Community Detection from a Random Graphs perspective

Souvik Dhara

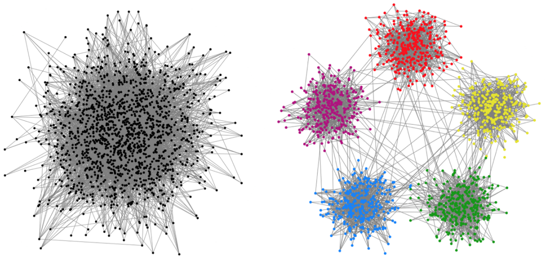
Purdue University, IE

Young European Probabilists Workshop 2024, EURANDOM

Day 1

Clustering/Community Detection Problem

➢ Communities are densely connected parts of a network



[Abbe '18]

➢ Divide the graph into communities from unlabelled graph. This is an unsupervised learning task

Applications

Community detection is a central problem in machine learning and data mining...

Numerous applications in ...

- ➜ Recommender system [Wu, Xu, Srikant, Massoulié, Lelarge, & Hajek '15]
- ➜ Webpage sorting [Kumar, Raghavan, Rajagopalan, & Tomkins '99]
- ➜ Functionalilty of Human Brain [Martinet, Kramer, Viles, Perkins, Spencer, Chu, Cash & Kolaczyk '20]
- ➜ Social networks [Goldenberg, Zheng, Fienberg, & Airoldi '10]

➢ Huge literature has developed in past two decades from TCS, ML, Stats..

Objective of this minicourse

Two high-level questions:

1. When is recovering clusters possible/impossible? (Information theoretic limit)
2. Are there fast and optimal algorithms to recover clusters

➢ Will dive deep into **a sharp phase transition for exact recovery**

References

1. E. Abbe, Community Detection and Stochastic Block Models: Recent Developments. *Journal of Machine Learning Research* (2018)

Stochastic Block Model

Parameters.

- Number of communities: $k \geqslant 2$
- Communities sizes: $p = (p_1, \ldots, p_k)$, a probability vector with $p_i > 0$
- Probability matrix: Q, a $k \times k$ symmetric matrix
- Sparsity ρ_n

Stochastic Block Model

Parameters.

- ➢ Number of communities: $k \geqslant 2$
- ➢ Communities sizes: $p = (p_1, \ldots, p_k)$, a probability vector with $p_i > 0$
- ➢ Probability matrix: Q, a $k \times k$ symmetric matrix
- ➢ Sparsity ρ_n

Generative Model. $\mathrm{SBM}(n, k, p, \rho_n Q)$

- ➢ Generate σ: $\sigma(u) \overset{iid}{\sim} p$ for each vertex $u \in [n]$
- ➢ Generate G: Add edge $\{u, v\}$ with probability $\rho_n Q_{\sigma(u)\sigma(v)}$ (independent)

➡ A Symmetric SBM corresponds to the case where $p_i = \frac{1}{k}$ and $Q_{ij} = a$ if $i = j$ and $Q_{ij} = b$ if $i \neq j$

Stochastic Block Model

Parameters.

➢ Number of communities: $k \geqslant 2$

➢ Communities sizes: $p = (p_1, \ldots, p_k)$, a probability vector with $p_i > 0$

➢ Probability matrix: Q, a $k \times k$ symmetric matrix

➢ Sparsity ρ_n

Generative Model. $\mathrm{SBM}(n, k, p, \rho_n Q)$

➢ Generate σ: $\sigma(u) \overset{iid}{\sim} p$ for each vertex $u \in [n]$

➢ Generate G: Add edge $\{u, v\}$ with probability $\rho_n Q_{\sigma(u)\sigma(v)}$ (independent)

➡ A Symmetric SBM corresponds to the case where $p_i = \frac{1}{k}$ and $Q_{ij} = a$ if $i = j$ and $Q_{ij} = b$ if $i \neq j$

Statistical Task. Suppose $(G, \sigma) \sim \mathrm{SBM}(n, K, p, \rho_n Q)$. We observe $G = g$, but σ is unknown.

$$\boxed{\text{Find a “good” estimator } \hat{\sigma}}$$

Different modes of recovery

Agreement: $A(\sigma, \hat{\sigma}) := \max_{\pi \in S_k} \frac{1}{n} \sum_{u=1}^{n} \mathbb{1}\{\sigma(u) = \pi(\hat{\sigma}(u))\}$, where S_k is the set of permutations of $[k] := \{1, \ldots, k\}$

