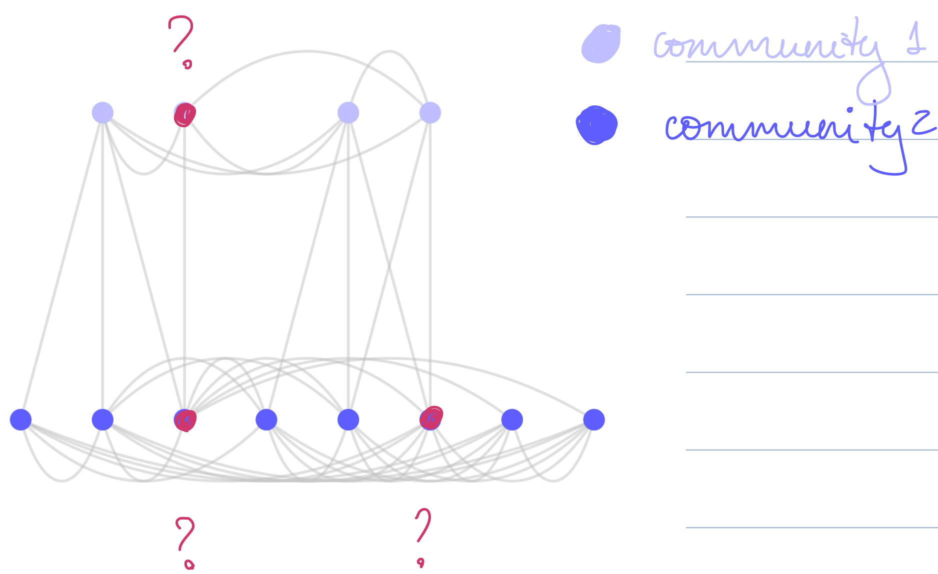
Lecture 6 EN.553.744 Prof. Luana Ruiz

Today: community detection (topic of HW1, Wed.)

- → recap SBM
- → spectral clustering
- → spectral embeddings
- → contextual SBM
- → GNNs for community detection



▶ The stochastic block model (recap)

An n-node SBM graph with C communities is given by:

$$P = YBY^T$$

$$A \sim P \qquad \text{(adjacency)}$$

where: •) $Y \in \{0,1\}^{n \times C}$ is community assignment matrix; if $Y_{ic} = 1$, node i belongs to comm. c

•) B is the matrix of intra & inter-comm. probabilities; $B_{c_1 c_2}$ is the edge probability for a node pair (i,j) s.t. $Y_{ic_1} = 1$, $Y_{jc_2} = 1$

(or symmetric → SSBM!)

→ We say the SBM is balanced when all comms. have the same size (n/C); it is unbalanced otherwise.

→ While B can take any value (as long as it's symmetric), often:

$$B_{c_1, c_2} = \begin{cases} p & \text{if } c_1 = c_2 \\ q & \text{if } c_1 \neq c_2 \end{cases}$$

- I.e., the probability of connecting to nodes in the same community is the same for all comms;
- the probability of connecting across communities is the same for all pairs c_1, c_2 w/ $c_1 \neq c_2$

→ Consider the 2-community balanced SBM w/ intra & intercommunity parameters p & q

$$B = \begin{bmatrix} p & q \\ q & p \end{bmatrix}$$

For an n-node graph, suppose the nodes are labeled such that the first $\frac{n}{2}$ are in community 1, and the remaining $\frac{n}{2}$ nodes are in community 2.

