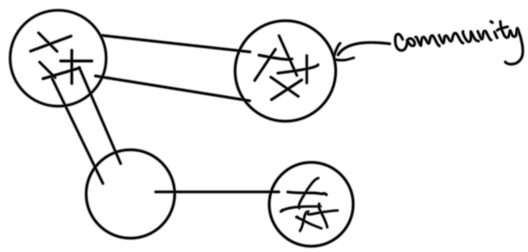
Lecture 6

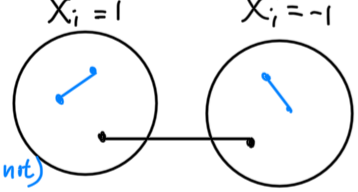
Community detection

Find clustering of population based on links between individuals.



Statistical model

determines the amount of info observed edges carry about community membership



Stochastic block model

unknown labels $X^\circ \in \{\pm 1\}^n$

parameters bias $\varepsilon \in [0,1]$ (same group or not)

degree $d \in \mathbb{N}_{\geq 1}$ (# of links) $\longrightarrow$ the average number of edges observed per node.

observe graph G

$$\forall i \neq j \in [n], \quad \mathbb{P}\{ij \in E(G)\} = \begin{cases} (1+\varepsilon)\frac{d}{n}, & X_i^\circ = X_j^\circ \text{ (in same community)} \\ (1-\varepsilon)\frac{d}{n}, & \text{otherwise} \end{cases}$$

Goal: compute $\hat{X}$ in poly time such that w.h.p.

$$\|\hat{X} - X^\circ\|_F^2 \leq 0.01 \cdot n^2, \quad \text{where } X^\circ = X^\circ (X^\circ)^T$$

$$\begin{array}{c|cc} & X_i^\circ = 1 & X_i^\circ = -1 \\ \hline X_j^\circ = 1 & 1 & -1 \\ X_j^\circ = -1 & -1 & 1 \end{array}$$

$\boxed{\text{Theorem Goal possible if } d \geq O\left(\frac{\log n}{\varepsilon^2}\right)}$

Strategy

- Compute Y with $\mathbb{E}[Y] = X^\circ = x_i^\circ \cdot x_j^\circ$
- Output best rank-1 approx. for Y

A = adj. matrix of G

$\mathbb{E}[A_{ij}] = (1 + \varepsilon \cdot X_i^\circ X_j^\circ)\frac{d}{n}$ for $i \neq j$

$Y_{ij} = \frac{n}{\varepsilon d}\left(A_{ij} - \frac{d}{n}\right), \quad Y_{ii} = 1$

$$Y_{ij} = \begin{cases} \frac{n}{\varepsilon d}\left(1 - \frac{d}{n}\right), & \text{if } ij \in E(G) \quad * A_{ij} = 1 \text{ edge exists} \\ -\frac{1}{\varepsilon}, & \text{if } ij \notin E(G) \quad * A_{ij} = 0 \text{ no edge} \end{cases}$$

Matrix Bernstein (variant)

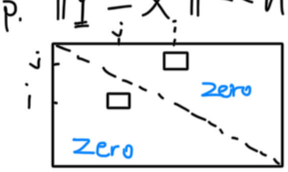
Let $Z_1, \cdots, Z_m$ indep. $n \times n$ matrices. Suppose $\mathbb{E}[Z_i] = 0$ and $\{Z_i\} = \{Z_i^T\}$ for all $i \in [m]$. Let $\sigma, b > 0$ satisfy $\|\sum_{i=1}^m \mathbb{E}[Z_i Z_i^T]\| \leq \sigma^2$ and $\|Z_i\| \leq b$, then w.h.p.,

$$\left\|\sum_{i=1}^m Z_i\right\| \lesssim \sigma\sqrt{\log n} + b \log n.$$

recall: $\|\hat{X} - X^\circ\|_F^2 \leq 8 \cdot \|Y - X^\circ\|^2$ * From matrix completion

to show if $d \geq O\left(\frac{\log n}{\varepsilon^2}\right)$, w.h.p. $\|Y - X^\circ\| \ll n$.

$$Y - X^\circ = \sum_{i \leq j} \underbrace{Z^{(ij)}}_{(Y_{ij} - X_{ij}^\circ)(e_i e_j^T + e_j e_i^T)} \qquad n \geq 100d$$