Partition associated with $\sigma, \hat{\sigma}$: $\Sigma_i = \{u : \sigma(u) = i\}$, and $\Sigma = \{\Sigma_1, \ldots, \Sigma_k\}$, and $\hat{\Sigma}$ is defined similarly for $\hat{\sigma}$

Different modes of recovery

Agreement: $A(\sigma, \hat{\sigma}) := \max_{\pi \in S_k} \frac{1}{n} \sum_{u=1}^{n} \mathbb{1}\{\sigma(u) = \pi(\hat{\sigma}(u))\}$, where S_k is the set of permutations of $[k] := \{1, \ldots, k\}$

Partition associated with $\sigma, \hat{\sigma}$: $\Sigma_i = \{u : \sigma(u) = i\}$, and $\Sigma = \{\Sigma_1, \ldots, \Sigma_k\}$, and $\hat{\Sigma}$ is defined similarly for $\hat{\sigma}$

Exact Recovery: $\lim_{n\to\infty} \mathbb{P}(A(\sigma, \hat{\sigma}) = 1) = 1$, or $\lim_{n\to\infty} \mathbb{P}(\Sigma = \hat{\Sigma}) = 1$

Sharp phase transition for $\rho_n = \frac{\log n}{n}$

Almost Exact Recovery: $\lim_{n\to\infty} \mathbb{P}(A(\sigma, \hat{\sigma}) \geqslant 1 - \varepsilon_n) = 1$ for some $\varepsilon_n \to 0$

Possibility depends on $n\rho_n \to \infty$ *or not*

Partial Recovery: Definition slightly tricky, but for symmetric SBMs...

$$\lim_{n\to\infty} \mathbb{P}(A(\sigma, \hat{\sigma}) \geqslant \alpha) \text{ for some } \alpha \in \left(\frac{1}{k}, 1\right)$$

Sharp phase transition for $\rho_n = \frac{1}{n}$

Exact Recovery

Maximum A Posteriori (MAP) Estimator

➢ Maximum A Posteriori (MAP) estimator, denoted by $\hat{\Sigma}_{\text{MAP}}$, solves the following maximization problem:

$$\text{maximize } \mathbb{P}(\hat{\Sigma} = S \mid G = g) \quad \text{over all partitions } S = \{S_1, \ldots, S_k\}$$

If there are multiple maximizers, pick one uniformly among them

Maximum A Posteriori (MAP) Estimator

➢ Maximum A Posteriori (MAP) estimator, denoted by $\hat{\Sigma}_{\text{MAP}}$, solves the following maximization problem:

$$\text{maximize } \mathbb{P}(\hat{\Sigma} = S \mid G = g) \quad \text{over all partitions } S = \{S_1, \ldots, S_k\}$$

If there are multiple maximizers, pick one uniformly among them

Lemma: For any estimator $\hat{\sigma}$, $\mathbb{P}(\hat{\Sigma}_{\text{MAP}} \neq \Sigma) \leqslant \mathbb{P}(\hat{\Sigma} \neq \Sigma)$

Maximum A Posteriori (MAP) Estimator

➢ Maximum A Posteriori (MAP) estimator, denoted by $\hat{\Sigma}_{\text{MAP}}$, solves the following maximization problem:

$$\text{maximize } \mathbb{P}(\hat{\Sigma} = S \mid G = g) \quad \text{over all partitions } S = \{S_1, \ldots, S_k\}$$

If there are multiple maximizers, pick one uniformly among them

Lemma: For any estimator $\hat{\sigma}$, $\mathbb{P}(\hat{\Sigma}_{\text{MAP}} \neq \Sigma) \leqslant \mathbb{P}(\hat{\Sigma} \neq \Sigma)$

Hence, Exact recovery is possible $\Longleftrightarrow$ $\hat{\Sigma}_{\text{MAP}}$ succeeds whp

Maximum A Posteriori (MAP) Estimator

➢ Maximum A Posteriori (MAP) estimator, denoted by $\hat{\Sigma}_{\text{MAP}}$, solves the following maximization problem:

$$\text{maximize } \mathbb{P}(\hat{\Sigma} = S \mid G = g) \quad \text{over all partitions } S = \{S_1, \ldots, S_k\}$$

If there are multiple maximizers, pick one uniformly among them

Lemma: For any estimator $\hat{\sigma}$, $\mathbb{P}(\hat{\Sigma}_{\text{MAP}} \neq \Sigma) \leqslant \mathbb{P}(\hat{\Sigma} \neq \Sigma)$

$$\text{Hence, Exact recovery is possible} \iff \hat{\Sigma}_{\text{MAP}} \text{ succeeds whp}$$

Let's try to find $\hat{\Sigma}_{\text{MAP}}$ in a simple scenario...

