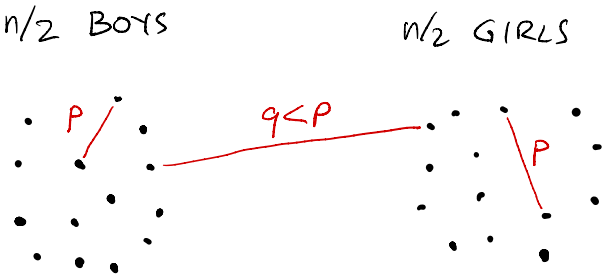
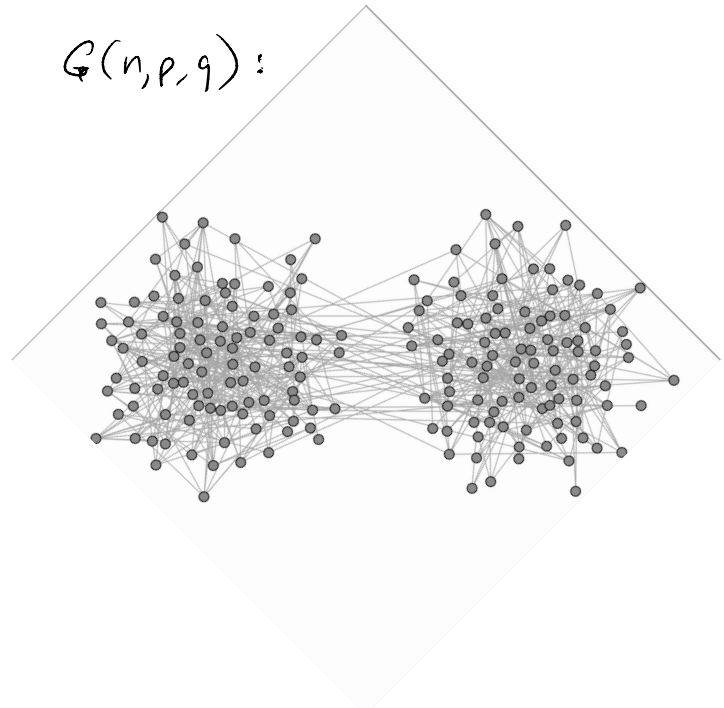
LECTURE 31

Last class: Stochastic Block Model $G(n,p,q)$:





COMMUNITY DETECTION PROBLEM:

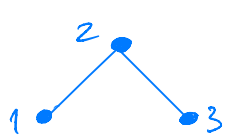
> Given a random graph drawn from a SBM, find the 2 communities.

Key to success: random matrix theory.

Def The adjacency matrix A of a graph $G=(V,E)$, $V = \{1,2,\dots,n\}$, has entries

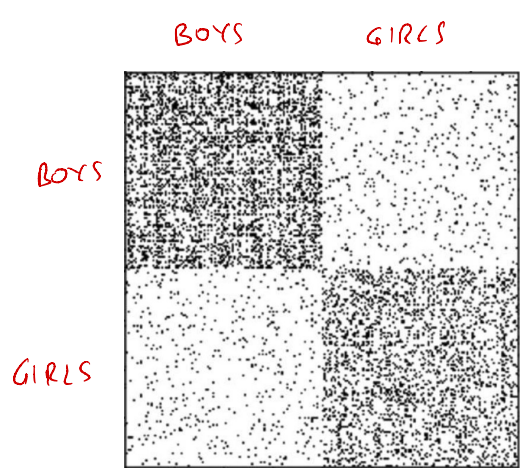
$$A_{ij} = \begin{cases} 1 & \text{if } (i,j) \in E \\ 0 & \text{if not} \end{cases}$$

- Ex:



$$A = \begin{array}{c|ccc} & 1 & 2 & 3 \\ \hline 1 & 0 & 1 & 0 \\ 2 & 1 & 0 & 1 \\ 3 & 0 & 1 & 0 \end{array}$$

- If G is a random graph (SBM), then A = symmetric random matrix with independent Bernoulli entries:



