Spectral Clustering and Community Detection

Arnab Sen

June, 2023

Part II

Erdös-Rényi random graph

$G(n,p)$: A graph on n vertices, where each pair of vertices are connected by an edge independently with probability p.

Recall that adjacency matrix of a graph G on n vertices is an $n \times n$ symmetric matrix A such that

$$A(i,j) = \begin{cases} 1 & i \sim j \\ 0 & i \nsim j. \end{cases}$$

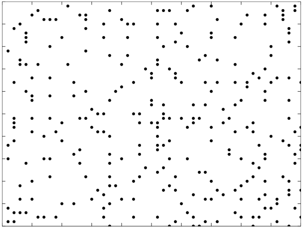


Figure: Adjacency matrix of $G(50, .125)$

Stochastic Block Model (balanced, with two communities)

Divide n vertices into two groups S_1 and S_2 such that $|S_1| = |S_2| = n/2$. Each vertex i has a label σ_i

$$\sigma_i = \begin{cases} +1 & i \in S_1 \\ -1 & i \in S_2. \end{cases}$$

$i \sim j$ with probability

$$= \begin{cases} p & \text{if } i \text{ \& } j \text{ are in same group, i.e, } \sigma_i = \sigma_j, \\ q & \text{if } i \text{ \& } j \text{ are in different groups, i.e, } \sigma_i \neq \sigma_j. \end{cases}$$

The above random graph is called the stochastic block model (SBM) and is denoted by $G(n, p, q)$. We assume that $p > q$.

A variant of SBM: choose the labels $\sigma_i \overset{\text{i.i.d.}}{\sim} \pm 1$ with probability $1/2$.

Community detection problem. Identify the (hidden) labels (possibly approximately) from $(\sigma_i)_{i \in [n]}$ from the adjacency matrix of $G(n, p, q)$.

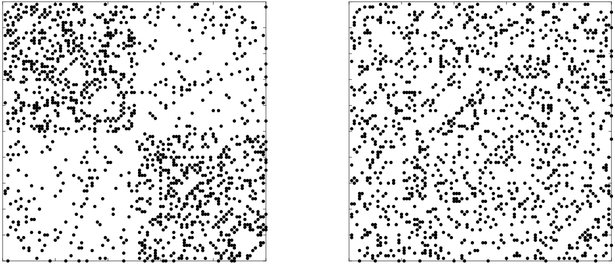


Figure: Adjacency matrix of $G(100, .2, .05)$ (Left) vertices are ordered into groups 1 and 2. (Right) vertices are unordered.

We have already seen that the second eigenvector of laplacian/adjacency matrix is useful in detecting the community. We will see its performance in this random graph model.

We write $A = \mathbb{E}[A] + E$ where

$$E = A - \mathbb{E}[A].$$

$$\mathbb{E}[A] = \left[\begin{array}{c|c} pJ_{n/2} & qJ_{n/2} \\ \hline qJ_{n/2} & pJ_{n/2} \end{array}\right] - pI_n,$$

where $J_{n/2}$ is $n/2 \times n/2$ matrix of all ones.

The eigenvalues of $\mathbb{E}[A]$ are

$$\frac{p+q}{2}n - p, \quad \frac{p-q}{2}n - p, \quad -p \quad \text{(multiplicity } n-2\text{)}.$$

The eigenvectors corresponding to top two eigenvalues

$$\begin{pmatrix} \mathbf{1}_{n/2} \\ \mathbf{1}_{n/2} \end{pmatrix}, \begin{pmatrix} \mathbf{1}_{n/2} \\ -\mathbf{1}_{n/2} \end{pmatrix}$$

The second eigenvector of $\mathbb{E}[A]$ perfectly recovers the labels $(\sigma_i)_{i\in[n]}$!

However, we only get to observe A not $\mathbb{E}[A]$.