Maximum A Posteriori (MAP) Estimator

➢ Maximum A Posteriori (MAP) estimator, denoted by $\hat{\Sigma}_{\mathrm{MAP}}$, solves the following maximization problem:

$$\text{maximize } \mathbb{P}(\hat{\Sigma} = S \mid G = g) \quad \text{over all partitions } S = \{S_1, \ldots, S_k\}$$

If there are multiple maximizers, pick one uniformly among them

Lemma: For any estimator $\hat{\sigma}$, $\mathbb{P}(\hat{\Sigma}_{\mathrm{MAP}} \neq \Sigma) \leqslant \mathbb{P}(\hat{\Sigma} \neq \Sigma)$

Hence, Exact recovery is possible $\iff$ $\hat{\Sigma}_{\mathrm{MAP}}$ succeeds whp

Let's try to find $\hat{\Sigma}_{\mathrm{MAP}}$ in a simple scenario...

1. $\hat{\Sigma}_{\mathrm{MAP}}$ is computationally intractable
2. The distribution of $\hat{\Sigma}_{\mathrm{MAP}}$ is difficult to analyze

Genie-based Estimator

Genie-based (hypothetical) estimator:

To estimate $\sigma(u)$

Genie-based Estimator

Genie-based (hypothetical) estimator:

To estimate $\sigma(u)$

➢ Suppose genie tells us $\sigma_{-u} := (\sigma(v))_{v \neq u}$

Genie-based Estimator

Genie-based (hypothetical) estimator:

To estimate $\sigma(u)$

- ➢ Suppose genie tells us $\sigma_{-u} := (\sigma(v))_{v \neq u}$
- ➢ Compute MAP with genie's added info

$$\hat{\sigma}_{\text{genie}}(u) := \operatorname{argmax}_{i \in [k]} \mathbb{P}(\sigma(u) = i \mid G = g, \sigma_{-u})$$

Genie-based Estimator

Genie-based (hypothetical) estimator:

To estimate $\sigma(u)$

➢ Suppose genie tells us $\sigma_{-u} := (\sigma(v))_{v \neq u}$

➢ Compute MAP with genie's added info

$$\begin{aligned}\hat{\sigma}_{\text{genie}}(u) &:= \operatorname{argmax}_{i \in [k]} \mathbb{P}(\sigma(u) = i \mid G = g, \sigma_{-u}) \\ &= \operatorname{argmax}_{i \in [k]} \mathbb{P}(G = g \mid \sigma(u) = i, \sigma_{-u}) p_i\end{aligned}$$

Genie-based Estimator

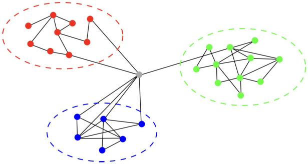
Genie-based (hypothetical) estimator:

To estimate $\sigma(u)$

➢ Suppose genie tells us $\sigma_{-u} := (\sigma(v))_{v \neq u}$

➢ Compute MAP with genie's added info

$$\begin{aligned}\hat{\sigma}_{\text{genie}}(u) &:= \operatorname{argmax}_{i\in[k]} \mathbb{P}(\sigma(u) = i \mid G = g, \sigma_{-u}) \\ &= \operatorname{argmax}_{i\in[k]} \mathbb{P}(G = g \mid \sigma(u) = i, \sigma_{-u}) p_i \\ &= \operatorname{argmax}_{i\in[k]} \mathbb{P}(D(u) = d \mid \sigma(u) = i, \sigma_{-u}) p_i\end{aligned}$$



➡ $D(u) = (D_1(u), \dots, D_k(u))$ is degree profile
$D_j(u)$:= # edges from u to community j

Error probability for the genie-based estimator

Fact: $\frac{1}{k-1}\tilde{P}_e \leqslant \mathbb{P}(\hat{\sigma}_{\text{genie}}(u) \neq \sigma(u) \mid \sigma_{-u}) \leqslant \tilde{P}_e$, where

$$\tilde{P}_e := \sum_{i<j} \sum_{d \in \mathbb{Z}_+^k} \min\{\mathbb{P}(D(u) = d \mid \sigma(u) = i, \sigma_{-u})p_i, \mathbb{P}(D(u) = d \mid \sigma(u) = j, \sigma_{-u})p_j\}$$