* e_i, e_j are canonical basis vectors

In 2D space, $e_1 = \binom{1}{0}$, $e_2 = \binom{0}{1}$

$\mathbb{E}[Z^{(ij)}] = 0$

$\|Z^{(ij)}\| = |Y_{ij} - X_{ij}^\circ| \leq 1 + \frac{1}{\varepsilon}\cdot\left(\frac{n}{d} - 1\right) = \frac{n}{\varepsilon d} = b$

$$Z^{(ij)}(Z^{(ij)})^T = (Y_{ij} - X_{ij})^2 \cdot (e_i e_i^T + e_j e_j^T)$$

$\ast\ \mathbb{E}[Y_{ij}] = X^*_{ij},\ (X^*_{ij})^2 = 1$

$Var(Y_{ij}) = \mathbb{E}[Y_{ij}^2] - (\mathbb{E}[Y_{ij}])^2$

$\mathbb{E}[(Y_{ij} - X_{ij})^2] = \mathbb{E}[Y_{ij}^2] - (X^*_{ij})^2 \leq \mathbb{E}[Y_{ij}^2]$

$\mathbb{E}[Y_{ij}^2] = P(edge)\cdot(\frac{n}{\varepsilon d}(1-\frac{d}{n}))^2 + (1-P(edge))\cdot(-\frac{1}{\varepsilon})^2$

$= P(edge)\cdot(\frac{n}{\varepsilon d}(1-\frac{d}{n}))^2 + (-\frac{1}{\varepsilon})^2 - (-\frac{1}{\varepsilon})^2 P(edge)$

$= \underbrace{P(edge)}_{=(1+\varepsilon X_i^0 X_j^0)\frac{d}{n}} \cdot (\frac{n^2}{\varepsilon^2 d^2}(1-\frac{d}{n})^2 - \frac{1}{\varepsilon^2}) + \frac{1}{\varepsilon^2}$

$$\mathbb{E}[(Y_{ij} - X^0_{ij})^2]$$
$$\leq (1+\varepsilon)\frac{d}{n}\cdot\frac{n^2}{\varepsilon^2 d^2}(1-\frac{d}{n})^2 + 1\cdot\frac{1}{\varepsilon^2}$$
$$\leq \frac{2n}{\varepsilon^2 d}$$
$$\mathbb{E}[Z^{(ij)}(Z^{(ij)})^T] \preceq \frac{2n}{\varepsilon^2 d}(e_i e_i^T + e_j e_j^T)$$
$$\sum_{i<j}\mathbb{E}[Z^{(ij)}(Z^{(ij)})^T] \preceq \frac{2n}{\varepsilon^2 d}\cdot(n-1)I_n$$
$$\longrightarrow \sigma = \sqrt{\frac{2n^2}{\varepsilon^2 d}}$$

$\ast\ (e_i e_j^T + e_j e_i^T)(e_i e_j^T + e_j e_i^T)^T$

$= e_i e_j^T e_j e_i^T + e_i \cancel{e_j^T e_i} e_j^T$

$+ e_j \cancel{e_i^T e_j} e_i^T + e_j e_i^T e_i e_j^T$

$= e_i e_i^T + e_j e_j^T$

$\sum_{i<j}(e_i e_i^T + e_j e_j^T)$

$- i=k, j>k,\ n-k$ times

$- i<k, j=k,\ k-1$ times

$\Rightarrow (n-k)+(k-1)\sum_{k=1}^{n} e_k e_k^T$

$= (n-1)\underbrace{\sum_{k=1}^{n} e_k e_k^T}_{I} = (n-1)I_n$

$$\xrightarrow{\text{Bernstein}} \text{w.h.p. } \|Y - X^0\| = \|\sum_{i<j} Z^{(ij)}\| \lesssim \sigma\sqrt{\log n} + b\cdot\log n$$

$d \geq O(\frac{\log n}{\varepsilon^2})$

$$\lesssim \frac{n}{\varepsilon\sqrt{d}}\sqrt{\log n} + \frac{n}{\varepsilon d}\cdot\log n$$
$$= \frac{n}{\varepsilon}\left(\underbrace{\sqrt{\frac{\log n}{d}}}_{\text{dominate}} + \frac{\log n}{d}\right)$$
$$\leq n\cdot\sqrt{\frac{\log n}{\varepsilon^2 d}}$$

recall
$$Y_{ij} = \begin{cases} 1, & \text{if } i=j \\ \frac{1}{\varepsilon}(\frac{n}{d}-1), & \text{if } ij \in E(G) \\ -\frac{1}{\varepsilon}, & \text{otherwise} \end{cases} \Longrightarrow \mathbb{E}[Y] = X^0$$

Class of estimators

For convex set C

$$\hat{X} = \hat{X}_C(G) = \operatorname{argmin}\{\|Y - X\|_F^2 \mid X \in C\}$$

$\ast$ The absolute maximum possible inner product between any matrix Z inside the combinatorial set C and target A.