View $A = \mathbb{E}[A] + E$ as a perturbation of $\mathbb{E}[A]$. Is

$$\text{2nd eigenvector of } A \approx \text{the 2nd eigenvector of } \mathbb{E}[A]$$

under this perturbation?

$\|E\| = \sup_{x:\|x\|_2=1} \|Ex\|_2$.

Theorem

Let $E = A - \mathbb{E}[A]$. *Then*

$$\|E\| \le C\sqrt{n} \quad \textit{with high probability.}$$

The above theorem is a corollary of the following result.

Theorem (Spectral norm bound of a non-symmetric matrix)

Let B *be an* $n \times n$ *(non-symmetric) matrix such that the entries are independent, mean zero, and* $|B_{ij}| \le 1$ *for all* i, j. *Then*

$$\|B\| \le C'\sqrt{n} \text{ with high probability.}$$

Decompose E into the upper-triangular part E^+ and lower-triangular part E^- such that

$$E = E^+ + E^-.$$

Apply the second theorem separately for E^+ and E^-. Then with high probability

$$\|E\| \le \|E^+\| + \|E^-\| \le 2C'\sqrt{n}.$$

Proof of spectral norm bound

$$\|B\| = \sup_{x,y\in\mathbb{R}^n : \|x\|_2 = \|y\|_2 = 1} \langle x, By\rangle.$$

Concentration bound: For *fixed* $x, y \in \mathcal{N}$ and for any $u > 0$,

$$\mathbb{P}\big(\langle x, By\rangle \geq u\big) \leq e^{-u^2/8}.$$

Problem: The above supremum is over an infinite set $S^{n-1} \times S^{n-1}$.

Solution: We can take supremum over a suitable finite set (called ϵ-net) of $S^{n-1} \times S^{n-1}$ by only paying a multiplicative constant factor.

ϵ-net

Definition

A subset $\mathcal{N} \subset (\mathbb{X}, d)$ is called an ϵ-net if for any $u \in \mathbb{X}$, there exists $v \in \mathcal{N}$ such that $d(u, v) \leq \epsilon$.

Lemma (size of ϵ-net)

There exists an ϵ-net of S^{n-1} of size at most $(1 + 2/\epsilon)^s$.

We build an ϵ-net as follows.

Start by adding points one by one (arbitrarily) in S^{n-1} such that any two pair of points are at least ϵ distance apart. Stop when no more points can be added. The resulting set $\mathcal{N}$ is an ϵ-net (why?).

To bound $|\mathcal{N}|$, we bound the n-dimensional volume of the set

$$\mathcal{N}^{\epsilon} := \bigcup_{u \in \mathcal{N}} \mathbb{B}(u, \epsilon/2)$$

from below and above.

Since the pairwise distance among the points in $\mathcal{N}$ is at least ϵ, the balls of radius $\epsilon/2$ around the points in $\mathcal{N}$ are disjoint. So,

$$\mathrm{Vol}(\mathcal{N}^{\epsilon}) \geq |\mathcal{N}| \mathrm{Vol}(\mathbb{B}(0, \epsilon/2)).$$

On the other hand, $\mathcal{N}^{\epsilon} \subset \mathbb{B}(0, 1+\epsilon/2)$ yielding that

$$\mathrm{Vol}(\mathcal{N}^{\epsilon}) \leq \mathrm{Vol}(\mathbb{B}(0, 1+\epsilon/2)).$$

Combining the two estimates

$$|\mathcal{N}| \leq \frac{\mathrm{Vol}(\mathbb{B}(0, 1+\epsilon/2))}{\mathrm{Vol}(\mathbb{B}(0, \epsilon/2))} = \left(\frac{1+\epsilon/2}{\epsilon/2}\right)^{n} = (1+2/\epsilon)^{n}.$$