Error probability for the genie-based estimator

Fact: $\frac{1}{k-1}\tilde{P}_e \leqslant \mathbb{P}(\hat{\sigma}_{\text{genie}}(u) \neq \sigma(u) \mid \sigma_{-u}) \leqslant \tilde{P}_e$, where

$$\tilde{P}_e := \sum_{i<j} \sum_{d \in \mathbb{Z}_+^k} \min\{\mathbb{P}(D(u) = d \mid \sigma(u) = i, \sigma_{-u})p_i, \mathbb{P}(D(u) = d \mid \sigma(u) = j, \sigma_{-u})p_j\}$$

Lemma 1 [Abbe, Sandon '15]

If $(G, \sigma) \sim \mathrm{SBM}(n, k, p, \rho_n Q)$ and $\rho_n = \frac{\log n}{n}$, then w.p. $\geqslant 1 - e^{-\Omega(n^c)}$

$$\tilde{P}_e = n^{-I(p,Q)+O\left(\frac{\log\log n}{\log n}\right)}$$

where

$$I(p, Q) = \min_{i<j} \mathrm{CH}((p_l Q_{il})_{l\in[k]} \parallel (p_l Q_{jl})_{l\in[k]})$$

$$\mathrm{CH}(\mu \parallel \nu) = \max_{t\in[0,1]} \sum_x \nu(x) f_t\left(\frac{\mu(x)}{\nu(x)}\right), f_t(y) = 1 - t + ty - y^t$$

Impossibility

Theorem 1

Suppose that $(G, \sigma) \sim \mathrm{SBM}(n, k, p, \rho_n Q)$ and $\rho_n = \frac{\log n}{n}$. If $I(p, Q) < 1$, then for any estimator $\hat{\sigma}$,

$$\lim_{n\to\infty} \mathbb{P}(\hat{\sigma} \neq \sigma) = 1 \qquad \text{(Exact recovery impossible)}$$

Impossibility

Theorem 1

Suppose that $(G, \sigma) \sim \mathrm{SBM}(n, k, p, \rho_n Q)$ and $\rho_n = \frac{\log n}{n}$. If $I(p, Q) < 1$, then for any estimator $\hat{\sigma}$,

$$\lim_{n \to \infty} \mathbb{P}(\hat{\sigma} \neq \sigma) = 1 \qquad \text{(Exact recovery impossible)}$$

Example: Consider the symmetric SBM, i.e., $Q_{ij} = a$ if $i = j$ and $Q_{ij} = b$ for $i \neq j$, and $p_i = \frac{1}{k}$ and $\rho_n = \frac{\log n}{n}$. Then

$$\frac{(\sqrt{a} - \sqrt{b})^2}{k} \qquad \Longrightarrow \qquad \text{Exact recovery impossible}$$

Finding good estimators

Can we find an estimator that is efficiently computable and achieves exact recovery whenever $I(p, Q) > 1$*?*

Finding good estimators

Can we find an estimator that is efficiently computable and achieves exact recovery whenever $I(p, Q) > 1$*?*

Idea (Two-step estimator):

- Step 1: Good Guess. Take a good enough initial estimator, i.e., take $\hat{\sigma}_1$ that achieves almost exact recovery
- Step 2: Clean up. Compute $\hat{\sigma}_2(u) := \hat{\sigma}_{\text{genie}}(u, G, \hat{\sigma}_{1,-u})$ for all u

Finding good estimators

Can we find an estimator that is efficiently computable and achieves exact recovery whenever $I(p, Q) > 1$*?*

Idea (Two-step estimator):

- ➡ Step 1: Good Guess. Take a good enough initial estimator, i.e., take $\hat{\sigma}_1$ that achieves almost exact recovery
- ➡ Step 2: Clean up. Compute $\hat{\sigma}_2(u) := \hat{\sigma}_{\text{genie}}(u, G, \hat{\sigma}_{1,-u})$ for all u

Advantage: $\hat{\sigma}_2$ is efficiently computable if $\hat{\sigma}_1$ is so

Challenges:

1. How to find a good $\hat{\sigma}_1$?
2. $\hat{\sigma}_1$ depends on G, which makes $\hat{\sigma}_{\text{genie}}(u, G, \hat{\sigma}_{1,-u})$ difficult to analyze.