$\|A\|_C = \max_{Z\in C}\langle Z, A\rangle$.

previous estimator

$C = \{uv^T \mid u, v \in \mathbb{R}^n\}$

$C = \{X \mid X \succeq 0, Tr(X) = n\} = conv\{uu^T \mid \|u\|^2 = n\}$

canonical choice

$C = \{xx^T \mid x \in \{\pm 1\}^n\}$

"inspired choice"

$C = \{uv^T \mid u, v \in \{\pm 1\}^n\}$

	$V_j=1$	$V_j=-1$
$u_i=1$	1	-1
$u_i=-1$	-1	1

$\|\hat{X} - X^0\|_F^2 \leq 2\langle\hat{X} - X^0, Y - X^0\rangle$

$\|\hat{X} - Y\|_F^2 = \|\hat{X} - X^0\|_F^2 + \cancel{\|Y - X^0\|_F^2} - 2\langle\hat{X} - X^0, Y - X^0\rangle$

analysis $\ast\ \|\hat{X} - X^0\|_F^2 \leq \|X^0 - Y\|_F^2$

$\|\hat{X} - Y\|_F^2 = \|(\hat{X} - X^0) - (Y - X^0)\|_F^2 \quad \leq \cancel{\|X^0 - Y\|_F^2}$

$$\|\hat{X} - X^0\|_F^2 \leq 2\langle Y - X^0, \hat{X} - X^0\rangle, \quad X^0 \in C$$
$$\leq 4\|Y - X^0\|_C$$

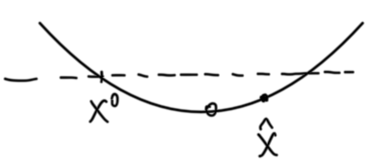
$\ast\ \|M\|_C = \max_{X\in C}|\langle X, M\rangle|$

$\ast\ \langle\hat{X} - X^0, Y - X^0\rangle$

$= \underbrace{\langle\hat{X}, Y - X^0\rangle}_{\leq\|Y - X^0\|_C} - \underbrace{\langle X^0, Y - X^0\rangle}_{\leq\|Y - X^0\|_C}$

$\leq 2\|Y - X^0\|_C$

$$f(X) = \|X - Y\|_F^2$$



$$f(X^0 + u) = f(X^0) + 2\langle Y - X^0, u\rangle + \|u\|_F^2$$

Claim

$\forall M, C \quad \|M\|_C = \|M\|_{conv(C)}$

$\|M\|_{cut} = \|M\|_{C_{cut}} = \max_{u,v\in\{\pm1\}^n} \langle \underbrace{uv^T, M}_{=\langle u, Mv\rangle} \rangle$

<u>to show</u> $\|\underline{Y} - X^\circ\|_{cut} \le 0.0001 \cdot n^2$ if $d \gtrsim O(1/\varepsilon^2)$, w.h.p.

$\le \frac{n^2}{\varepsilon\sqrt{d}}$

<u>proof</u> $\forall u,v \in \{\pm1\}^n$

$\mathbb{P}\left(\langle u, (\underline{Y} - X^\circ)v\rangle > \frac{n^2}{\varepsilon\sqrt{d}}\right) \le 2^{-3n}$

$\xrightarrow{\text{union bound}} \mathbb{P}\left(\exists u,v \in \{\pm1\}^n > \frac{n^2}{\varepsilon\sqrt{d}}\right) \le 2^{2n}\cdot 2^{-3n} = 2^{-n}$

<u>Grothendieck's inequality</u>

$-A \in \mathbb{R}^{n\times n}$

$-\max \langle x, Ay\rangle = \langle A, xy^T\rangle = \mathrm{Tr}\, A\underbrace{(xy^T)^T}_{yx^T}$ NP-hard

$-x, y \in \{\pm1\}^n$ "quadratic programming" $\exists$ polytime const. factor approx. alg.

$$\underbrace{\max_{\substack{W\in\mathbb{R}^{2n\times 2n}\\ W\succeq 0\\ W_{11}=\cdots=W_{2n\,2n}=1}} \left\langle W, \begin{pmatrix} 0 & A \\ 0 & 0\end{pmatrix}\right\rangle}_{\|A\|_G}, \quad W = \begin{pmatrix} W^{(1,1)} & W^{(1,2)} \\ W^{(2,1)} & W^{(2,2)}\end{pmatrix}$$

$$\longrightarrow \left\langle W, \begin{pmatrix} 0 & A \\ 0 & 0\end{pmatrix}\right\rangle = \langle W^{(1,2)}, A\rangle$$

$$W = \begin{pmatrix} x \\ y\end{pmatrix}\begin{pmatrix} x \\ y\end{pmatrix}^T = \begin{pmatrix} xx^T & xy^T \\ yx^T & yy^T\end{pmatrix}$$