Let $\mathcal{N}$ be a $1/4$-net of the sphere S^{n-1} of size 9^n. Then (exercise)

$$\|B\| \le 2 \sup_{x,y \in \mathcal{N}} \langle x, By \rangle.$$

Union bound over the net:

$$\begin{aligned}
\mathbb{P}(\|B\| > C'\sqrt{n}) &\le \mathbb{P}\Big(\sup_{x,y \in \mathcal{N}} \langle x, By \rangle > (C'/2)\sqrt{n} \Big) \\
&\le \sum_{x,y \in \mathcal{N}} \mathbb{P}\Big(\langle x, By \rangle > (C'/2)\sqrt{n} \Big) \\
&\le (9^n)^2 \cdot e^{-C'^2 n/32},
\end{aligned}$$

which can be made exponentially small in n by choosing sufficiently large constant $C' > 0$. $\square$

Perturbation of eigenvalues

Let M and E be symmetric matrices. Set

$$\widehat{M} = M + E.$$

Let $\lambda_i(M)$ be the i-th largest eigenvalue of M with unit eigenvector $v_i(M)$ (and similarly for $\widehat{M}$).

Theorem (Weyl's law)

$$|\, \lambda_i(\widehat{M}) - \lambda_i(M)\, | \le \|E\| \qquad \textit{for each } i.$$

Hence for $G(n, p, q)$, with high probability

$$\lambda_1(A) \approx \frac{p+q}{2}n, \qquad \lambda_2(A) \approx \frac{p-q}{2}n, \qquad \max_{i \ge 2} |\lambda_i(A)| \le C\sqrt{n}.$$

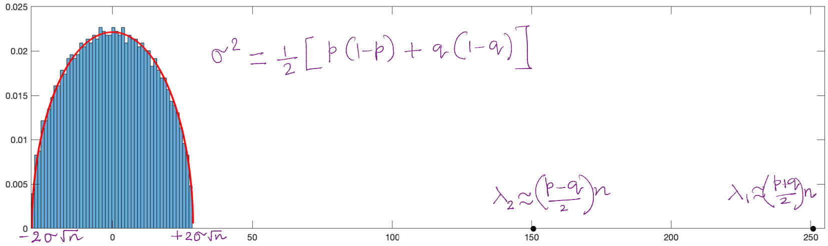


Figure: Eigenvalues of SBM $G(n = 2000, p = .2, q = .05)$.

Perturbation of eigenvectors

The eigenvectors of M and $\widehat{M}$ may not be close to each other even if $\|E\|$ is small.

Example: Let $\epsilon > 0$ be small.

$$M = \begin{bmatrix} 1+\epsilon & 0 \\ 0 & 1-\epsilon \end{bmatrix}, \quad \widehat{M} = \begin{bmatrix} 1 & \epsilon \\ \epsilon & 1 \end{bmatrix}.$$

Check that $\|\widehat{M} - M\| = \sqrt{2}\epsilon$. Also,

$$\lambda_1(M) = \lambda_1(\widehat{M}) = 1+\epsilon, \qquad \lambda_2(M) = \lambda_2(\widehat{M}) = 1-\epsilon.$$

However, the eigenvectors are totally different:

$$v_1(M) = \begin{bmatrix} 1 \\ 0 \end{bmatrix}, v_2(M) = \begin{bmatrix} 0 \\ 1 \end{bmatrix}, \quad v_1(\widehat{M}) = \frac{1}{\sqrt{2}}\begin{bmatrix} 1 \\ 1 \end{bmatrix}, v_2(\widehat{M}) = \frac{1}{\sqrt{2}}\begin{bmatrix} 1 \\ -1 \end{bmatrix}.$$

The instability of eigenvectors of M is caused by the lack of separation between $\lambda_1(M)$ and $\lambda_2(M)$.

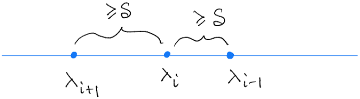
Theorem (Davis-Kahan)

Fix i. Let

$$\delta = \min_{j \neq i} |\lambda_j(M) - \lambda_i(M)| > 0.$$

Then there exists $\theta \in \{-1, +1\}$ such that

$$\|v_i(M) - \theta v_i(\widehat{M})\|_2 \leq \frac{4\|E\|}{\delta}.$$



Theorem

Let A be the adjacency matrix of SBM $G(n, p, q)$. Let $\mu = \min(q, \frac{p-q}{2}) > 0$. Then with high probability $\mathrm{sgn}(v_2(A))$ identifies the two communities of G, except for C/μ^2 misclassified vertices for some constant $C > 0$.