Graph-splitting

- (G_1, G_2) is constructed from G as follows:
  1. Include each edge of G in G_1 w.p. γ_n (independently)
  2. $G_2 = G \setminus G_1$ contains rest of the edges

Graph-splitting

- (G_1, G_2) is constructed from G as follows:
  1. Include each edge of G in G_1 w.p. γ_n (independently)
  2. $G_2 = G \setminus G_1$ contains rest of the edges

Modified Two-step estimator: Compute $\hat{\sigma}_{\text{genie}}(u, G_2, \hat{\sigma}_{1,-u}(G_1))$ for all u

Graph-splitting

- (G_1, G_2) is constructed from G as follows:
  1. Include each edge of G in G_1 w.p. γ_n (independently)
  2. $G_2 = G \setminus G_1$ contains rest of the edges

Modified Two-step estimator: Compute $\hat{\sigma}_{\text{genie}}(u, G_2, \hat{\sigma}_{1,-u}(G_1))$ for all u

Lemma 2

Suppose that $(G, \sigma) \sim \mathrm{SBM}(n, k, p, \rho_n Q)$ and $\rho_n = \frac{\log n}{n}$, and take graph-splitting (G_1, G_2) as above with $\frac{1}{\log n} \ll \gamma_n \ll \frac{\log \log n}{\log n}$.

- Suppose $\hat{\sigma}_1 = \hat{\sigma}_1(G_1)$ achieve almost exact recovery
- Take $\tilde{G}_2 \sim \mathrm{SBM}(n, k, p, \frac{(1-\gamma_n)\log n}{n} Q)$

Then for any u and $d \in \mathbb{Z}_+^k$, with high probability,

$$\begin{aligned} &\mathbb{P}(D(u; \hat{\sigma}_1, G_2) = d \mid G_1, \hat{\sigma}_{1,-u}, \sigma(u) = i) \\ &\leqslant (1 + o(1))\mathbb{P}(D(u; \hat{\sigma}_1, \tilde{G}_2) = d \mid \hat{\sigma}_{1,-u}, \sigma(u) = i) + n^{-\omega(1)} \end{aligned}$$

Achievability

Theorem 2

Suppose that $(G, \sigma) \sim \mathrm{SBM}(n, k, p, \rho_n Q)$ and $\rho_n = \frac{\log n}{n}$, and take graph-splitting (G_1, G_2) as above with $\frac{1}{\log n} \ll \gamma_n \ll \frac{\log \log n}{\log n}$.

- Suppose $\hat{\sigma}_1 = \hat{\sigma}_1(G_1)$ achieve almost exact recovery

Then,

$$I(p, Q) > 1 \implies \hat{\sigma}_{\text{genie}}(G_2, \hat{\sigma}_{1,-u}(G_1)) \text{ achieves exact recovery}$$

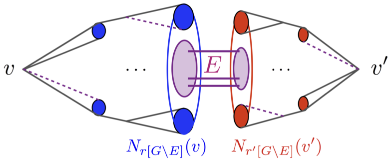
- *Exact recovery is possible up to the information theoretic threshold if almost exact recovery is possible for* $n\rho_n \to \infty$

How to produce a good initial estimator?

Approach 1: Sphere comparison

➢ Choose E by sampling from all edges with probability c (fixed)

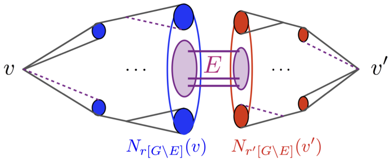
➢ Let $N_{r,r'}(v, v', E)$ be the number of pairs of vertices (v_1, v_2) such that $v_1 \in N_r(v, G \setminus E)$, $v_2 \in N_r(v', G \setminus E)$, and $\{v_1, v_2\} \in E$



Approach 1: Sphere comparison

➢ Choose E by sampling from all edges with probability c (fixed)

➢ Let $N_{r,r'}(v, v', E)$ be the number of pairs of vertices (v_1, v_2) such that $v_1 \in N_r(v, G \setminus E)$, $v_2 \in N_r(v', G \setminus E)$, and $\{v_1, v_2\} \in E$



➢ Idea of [Abbe, Sandon '15] is to use $N_{r,r'}(v, v', E)$ to come up with tests for deciding whether v, v' are in same community or not. For example, for $k = 2$, compute

$$I_{r,r'}(v, v', E) = N_{r+2,r'}(v, v', E) \times N_{r,r'}(v, v', E) - N_{r+1,r'}(v, v', E)^2$$

Result [Abbe, Sandon '15]. Sphere comparison achieves almost exact recovery for a suitable choice of r, c whenever $n\rho_n \to \infty$, Q is irreducible, and no two rows of Q are identical

Day 2

Recap

- Interested in **exact recovery** of σ when $(G, \sigma) \sim \mathrm{SBM}(n, k, p, \rho_n Q)$