- Then,

$$Y = \begin{array}{c} \left[\begin{array}{cc} 1 & 0 \\ 1 & 0 \\ \vdots & \vdots \\ 1 & 0 \\ 0 & 1 \\ \vdots & \vdots \\ 0 & 1 \end{array}\right] \end{array} \begin{array}{l} \left.\vphantom{\begin{array}{c}1\\1\\\vdots\\1\end{array}}\right] \frac{n}{2} \quad \text{community 1} \\ \left.\vphantom{\begin{array}{c}0\\\vdots\\0\end{array}}\right] \frac{n}{2} \quad \text{community 2} \end{array}$$

and P is given by:

$$
P = \left[\begin{array}{ccc|ccc}
p & \cdots & p & q & \cdots & q \\
\vdots & & & \vdots & & \\
p & \cdots & p & q & \cdots & q \\
\hline
q & \cdots & q & p & \cdots & p \\
\vdots & & \vdots & \vdots & & \vdots \\
q & \cdots & q & p & \cdots & p
\end{array}\right]
\begin{array}{l}
\left.\vphantom{\begin{array}{c}p\\ \vdots\\ p\end{array}}\right\} n/2 \\
\left.\vphantom{\begin{array}{c}q\\ \vdots\\ q\end{array}}\right\} n/2
\end{array}
$$

$$
\underbrace{\hphantom{q \cdots q}}_{n/2} \quad \underbrace{\hphantom{p \cdots p}}_{n/2}
$$

- Recall $A \sim P$, i.e.,

$$
A_{ij} = \begin{cases} 1 & \text{with probability } P_{ij} \\ 0 & \text{with prob. } 1 - P_{ij} \end{cases}
$$

- Hence, $\mathbb{E}(A) = P$. Let's compute the eigenvalues & eigenvectors of $P / \mathbb{E}(A)$

$$
v_1 = \begin{bmatrix} \frac{1}{\sqrt{n}} \\ \vdots \\ \frac{1}{\sqrt{n}} \end{bmatrix}
$$

since the rows of P all have the same sum

?

$$\begin{pmatrix} \boxed{P} & \boxed{Q} \\ \boxed{Q} & \boxed{P} \end{pmatrix} \begin{pmatrix} \frac{1}{\sqrt{n}} \\ \vdots \\ \frac{1}{\sqrt{n}} \end{pmatrix} = \frac{(p+q)}{2} \cdot \begin{pmatrix} \frac{1}{\sqrt{n}} \\ \vdots \\ \frac{1}{\sqrt{n}} \end{pmatrix}$$

$$\Rightarrow \lambda_1 = p+q$$

Further, note:

$$\begin{pmatrix} \boxed{P} & \boxed{Q} \\ \boxed{Q} & \boxed{P} \end{pmatrix} = \frac{(p+q)}{2} \overbrace{\begin{bmatrix} \boxed{1} \end{bmatrix}}^{\mathbb{1}\,\mathbb{1}^T} + \frac{(p-q)}{2} \begin{pmatrix} \boxed{1} & \boxed{-1} \\ \boxed{-1} & \boxed{1} \end{pmatrix}$$

$$= \frac{(p+q)}{2} \begin{bmatrix} 1 \\ \vdots \\ 1 \end{bmatrix} \begin{bmatrix} 1 & \dots & 1 \end{bmatrix}$$

$$+ \frac{(p-q)}{2} \begin{bmatrix} 1 \\ \vdots \\ 1 \\ -1 \\ \vdots \\ -1 \end{bmatrix} \begin{bmatrix} \overbrace{1 \dots 1}^{n/2} & \overbrace{-1 \dots -1}^{n/2} \end{bmatrix}$$

$$\Rightarrow v_2 = [\underbrace{\tfrac{1}{\sqrt{n}} \cdots \tfrac{1}{\sqrt{n}}}_{n/2}\ \underbrace{-\tfrac{1}{\sqrt{n}} \cdots -\tfrac{1}{\sqrt{n}}}_{n/2}]$$

$$\lambda_2 = \frac{p-q}{2}$$

The second eigenvector associated w/ the second largest eigenvalue (in abs. value) of $E(A) = P$

reveals the comm. assignment

This is the intuition behind spectral clustering; For an arbitrary graph with adjacency matrix A, we assume $A = A_{SBM} + \mathcal{E}$, where $\mathcal{E}$ is random noise w/ $E(\mathcal{E}) = 0$; since $E(A) = E(A_{SBM}) = P$, we can estimate the community assignment from the eigenvectors of A, which is an unbiased estimator of P.