$\rightarrow \forall A, \ \|A\|_{cut} \le \|A\|_G$

Theorem

$\exists K_G > 1, \ \forall A \ \ \|A\|_{cut} \le \|A\|_G \le K_G \cdot \|A\|_{cut}$

<u>Claim:</u> $\langle x, Ay\rangle \le \|x\|_\infty \cdot \|y\|_\infty \cdot \|A\|_G, \ \forall A, x, y$

$\forall x, y, \ \|x\|_\infty = 1, \ \|y\|_\infty = 1 \longrightarrow \langle x, Ay\rangle \le \|A\|_{cut}$

find $x' \in \{\pm1\}^n$, $\langle x, Ay\rangle \le \langle x', Ay\rangle$

$\le \langle x', Ay'\rangle$

$\le \|A\|_{cut}$

<u>Claim:</u> $\forall$ jointly distributed random vectors $\underline{x}, \underline{y}$

$\mathbb{E}[\langle \underline{x}, A\underline{y}\rangle] = \max\{\mathbb{E}[\underline{x}_1^2], \cdots, \mathbb{E}[\underline{x}_n^2], \mathbb{E}[\underline{y}_1^2], \cdots, \mathbb{E}[\underline{y}_n^2]\}\cdot \|A\|_G$.

and equality achieved by jointly Gaussian vectors $\underline{x}, \underline{y}$ with

$\mathbb{E}[\underline{x}_1^2], \cdots, \mathbb{E}[\underline{x}_n^2], \mathbb{E}[\underline{y}_1^2], \cdots, \mathbb{E}[\underline{y}_n^2] = 1$, $\langle x, Ay\rangle$

$W = \mathbb{E}\left[\binom{x}{y}\binom{x}{y}^T\right]$, $W^{(1,2)} = \mathbb{E}[x y^T]$, $\langle A, W^{(1,2)} \rangle = \mathbb{E}\left[\overbrace{\langle A, x y^T \rangle}^{\langle x, Ay \rangle}\right]$

Consider $w \sim N(0, W)$, $w = (x, y)$, $W = \mathbb{E}[w w^T] = \mathbb{E}\begin{bmatrix} x x^T & x y^T \\ y x^T & y y^T \end{bmatrix}$

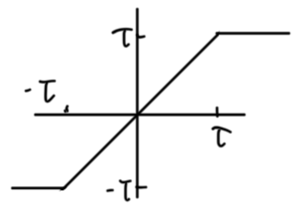
Proof (inequality)

Take jointly Gaussian vectors x, y, we have

- $\mathbb{E}[\langle x, Ay \rangle] = \|A\|_G$
- $\mathbb{E}[x_i^2] = \mathbb{E}[y_i^2] = 1$

Choose $\tau > 1$ large enough, $\overbrace{x_i^{\le\tau} = \mathrm{clamp}_\tau(x_i), \quad x_i^{>\tau} = x_i - x_i^{\le\tau}}^{x = x^{\le\tau} + x^{>\tau}}$

same for y_i



$$\|A\|_G = \mathbb{E}[\langle x, Ay \rangle]$$
$$= \underbrace{\mathbb{E}[\langle x^{\le\tau}, A y^{\le\tau} \rangle]}_{\le \tau^2 \cdot \|A\|_{cut}} + \underbrace{\mathbb{E}[\langle x^{>\tau}, A y^{\le\tau} \rangle] + \mathbb{E}[\langle x^{\le\tau}, A y^{>\tau} \rangle] + \mathbb{E}[\langle x^{>\tau}, A y^{>\tau} \rangle]}_{\le 3 e^{\tau^2/4} \cdot \|A\|_G}$$

To show: $\mathbb{E}[(x_i^{>\tau})^2] \le e^{-\tau^2/2}$ (also for y)

$$\mathbb{E}[(x_i^{>\tau})^2] = \mathbb{E}[(|x_i| - \tau)_+^2]$$
$$= 2\int_\tau^\infty (x-\tau)^2 \gamma(x)\, dx \qquad \gamma(x) \propto e^{-x^2/2}$$
$$\le \underbrace{2\int_\tau^\infty (x-\tau)^2 \gamma(x-\tau)\, dx}_{=1} \cdot \max_{x \ge \tau} \underbrace{\frac{\gamma(x)}{\gamma(x-\tau)}}_{= \frac{e^{-x^2/2}}{e^{-(x-\tau)^2/2}} = e^{-\frac{1}{2}\overbrace{(x^2 - (x-\tau)^2)}^{2x\tau - \tau^2 \ge \tau^2}}}$$