Recap

➢ Interested in **exact recovery** of σ when $(G, \sigma) \sim \mathrm{SBM}(n, k, p, \rho_n Q)$

➢ Used Genie-based (hypothetical) estimator

$$\begin{aligned} \hat{\sigma}_{\text{genie}}(u) &:= \operatorname{argmax}_{i \in [k]} \mathbb{P}(\sigma(u) = i \mid G = g, \sigma_{-u}) p_i \\ &= \operatorname{argmax}_{i \in [k]} \mathbb{P}(D(u) = d \mid \sigma(u) = i, \sigma_{-u}) p_i \end{aligned}$$

Recap

- Interested in **exact recovery** of σ when $(G, \sigma) \sim \mathrm{SBM}(n, k, p, \rho_n Q)$

- Used Genie-based (hypothetical) estimator

$$\begin{aligned} \hat{\sigma}_{\text{genie}}(u) &:= \operatorname{argmax}_{i \in [k]} \mathbb{P}(\sigma(u) = i \mid G = g, \sigma_{-u}) p_i \\ &= \operatorname{argmax}_{i \in [k]} \mathbb{P}(D(u) = d \mid \sigma(u) = i, \sigma_{-u}) p_i \end{aligned}$$

- $\mathbb{P}(\hat{\sigma}_{\text{genie}}(u) \neq \sigma(u)) = n^{-I(p,Q)+o(1)}$

$I(p, Q) < 1 \implies$ *Impossibility*, $\qquad I(p, Q) > 1 \implies$ *Possible?*

Recap

➢ Interested in **exact recovery** of σ when $(G, \sigma) \sim \mathrm{SBM}(n, k, p, \rho_n Q)$

➢ Used Genie-based (hypothetical) estimator

$$\begin{aligned}\hat{\sigma}_{\text{genie}}(u) &:= \mathrm{argmax}_{i \in [k]} \, \mathbb{P}(\sigma(u) = i \mid G = g, \sigma_{-u}) p_i \\ &= \mathrm{argmax}_{i \in [k]} \, \mathbb{P}(D(u) = d \mid \sigma(u) = i, \sigma_{-u}) p_i\end{aligned}$$

➢ $\mathbb{P}(\hat{\sigma}_{\text{genie}}(u) \neq \sigma(u)) = n^{-I(p,Q)+o(1)}$

$$I(p, Q) < 1 \implies \textit{Impossibility}, \qquad I(p, Q) > 1 \implies \textit{Possible?}$$

➢ **Graph-splitting:** Compute $\hat{\sigma}_2(u) = \hat{\sigma}_{\text{genie}}(u; G_2, \hat{\sigma}_{-u}(G_1))$ for all u

$$\begin{aligned}&I(p, Q) > 1, \text{ and } \hat{\sigma}_1 \text{ achieves almost exact recovery for } n\rho_n \to \infty \\ &\implies \hat{\sigma}_2 \text{ achieves exact recovery}\end{aligned}$$

How to produce a good initial estimator?

Approach 1: Sphere comparison. [Abbe, Sandon '15] use $N_{r,r'}(v, v', E)$ to come up with tests for deciding whether v, v' are in same community or not. For example, for $k = 2$, compute

$$I_{r,r'}(v, v', E) = N_{r+2,r'}(v, v', E) \times N_{r,r'}(v, v', E) - N_{r+1,r'}(v, v', E)^2$$
$$\approx \frac{c^2(1-c)^{2r+2r'+2}}{n^2}\left(d - \frac{a-b}{2}\right)^2 d^{r+r'+1}\left(\frac{a-b}{2}\right)^{r+r'+1}(2\mathbb{1}\{\sigma(u) = \sigma(v)\} - 1)$$

[Abbe, Sandon '15] proved that such sphere comparison achieves almost exact recovery for a suitable choice of r, c whenever $n\rho_n \to \infty$, Q is irreducible, and no two rows of Q are identical

How to produce a good initial estimator?