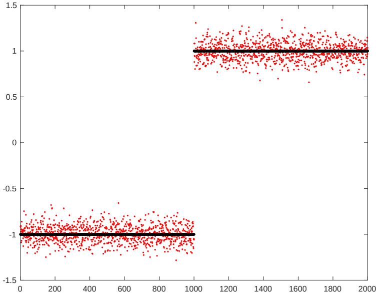


Figure: 2nd eigenvector of SBM $G(n = 2000, p = .2, q = .05)$.

Proof

We apply Davis-Kahan theorem to compare $v_2(\mathbb{E}[A])$ with $v_2(A)$. Here $M = \mathbb{E}[A]$ and $\widehat{M} = A = \mathbb{E}[A] + E$.

For $\mathbb{E}[A]$, the eigenvalue gap around λ_2 is

$$\delta = \min\left(n\frac{p-q}{2}, nq\right) = n\mu > 0.$$

By Davis Kahan, there exists $\theta \in \{-1, 1\}$ such that

$$\|v_2(\mathbb{E}[A]) - \theta v_2(A)\|_2 \leq \frac{4\|E\|}{n\mu} \leq \frac{C''}{\sqrt{n\mu}},$$

with high probability.

$$\sum_i \left| \sqrt{n} v_2(\mathbb{E}[A])_i - \sqrt{n}\theta v_2(A)_i \right|^2 \leq \frac{(C'')^2}{\mu^2}.$$

This implies that

$$\sum_i \left| \sigma_i - \sqrt{n}\theta v_2(A)_i \right|^2 \leq \frac{(C'')^2}{\mu^2}.$$

If $\sigma_i \neq \operatorname{sgn}(\theta v_2(A)_i)$, then the i-th term in the sum is bigger than 1.

$$\sum_i \mathbf{1}(\sigma_i \neq \operatorname{sgn}(\theta v_2(A)_i)) \leq \frac{(C'')^2}{\mu^2},$$

with high probability. $\square$

Question. How can we estimate p and q from the adjacency matrix?

SBM in sparse case

We will consider SBM $G(n, p = \frac{a}{n}, q = \frac{b}{n})$ where $a > b > 0$ are constants.

The mean degree of a vertex is $\approx d := \frac{a+b}{2}$.

The eigenvalues of $\mathbb{E}[A]$ are

$$\frac{a+b}{2} - \frac{a}{n} \approx d, \quad \frac{a-b}{2} - \frac{a}{n} \approx \frac{a-b}{2}, \quad -\frac{a}{n} \approx 0 \quad \text{(multiplicity } n-2\text{)}.$$

The eigenvectors corresponding to top two eigenvalues

$$\begin{pmatrix} \mathbf{1}_{n/2} \\ \mathbf{1}_{n/2} \end{pmatrix}, \begin{pmatrix} \mathbf{1}_{n/2} \\ -\mathbf{1}_{n/2} \end{pmatrix}$$

Even in the sparse case, the second eigenvector of $\mathbb{E}[A]$ exactly recovers the community labels.

However ...