Recall unknown $X^0 = (x^0)(x^0)^T$, $x^0 \in \{\pm 1\}^n$

$$Y_{ij} = \begin{cases} 1 & \text{, if } i=j \\ \frac{1}{\varepsilon}\left(\frac{n}{d} - 1\right) & \text{, with prob. } (1 + \varepsilon X^0_{ij}) \cdot \frac{d}{n} \\ -\frac{1}{\varepsilon} & \text{, otherwise} \end{cases} \qquad * \ \mathbb{E}[Y] = X^0$$

$\hat{X} = \mathrm{argmin}\{\|Y - X\|_F^2 \mid X \in C\}$

- $C_{cut} = \{x y^T \mid x, y \in \{\pm 1\}^n\}$
- $C_G = \{W^{(1,2)} \mid W \succeq 0, W_{11} = \cdots = W_{2n2n} = 1\}$

Claim 1: $\|\hat{X} - X^0\|_F^2 \le 4\|Y - X^0\|_G$

Claim 2: $\|Y - X^0\|_{cut} \le 0.001 \cdot n^2$ for $d \gg 1/\varepsilon^2$ w.h.p.

Proof: (claim 1)

Together with Grotendieck ineq. $\|A\|_G \le 2 \cdot \|A\|_{cut}$

$$\|\hat{X} - X^0\|_F^2 \le 4\|Y - X^0\|_G \le 8\|Y - X^0\|_{cut} \le 8 \cdot 0.001 n^2 \le 0.01 n^2$$

$$\|\hat{X} - X^0\|_F^2 = \underbrace{\|Y - \hat{X}\|_F^2 - \|Y - X^0\|_F^2}_{\le 0} - 2\langle X^0 - \hat{X}, Y - X^0 \rangle \quad * \ Y - \hat{X} = (Y - X^0) + X^0 - \hat{X}$$

$* \ C_G = -C_G$

$$\le 2|\langle X^0 - \hat{X}, Y - X^0 \rangle|$$
$$\le 2(|\langle X^0, Y - X^0 \rangle| + |\langle \hat{X}, Y - X^0 \rangle|) \quad * \ X^0, \hat{X} \in C_G$$

$$\le 4 \| \underline{Y} - X^0 \|_G$$

Proof. (claim2)

Recall Bernstein inequality, $|\underline{z}_1 + \cdots + \underline{z}_N| \le \sigma \cdot \sqrt{\log(1/\delta)} + b \cdot \log(1/\delta)$, where $\underline{z}_1, \cdots, \underline{z}_n$ indep.

$\mathbb{E}[\underline{z}_i] = 0$, $\sigma, b > 0$ with $\mathbb{E}[\underline{z}_i^2] \le \sigma^2$, $|\underline{z}_i| \le b$, with prob. $1-\delta$.

Let $u, v \in \{\pm 1\}^n$

$$\langle u, (\underline{Y} - X^0) v \rangle = \sum_{i \ne j} u_i (\underline{Y} - X^0)_{ij} v_j$$

* if $i=j$, the matrix is 0.

$$= \sum_{i<j} \underbrace{(\underline{Y} - X^0)_{ij} (u_i v_j + u_j v_i)}_{= \underline{z}_{ij}}$$

* $\mathbb{E}[\underline{z}_{ij}] = 0$ because $Y_{ij} = X^0_{ij}$

* $|\underline{z}_{ij}| \le 2\left(\frac{1}{\varepsilon}\left(\frac{n}{d} - 1\right) + 1\right)$

$\le \frac{2}{\varepsilon} \cdot \frac{n}{d} = b$

* $\mathbb{E}[\underline{z}_{ij}^2] \le \frac{3n}{\varepsilon^2 \cdot d} \rightarrow \sigma^2 = n^2 \cdot \frac{3n}{\varepsilon^2 \cdot d}$

* $\delta = 2^{-3n}$

$\rightarrow \mathbb{P}\left(\| \underline{Y} - X^0 \|_{cut} > \sigma\sqrt{\log(1/\delta)} + b \cdot \log(1/\delta)\right) \le 2^{2n} \cdot \delta = 2^{-n}$

$$\frac{n^{3/2}}{\varepsilon\sqrt{d}} \sqrt{n} + \frac{n^2}{\varepsilon d} = n^2 \underbrace{\left(\frac{1}{\varepsilon\sqrt{d}} + \frac{1}{\varepsilon d}\right)}_{\le \frac{2}{\varepsilon\sqrt{d}}}$$