Approach 1: Sphere comparison. [Abbe, Sandon '15] use $N_{r,r'}(v,v',E)$ to come up with tests for deciding whether v, v' are in same community or not. For example, for $k = 2$, compute

$$I_{r,r'}(v,v',E) = N_{r+2,r'}(v,v',E) \times N_{r,r'}(v,v',E) - N_{r+1,r'}(v,v',E)^2$$
$$\approx \frac{c^2(1-c)^{2r+2r'+2}}{n^2}\left(d - \frac{a-b}{2}\right)^2 d^{r+r'+1}\left(\frac{a-b}{2}\right)^{r+r'+1}(2\mathbb{1}\{\sigma(u)=\sigma(v)\}-1)$$

[Abbe, Sandon '15] proved that such sphere comparison achieves almost exact recovery for a suitable choice of r, c whenever $n\rho_n \to \infty$, Q is irreducible, and no two rows of Q are identical

Approach 2: Spectral Algorithms. Today's focus

- ➡ Almost recovery when $n\rho_n \to \infty$
- ➡ Exact recovery up to information theoretic threshold

Intuition: Spectral algorithm

Let A be the the adjacency matrix of G and $A^\star = \mathbb{E}[A \mid \sigma]$

$$A = \underbrace{A^\star}_{\text{signal}} + \underbrace{(A - A^\star)}_{\text{noise}}$$

➢ $(\lambda_i, \phi_i)_{i=1}^k$ and $(\lambda_i^\star, \phi_i^\star)_{i=1}^k$ be the top k eigenpairs of $A, A^\star$ resp.

Intuition: Spectral algorithm

Let A be the the adjacency matrix of G and $A^\star = \mathbb{E}[A \mid \sigma]$

$$A = \underbrace{A^\star}_{\text{signal}} + \underbrace{(A - A^\star)}_{\text{noise}}$$

➢ $(\lambda_i, \phi_i)_{i=1}^k$ and $(\lambda_i^\star, \phi_i^\star)_{i=1}^k$ be the top k eigenpairs of $A, A^\star$ resp.

Fact: If $\Phi^\star = (\phi_1^\star, \dots, \phi_k^\star)$ be $n \times k$ matrix, then

$$(\Phi^\star)_{u,\cdot} \begin{cases} = (\Phi^\star)_{v,\cdot} & \text{if } \sigma(u) = \sigma(v) \\ \neq (\Phi^\star)_{v,\cdot} & \text{if } \sigma(u) \neq \sigma(v) \end{cases}$$

Intuition: Spectral algorithm

Let A be the the adjacency matrix of G and $A^\star = \mathbb{E}[A \mid \sigma]$

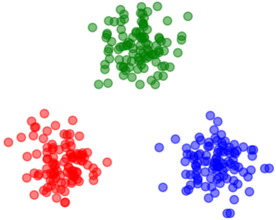
$$A = \underbrace{A^\star}_{\text{signal}} + \underbrace{(A - A^\star)}_{\text{noise}}$$

- $(\lambda_i, \phi_i)_{i=1}^k$ and $(\lambda_i^\star, \phi_i^\star)_{i=1}^k$ be the top k eigenpairs of $A, A^\star$ resp.

Fact: If $\Phi^\star = (\phi_1^\star, \dots, \phi_k^\star)$ be $n \times k$ matrix, then

$$(\Phi^\star)_{u,\cdot} \begin{cases} = (\Phi^\star)_{v,\cdot} & \text{if } \sigma(u) = \sigma(v) \\ \neq (\Phi^\star)_{v,\cdot} & \text{if } \sigma(u) \neq \sigma(v) \end{cases}$$

If $\|A - A^\star\|$ small $\implies \Phi \approx \Phi^\star$



Matrix perturbation theory

Let $(\lambda_i(X), \phi_i(X))$ be the i-th top eigenpair of X

Theorem [Davis, Kahan '70]

Suppose X, X_0 are symmetric matrices with X_0 has k distinct non-zero eigenvalues. Then

$$\min_{s\in\{\pm 1\}} \|\phi_i(X) - s\phi_i(X_0)\|_2 \leqslant \frac{c\|X - X_0\|_{2\to 2}}{\min\{\{\lambda_{i-1}(X_0) - \lambda_i(X_0), \lambda_i(X_0) - \lambda_{i+1}(X_0)\}}$$

with $\lambda_0(X_0) = +\infty$, $\lambda_k(X_0) = -\infty$, for some absolute constant $c > 0$

Spectral algorithm for almost exact recovery

Is $\|A - A^\star\|_{2\to 2} \ll n\rho_n$ *whp whenever* $n\rho_n \to \infty$*?*

Spectral algorithm for almost exact recovery

Is $\|A - A^\star\|_{2\to2} \ll n\rho_n$ *whp whenever* $n\rho_n \to \infty$*?*

— NO, instead we can look at the trimmed matrix

Spectral algorithm for almost exact recovery

Is $\|A - A^\star\|_{2\to 2} \ll n\rho_n$ *whp whenever* $n\rho_n \to \infty$*?*

— NO, instead we can look at the trimmed matrix

Spectral clustering: Compute $\hat{\sigma}_1$

(S1) Construct $\tilde{A}$ as

$$\tilde{A}_{ij} = \begin{cases} A_{ij} & \text{if } d_i \text{ or } d_j < 2\|Q\|_\infty n\rho_n \\ 0 & \text{otherwise} \end{cases}$$