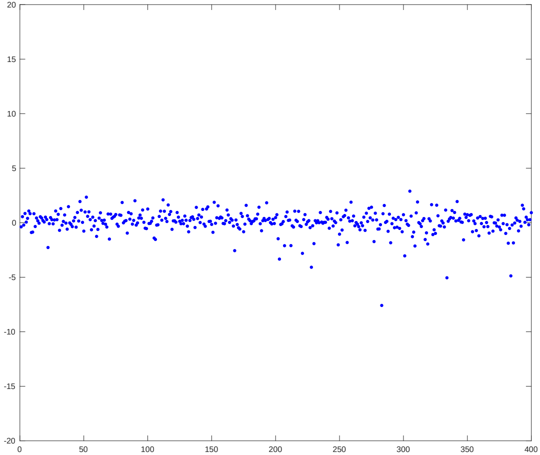


Figure: 2nd eigenvector of SBM $G(n = 400, p = 4/n, q = 2/n)$

Failure of spectral method

In the sparse case, the noise matrix $E := A - \mathbb{E}[A]$ is too big.

In fact,

$$\|A\|, \|E\| \sim \sqrt{\frac{\log n}{\log\log n}} \quad \text{vs} \quad \|\mathbb{E}[A]\| \sim d.$$

Hence, we do not expect $v_2(A)$ and $v_2(\mathbb{E}(A))$ are close to each other.

effect of high degree vertices

For simplicity, let us consider Erdos-Renyi graph $G(n, d/n)$.

If we pretend the degree of vertices are i.i.d. $\mathrm{Bin}(n-1, d/n) \approx \mathrm{Poi}(d)$ random variables, then

$$\text{max degree } = d_{\max} \sim c_n$$

where $\mathbb{P}(\mathrm{Poi}(d) \geq c_n) = 1/n$. A calculation yields $c_n \sim \frac{\log n}{\log \log n}$.

We would like to argue that with high probability

$$\|A\| = \lambda_1(A) \sim \sqrt{d_{\max}} \sim \sqrt{\frac{\log n}{\log \log n}},$$

Heuristics for $\lambda_1(A) \sim \sqrt{d_{\max}}$

Lower bound. Let i be a vertex with degree $d_{\max}$.

$$\lambda_1(A)^2 \geq \langle e_i, A^2 e_i \rangle = (A^2)_{ii} = d_{\max}.$$

Upper bound. For any $k \geq 1$

$$\begin{aligned} \lambda_1(A)^{2k} &\leq \sum_j \lambda_j(A)^{2k} = \operatorname{tr}(A^{2k}) \\ &= \sum_j (A^{2k})_{jj} \leq n \max_j (A^{2k})_{jj}. \end{aligned}$$

$$\begin{aligned} (A^{2k})_{jj} &= \sum_{j_1, j_2, \ldots, j_{2k-1}} A_{jj_1} A_{j_1 j_2} \cdots A_{j_{2k-1} j} \\ &= \text{ number of closed walks of length } 2k \text{ from } j \text{ to } j \end{aligned}$$

If j is a high degree vertices, the number of closed walks of length $2k$ from j is dominated by the closed walks of the form

$$j \to i_1 \to j \to i_2 \to \cdots \to j \to i_k \to j, \quad i_1, i_2, \ldots, i_k \text{ are neighbors of } j,$$

where we allow repetition. There are exactly $\deg(j)^k$ of them.

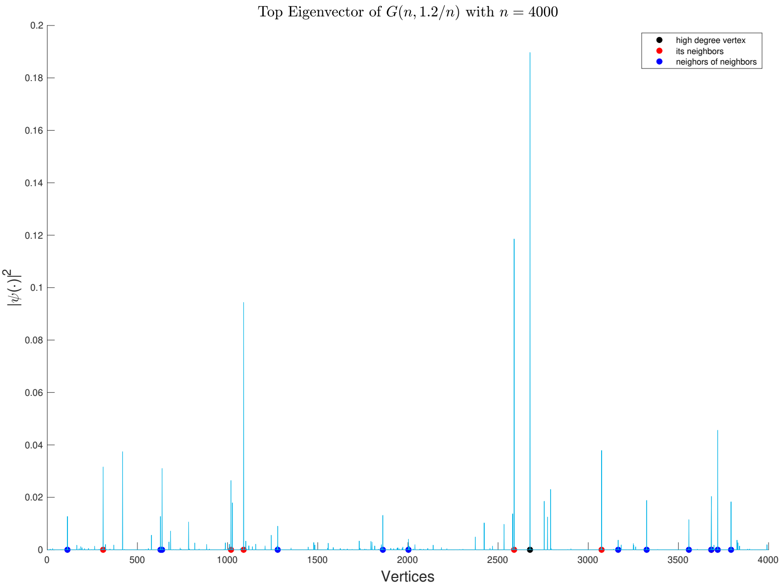
$$\max_j A_{jj}^{2k} \leq ((1+\epsilon)d_{\max})^k.$$

