive and estimate. So we need to impose four other conditions in addition to Condition C1-C3:

C4 $\exists c_2 \in [0, 1/2), c_3 \in (0, 1-2c_2), c_4 > 0, c_5 \in (0,1)$ s.t.

$$\lambda_K(\mathbf{P}) \ge n^{-c_2} \qquad \min_{1\le k\le K} |\mathcal{N}_k| \ge c_5 n \qquad \theta_{\max} \le c_4 \theta_{\min} \qquad \theta_{\min}^2 \ge n^{-c_3}$$

C5 $\mathbf{P} = (p_{kl})$ is positive definite, irreducible and has unit diagonal entries, moreover

$$n \min_{1\le k\le K, t=i,j} \operatorname{Var}\left(\mathbf{e}_t' \mathbf{W} \mathbf{v_k}\right) \sim n\theta_{\max}^2 \to \infty$$

C6 all the eigenvalues of

$$\left(n\theta_{\max}^2\right)^{-1} \mathbf{D} \mathrm{Cov}(f) \mathbf{D}$$

are bounded away from 0 and ∞

C7 Let η_1 be the first right singular vector of $\mathbf{P\Pi'\Theta^2\Pi}$, it holds that

$$\min_{1\le k\le K} \eta_1(k) > 0 \qquad \frac{\max_{1\le k\le K} \eta_1(k)}{\min_{1\le k\le K} \eta_1(k)} \le C$$

for some positive C, where $\eta_1(k)$ is the k-th entry of η_1.

Then we have

Theorem 19.2.2: Test Statistic Distribution with Degree Heterogeneity

Under Condition **C1, C4-C7**, with degree heterogeneity,

- under the null,

$$G_{ij} \xrightarrow{\mathcal{D}} \chi_{K-1}^2$$

- under the alternative with $\lambda_2\left(\boldsymbol{\pi}_i\boldsymbol{\pi}_i' + \boldsymbol{\pi}_i\boldsymbol{\pi}_j'\right) \gg \frac{1}{n^{1-2c}\theta_{\min}^2}$, for any arbitrarily large constant $C > 0$,

$$\Pr(G_{ij} > C) \xrightarrow{n\to\infty} 1$$

notice that K and $\mathbf{\Sigma}_2$ are both unknown, we must have

- for estimator $\hat{\mathbf{S}}_2$ of $\mathbf{\Sigma}_2$, we need

$$\left(n\theta_{\max}^2\right)^{-1} \left\|\mathbf{D}\left(\hat{\mathbf{S}}_2 - \mathbf{\Sigma}^2\right)\mathbf{D}\right\|_2 = o_p(1)$$

  replace $\mathbf{\Sigma}_2$ with $\hat{\mathbf{S}}_2$

- for K, under Condition **C1**, and $|d_K| \gg \sqrt{\log(n)}\alpha_n$ and $\alpha_n \geq n^{c_5}$ for some positive constant c_5, a consistent thresholding estimator is defined

$$\hat{K} = \left|\left\{\hat{d}_i : \hat{d}_i^2 > 2.01(\log n)\check{d}_n, i \in [n]\right\}\right|$$

  where the constant 2.01 can be replace with any other constant slightly larger than 2, and

$$\check{d}_n = \max_{1\leq l\leq n} \sum_{j=1}^{n} \mathbf{X}_{lj}$$

  is the maximum degree of the network. For $\hat{K}$ to be consistent, we need
  - Condition **C1** holds
  - $|d_K| \gg \sqrt{\log(n)}\alpha_n$, where $\alpha_n \geq n^{c_5}$ for some constant $c_5 > 0$

19.3 Rank Inference via Residual Subsampling

Again, consider $n \times n$ symmetric random matrix $\tilde{\mathbf{X}}$ and its decomposition

$$\tilde{\mathbf{X}} = \mathbf{H} + \mathbf{W}$$

where

- $\mathbf{H} = \mathbb{E}\left(\tilde{\mathbf{X}}\right)$ with some fixed but unknown rank $K \ll n$, it can be eigen-decomposed as

$$\mathbf{H} = \mathbf{VDV}'$$

  where $\mathbf{D} = \operatorname{diag}(d_1, \cdots, d_K)$ are the non-zero eigenvalues of $\mathbf{H}$ in decreasing magnitude and $\mathbf{V} = (\mathbf{v}_1, \cdots, \mathbf{v}_K)$ are the corresponding eigenvectors
- $\mathbf{W}$ has bounded and independent entries on and above the diagonals

for a simple case (networks with self loops), when the observed data matrix $\mathbf{X} = \tilde{\mathbf{X}} = \mathbf{H} + \mathbf{W}$, then we have

$$\begin{array}{l} \dfrac{\sum_{i=1}^n w_{ii}}{\sqrt{\sum_{i=1}^n \mathbb{E}w_{ii}^2}} \xrightarrow{d} \mathcal{N}(0,1) \\ \dfrac{\sum_{i=1}^n \mathbb{E}w_{ii}^2}{\sum_{i=1}^n w_{ii}^2} \xrightarrow{p} 1 \end{array} \quad \Rightarrow \frac{\sum_{i=1}^n w_{ii}}{\sqrt{\sum_{i=1}^n w_{ii}^2}} \xrightarrow{d} \mathcal{N}(0,1)$$

Let $\hat{\mathbf{V}}\hat{\mathbf{D}}\hat{\mathbf{V}}' = \sum_{k=1}^{K_0} \hat{d}_k \hat{\mathbf{v}}_k \hat{\mathbf{v}}_k'$ be an estimate of $\mathbf{H}$ with rank $K = K_0$, then

$$\hat{\mathbf{W}} = \left(\hat{w}_{ij}\right) = \mathbf{X} - \sum_{k=1}^{K_0} \hat{d}_k \hat{\mathbf{v}}_k \hat{\mathbf{v}}_k'$$

a test of the form

$$\tilde{T}_n = \frac{\sum_{i=1}^n \hat{\mathbf{w}}_{ij}}{\sqrt{\sum_{i=1}^n \hat{\mathbf{w}}_{ij}^2}}$$

can be used. But what if $\operatorname{diag}(\mathbf{X}) = 0$, that is, the network doesn't contain **self loops**?

Network without self loops In the absence of self loops, the observed data matrix $\mathbf{X}$ takes the form

$$\tilde{\mathbf{X}} - \operatorname{diag}\left(\tilde{\mathbf{X}}\right)$$

and thus $\hat{\mathbf{W}}$ actually estimates

$$\mathbf{W} - \operatorname{diag}\left(\tilde{\mathbf{X}}\right)$$

also, the entries of $\hat{\mathbf{W}}$ are all correlated: aggregating too many entries will cause too much noise accumulation and invalidate the normality assumption.

Rank inference via residual subsampling (RIRS) Define i.i.d. Bernoulli random variables Y_{ij} with $\Pr\left(Y_{ij} = 1\right) = \frac{1}{m}, 1 \leq i < j \leq n$, then the universal RIRS test takes the form

$$T_n = \frac{\sqrt{m}\sum_{i \neq j} \hat{\mathbf{w}}_{ij}\mathbf{Y}_{ij}}{\sqrt{2\sum_{i \neq j} \hat{\mathbf{w}}_{ij}^2}}$$

m controls (on average) how many entries of the residual matrix to use in calculating the statistic, and it needs to diverge fast for CLT, but not too fast to avoid noise accumulation

Conditions there are several conditions necessary for this method

References

Emmanuel Abbe. Community detection and stochastic block models: recent developments. *The Journal of Machine Learning Research*, 18(1):6446–6531, 2017.