(S2) Construct $\tilde{\Phi}$ with top k eigenvectors of $\tilde{A}$

(S3) Apply k-means clustering on the rows of $\tilde{\Phi}$

Theorem

$\hat{\sigma}_1$ achieves almost exact recovery whenever $n\rho_n \to \infty$

$\implies$ $\hat{\sigma}_2 = \hat{\sigma}_{\text{genie}}(G_1, \hat{\sigma}_1)$ achieves exact recovery whenever $I(p, Q) > 1$

We have shown

Spectral clustering + clean-up achieves exact recovery for $I(p, Q) > 1$

Do direct spectral algorithms achieve this optimal recovery?

Need an entrywise perturbation bound...

Entrywise perturbation bound

Assumptions.

(A1) *Well-behaved mean matrix.* $A^\star$ has k distinct, non-zero eigenvalues $\lambda_1^\star > \cdots > \lambda_k^\star$ with $\lambda_k^\star = \Theta(\lambda_1^\star)$. Moreover, for some $\gamma \to 0$ and $\Delta = \min_i\{\lambda_{i-1}^\star - \lambda_i^\star\}$ with $\lambda_0^\star = \infty$

$$\|A^\star\|_{2\to\infty} \leqslant \gamma\Delta$$

(A2) *Row-wise and column-wise independence.* For any $m \in [n]$, $(A_{ij} : i = m \text{ or } j = m)$ is independent of $(A_{ij} : i \neq m \text{ or } j \neq m)$

(A3) *Spectral norm concentration.* $\|A - A^\star\|_{2\to 2} \leqslant \gamma\Delta$ whp

(A4) *Row concentration.* $\exists\psi : \mathbb{R}_+ \mapsto \mathbb{R}_+$ (continuous, non-decreasing, possibly depending on n) such that $\psi(0) = 0, \psi(1) = O(1)$, $\frac{\psi(x)}{x}$ is non-increasing and for any $m \in [n], w \in \mathbb{R}^n$

$$|\langle (A - A)_m, w\rangle| \leqslant \Delta\|w\|_\infty \psi\left(\frac{\|w\|_2}{\sqrt{n}\|w\|_\infty}\right) \quad \text{w.p. } \geqslant 1 - o(n^{-1})$$

Entrywise perturbation bound

Assumptions.

(A1) *Well-behaved mean matrix.* $A^\star$ has k distinct, non-zero eigenvalues $\lambda_1^\star > \cdots > \lambda_k^\star$ with $\lambda_k^\star = \Theta(\lambda_1^\star)$. Moreover, for some $\gamma \to 0$ and $\Delta = \min_i\{\lambda_{i-1}^\star - \lambda_i^\star\}$ with $\lambda_0^\star = \infty$

$$\|A^\star\|_{2\to\infty} \leqslant \gamma\Delta$$

(A2) *Row-wise and column-wise independence.* For any $m \in [n]$, $(A_{ij} : i = m \text{ or } j = m)$ is independent of $(A_{ij} : i \neq m \text{ or } j \neq m)$

(A3) *Spectral norm concentration.* $\|A - A^\star\|_{2\to 2} \leqslant \gamma\Delta$ whp

(A4) *Row concentration.* $\exists\psi : \mathbb{R}_+ \mapsto \mathbb{R}_+$ (continuous, non-decreasing, possibly depending on n) such that $\psi(0) = 0, \psi(1) = O(1)$, $\frac{\psi(x)}{x}$ is non-increasing and for any $m \in [n], w \in \mathbb{R}^n$

$$|\langle (A - A)_m, w\rangle| \leqslant \Delta\|w\|_\infty \psi\left(\frac{\|w\|_2}{\sqrt{n}\|w\|_\infty}\right) \quad \text{w.p. } \geqslant 1 - o(n^{-1})$$

Theorem [Abbe, Fan, Wang, Zhong '20]

$\min_{s\in\{\pm 1\}} \left\|\phi_i - s\frac{A\phi_i^\star}{\lambda_i^\star}\right\|_\infty = o(\|\phi_i^\star\|_\infty)$ whp for all $i \in [k]$