$= A$

HEURISTICS

$$A \approx \mathbb{E}A = \begin{array}{c} \text{BOYS} \quad \text{GIRLS} \\ \left[\begin{array}{cc|cc} 0 & p & q & q \\ p & 0 & q & q \\ \hline q & q & 0 & p \\ q & q & p & 0 \end{array}\right] \end{array} \begin{array}{l} \text{BOYS} \\ \\ \text{GIRLS} \end{array} \approx \left[\begin{array}{cc|cc} p & p & q & q \\ p & p & q & q \\ \hline q & q & p & p \\ q & q & p & p \end{array}\right] =: B.$$

matrix concentration

$$\lambda_1(B) = \left(\frac{p+q}{2}\right)n, \quad v_1(B) = \frac{1}{\sqrt{n}}\left[\begin{array}{c} 1 \\ 1 \\ \hline 1 \\ 1 \end{array}\right] \qquad \lambda_2(B) = \left(\frac{p-q}{2}\right)n, \quad v_2(B) = \left[\begin{array}{c} 1 \\ 1 \\ \hline -1 \\ -1 \end{array}\right] \quad (*)$$

coeffs. determine the community membership!

$\Rightarrow$ by Davis-Kahan inequality,

$$v_2(A) \approx v_2(B).$$

$\Rightarrow$ the signs of the coefficients of $v_2(A)$ recover the community membership.

each person has $\sim a$ friends in own community, b friends in the other.

THM (Community detection) Let $p = a/n$, $q = b/n$, where $b < a < 3a$

Assume $(a-b)^2 \geq C a \log n$

absolute constant, large enough

Then, with probability ≥ 0.99, the signs of the coefficients of $v_2(A)$, top eigenvector of the adjacency matrix of $G(n,p,q)$ recover the communities with 99% accuracy

i.e. with $< 0.01\,n$ misclassified vertices

Proof WTS $A \approx B$.

$$A - B = \underbrace{(A - \mathbb{E}A)}_{R} - \underbrace{(B - \mathbb{E}A)}_{p \cdot I_n}. \qquad (\circ)$$

$\|p \cdot I_n\| = p \le 1$

- $\|D\| = \left\| \begin{bmatrix} p & & & 0 \\ & p & & \\ & & q & \\ 0 & & & q \end{bmatrix} \right\| = p \le 1.$
- Decompose R entry-by-entry:
- $R = \sum_{i<j} R_{ij} Z_{ij}$ where $Z_{ij} = \begin{array}{c|cccc} & & i & j & \\ \hline & 0 & 0 & 0 & 0 \\ i & 0 & 0 & 1 & 0 \\ j & 0 & 1 & 0 & 0 \\ & 0 & 0 & 0 & 0 \end{array} = e_i e_i^T + e_j e_j^T$ and $e_i = \begin{array}{c|c} 0 & \\ 1 & i \\ 0 & \\ 0 & \end{array}$

  = sum of indep. random matrices

  canonical basis of $\mathbb{R}^n$

- Apply Matrix Bernstein inequality (lec. 30):

$$\boxed{\|R\| \le C\sigma\sqrt{\log n} + CK\log n \quad \text{with prob} \ge 0.99} \qquad (*)$$

where $\sigma^2 = \underbrace{\left\| \mathbb{E} \sum_{i<j} (R_{ij} Z_{ij})^2 \right\|}_{?}$; $\underbrace{\|Z_{ij}\|}_{\|(\text{DIY})\ 1} \le K \Rightarrow K = 1$

- $Z_{ij}^2 = \begin{array}{c|cccc} & & i & j & \\ \hline & 0 & 0 & 0 & 0 \\ i & 0 & 1 & 0 & 0 \\ j & 0 & 0 & 1 & 0 \\ & 0 & 0 & 0 & 0 \end{array} = e_i e_i^T + e_j e_j^T$ (Check: $(e_i e_i^T + e_j e_j^T)^2 = \dots \leftarrow$ DIY.)