By choosing $k \gg \log n$ such that $n^{1/2k} \to 1$, we see that

$$\lambda_1(A) \leq (1+\epsilon)\sqrt{d_{\max}}.$$

The leading eigenvalues of A are all close to $\sqrt{\frac{\log n}{\log\log n}}$ and the corresponding leading eigenvectors tend to localize around high degree nodes.

For SBM, the leading eigenvectors are again created by the high degree nodes and do not contain information about the community labels.



Top Eigenvector of $G(n, 1.2/n)$ with $n = 4000$
high degree vertex
its neighbors
neighors of neighbors
$|\psi(\cdot)|^2$
Vertices
0
0.02
0.04
0.06
0.08
0.1
0.12
0.14
0.16
0.18
0.2
500
1000
1500
2000
2500
3000
3500
4000

Big theorem: phase transition in SBM

Recall that $d = \frac{a+b}{2}$.

Theorem

(a) *(no recovery) If* $(a-b)/2 < \sqrt{d}$ *or equivalently,* $(a-b)^2 < 2(a+b)$, *then any estimate* $\hat{\sigma} = \sigma(A)$ *will fail to perform better than random guess, i.e.,*

$$\frac{1}{n}\sum_{i=1}^{n} \mathbf{1}(\hat{\sigma}_i = \sigma_i) \to \frac{1}{2}.$$

(b) *(partial recovery) If* $(a-b)/2 > \sqrt{d} \Leftrightarrow (a-b)^2 < 2(a+b)$, *then there exists an estimate* $\hat{\sigma}$ *that performs better than random guess, i.e., there exists* $c > 0$ *such that*

$$\frac{1}{n}\sum_{i=1}^{n} \mathbf{1}(\hat{\sigma}_i = \sigma_i) \geq \frac{1}{2} + c \quad \textit{for large } n.$$

Remark. In the sparse regime, the graph has a linear number of isolated vertices. So, even if a is much larger than b, it is still not possible to come up with an estimate that gives (almost) exact recovery, i.e.,

$$\frac{1}{n}\sum_{i=1}^{n} \mathbf{1}(\hat{\sigma}_i = \sigma_i) \to 1.$$

Consistent estimates for a and b

Theorem

(a) *If $(a-b)/2 > \sqrt{d}$, then there exist consistent estimators*

$$\hat{a}_n \to a \quad \text{and} \quad \hat{b}_n \to b.$$

Moreover, these estimators can be computed in polynomial time.

(b) *If $(a-b)/2 < \sqrt{d}$, then there are no consistent estimators of a and b.*

No detection if $(a-b)/2 < \sqrt{d}$

We will give an argument that if $(a-b)/2 < \sqrt{d}$, then it is not possible to distinguish two hypotheses

$$H_0 : A_n \sim G(n, d/n) \quad \text{vs} \quad H_1 : A_n \sim G(n, a/n, b/n),$$

where $\boldsymbol{\sigma} = (\sigma_i)_{i\in[n]}$ be i.i.d. ± 1 symmetric labels in SBM.

This means there does not exist a test statistics $T_n = T_n(A_n)$ ($T_n = 0$ if we accept H_0 and $T_n = 1$ if we accept H_1) such that

$$\mathbb{P}_{H_0}(T_n = 1) + \mathbb{P}_{H_1}(T_n = 0) \to 0.$$

- Non-detection strongly indicates (but it does not prove) non-recovery.
- If $(a-b)/2 < \sqrt{d}$, then we can not distinguish between $G(n, a/n, b/n)$ and $G(n, \alpha/n, \beta/n)$ if $a + b = \alpha + \beta$ and $(\alpha - \beta)/2 < \sqrt{(\alpha + \beta)/2}$. So, we can not consistently estimate a and b.