▶ Spectral clustering

↳ what happens for $c > 2$?

- the community assignment is a linear combination of the top c eigenvectors (in absolute value)

↳ suppose we are given A and C. How to estimate Y?

The spectral clustering algorithm:

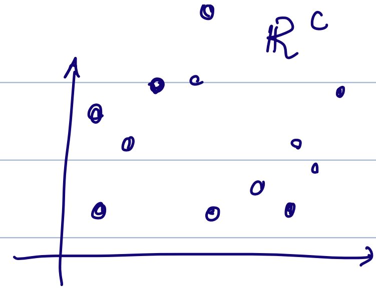
1. Diagonalize $A = V \Lambda V^T$

2. Order eigenvectors by decreasing eigenvalue magnitude

$$V_C = \begin{bmatrix} | & | & | & & | \\ v_1 & v_2 & v_3 & \cdots & v_c \\ | & | & | & & | \end{bmatrix} = \begin{bmatrix} \text{---} & u_1 & \text{---} \\ \text{---} & u_2 & \text{---} \\ & \vdots & \\ \text{---} & u_n & \text{---} \end{bmatrix} \quad u_j \in \mathbb{R}^c$$

$v_i \in \mathbb{R}^n$

u_j is the embedding of node j in $\mathbb{R}^c$



(3) Cluster the u_i into C groups.
There are multiple clustering algos; most common is K-means clustering

▶ Semi-supervised learning with spectral embeddings

Spectral clustering is fully unsupervised – it only assumes knowledge of A.

In the case where the labels of a node subset $\mathcal{J} \subset \mathcal{V}$ are known, these can be used to enhance predictive power.

Explicitly, we aim to solve the problem:

(*)

indicator fcn

$$\min_{f} \sum_{i \in \mathcal{J}} \mathbb{I}\left([f(A)]_i = y_i \right)$$

some parametric function

one-hot comm. vector

$[y_{ic}] = 1$ if i is in comm. c

(in practice the indicator/0-1 loss is substituted for the cross-entropy loss)

Spectral embeddings: $f(A) = \sigma(V_c W)$
(1-layer)

$W \in \mathbb{R}^{c \times c}$ linear map

σ nonlinearity

↳ can be thought of as an FCNN on the c-dimensional embeddings $u_1, \ldots, u_n$ of nodes in V

To find f^*, i.e., W^*, we solve (*) using gradient-descent methods

▶ Information-theoretic threshold

→ What happens when $p \approx q$ in the balanced SBM with $c = 2$?

When $p = q$, we actually have an Erdős-Rényi graph, in which the edge probability is constant for all nodes; there are no communities to distinguish

But even when $p \neq q$, there is a region around $p = q$ where detection of communities is impossible in an information theoretic sense:

- Let $SNR = \frac{(p-q)^2}{2(p+q)}$ (signal-to-noise ratio)

degree discrepancy in ≠ subgraphs

~ avg degree ("noise")

- If $SNR < 1/n$, almost exact recovery

$$\left\{ P\left(\frac{1}{n} \mathbb{I}\left([f(A)]_i = y_i \right) = 1 - o(1) \right) = 1 - o(1) \right\}$$

is impossible. Even with infinite time & resources, there is no algorithm that can recover the true communities given A.

For the pf, check the works of Massoulié (2014)
Mossel (2014)
Abbe (survey)

E.g.: sparse graphs

$p = \frac{a}{n}$; $q = \frac{b}{n}$ $\qquad \frac{|E|}{n^2} \longrightarrow 0$ as $n \to \infty$

(avg. degree vanishes)

$$SNR = \frac{\left(\frac{a}{n} - \frac{b}{n}\right)^2}{2\left(\frac{a+b}{n}\right)} = \boxed{\frac{1}{2n} \frac{(a-b)^2}{(a+b)}}$$