renamed $i \leftrightarrow j$

$$\Rightarrow \sum_{i<j} (R_{ij} Z_{ij})^2 = \sum_{i<j} R_{ij}^2 (e_i e_i^T + e_j e_j^T) = \sum_{i<j} R_{ij}^2 e_i e_i^T + \sum_{j<i} \underbrace{R_{ji}^2}_{= R_{ij}^2 \text{ (symmetry)}} e_i e_i^T$$

$$= \sum_{i,j=1}^n R_{ij}^2 e_i e_i^T = \sum_{i=1}^n \left(\sum_{j=1}^n R_{ij}^2 \right) e_i e_i^T.$$

since $A_{ii} = 0 \Rightarrow R_{ii} = 0$

$$\Rightarrow \sigma^2 = \left\| \sum_{i=1}^n \underbrace{\left(\sum_{j=1}^n \mathbb{E} R_{ij}^2 \right)}_{} e_i e_i^T \right\| = \max_{i=1,\dots,n} \sum_{j=1}^n \mathbb{E} R_{ij}^2$$

diagonal matrix with these entries

$$\begin{array}{c|cc} & n/2 & n/2 \\ \hline i & \mathrm{Ber}(p) & \mathrm{Ber}(q) \\ & \mathrm{Ber}(q) & \mathrm{Ber}(p) \end{array}$$

$$= \frac{n}{2} \cdot \mathrm{Var}(\mathrm{Ber}(p)) + \frac{n}{2} \mathrm{Var}(\mathrm{Ber}(q))$$

$$\le n(p+q) = a + b \le 2a. \qquad (\text{since } a > b)$$

this notation hides an absolute constant factor

(*)

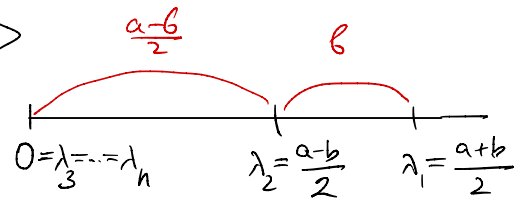
- $\xrightarrow[(*\,p.3)]{MBI} \quad \|R\| \lesssim \sqrt{(a+b)\log n} + \log n.$

- $\xrightarrow{(o\,p.3)} \|A-B\| \le \|R\| + 1 \lesssim \sqrt{a\log n} + \log n + 1 \lesssim \boxed{\sqrt{a\log n}}.$

(since $\log n \le \frac{(a-b)^2}{a} \le a$ by assumption.
$\Rightarrow \log n \le \sqrt{a\log n}$.)

(2) Davis–Kahan (HW 10 Problem 1) $\Rightarrow$

$$\|\underbrace{V_2(A)}_{u} - \underbrace{V_2(B)}_{v}\|_2 \le \frac{\|A-B\|}{gap_2(B)}$$



(*p.2)

$gap_2(B) = \min(\lambda_2 - 0, \lambda_1 - \lambda_2) = \min\left(\frac{a-b}{2}, b\right) = \boxed{\frac{a-b}{2}}$ (since $a \le 3b$ by assumption).

$\Rightarrow \|u-v\|_2 \lesssim \frac{\sqrt{a\log n}}{a-b} < \frac{1}{10}$ (choosing const C in the assumption large enough).

$\Downarrow$

(3) $\sum_{i=1}^{n}(u_i - v_i)^2 \le \frac{1}{100} \Rightarrow$ at most $\frac{n}{100}$ of $(u_i-v_i)^2$ can be $> \frac{1}{n}$

$\Rightarrow$ 99% of $(u_i - v_i)^2 \le \frac{1}{n}$

$\Updownarrow$

$|\sqrt{n}\,u_i - \underbrace{\sqrt{n}\,v_i}_{\pm 1}| < 1$ (*p.2)

$\Rightarrow \operatorname{sign}(u_i) = \operatorname{sign}(v_i)$.

<u>Summary</u>: With prob. ≥ 0.99 (*p.2), at least 99% of coefficients of $u = V_2(A)$ agree with the coeff. of $v = V_2(B)$, and thus correctly determine the community membership.

QED.