We will show that $A_n \sim H_1$ is contiguous to $A_n \sim H_0$, i.e., for any sequence of events F_n

$$\mathbb{P}_{H_0}(A_n \in F_n) \to 0 \Rightarrow \mathbb{P}_{H_1}(A_n \in F_n) \to 0.$$

It is easy to see that contiguity implies non-detection: take $F_n = \{T_n = 1\}$.

Let $\mathbf{A}_n = (A_n(i,j))_{i<j}$ be the collection of the upper triangular entries of A_n.

Let f_n and g_n be the p.m.f. of $\mathbf{A}_n$ under H_1 and H_0 respectively, i.e.,

$$f_n(\mathbf{a}) = \mathbb{P}_{H_1}(\mathbf{A}_n = \mathbf{a}), \quad g_n(\mathbf{a}) = \mathbb{P}_{H_0}(\mathbf{A}_n = \mathbf{a}).$$

Also, $f_n(\mathbf{a}|\boldsymbol{\sigma}) = \mathbb{P}_{H_1}(\mathbf{A}_n = \mathbf{a}|\boldsymbol{\sigma})$ denotes the conditional p.m.f. given the labels. So, we have

$$f_n(\mathbf{a}) = \mathbb{E}_{\boldsymbol{\sigma}} f_n(\mathbf{a}|\boldsymbol{\sigma}).$$

Define

$$\underbrace{\chi^2(H_1 \| H_0)}_{\chi^2 \text{ divergence of } H_1 \text{ w.r.t. } H_0} := \mathbb{E}_{H_0}\left(\frac{f_n(\mathbf{A}_n)}{g_n(\mathbf{A}_n)} - 1\right)^2$$

$$= \sum_{\mathbf{a}}\left(\frac{f_n(\mathbf{a})}{g_n(\mathbf{a})} - 1\right)^2 g(\mathbf{a}) = \sum_{\mathbf{a}} \frac{f_n(\mathbf{a})^2}{g_n(\mathbf{a})} - 1$$

Observation. If $\chi^2(H_1 \| H_0) \leq C$, then

$$\mathbb{P}_{H_0}(A_n \in F_n) \to 0 \Rightarrow \mathbb{P}_{H_1}(A_n \in F_n) \to 0.$$

Proof.

$$\begin{aligned}
\mathbb{P}_{H_1}(A_n \in F_n) &= \sum_{\mathbf{a}} f_n(\mathbf{a}) \mathbf{1}_{(\mathbf{a} \in F_n)} \\
&= \sum_{\mathbf{a}} \frac{f_n(\mathbf{a})}{g_n(\mathbf{a})} g_n(\mathbf{a}) \mathbf{1}_{(\mathbf{a} \in F_n)} \\
&\leq \sqrt{\left(\sum_{\mathbf{a}} \left(\frac{f_n(\mathbf{a})}{g_n(\mathbf{a})} \right)^2 g_n(\mathbf{a}) \right) \left(\sum_{\mathbf{a}} \mathbf{1}_{(\mathbf{a} \in F_n)} g_n(\mathbf{a}) \right)} \quad \text{(Cauchy-Schwarz)} \\
&\leq (C+1)^{1/2} \sqrt{\mathbb{P}_{H_0}(A_n \in F_n)} \to 0.
\end{aligned}$$

Lemma

If $(a-b)/2 < \sqrt{d}$, *then* $\chi^2(H_1 \| H_0) \le C$.

Proof. Replica trick.

$$\chi^2(H_1\|H_0) + 1 = \sum_{\mathbf{a}} \frac{f_n(\mathbf{a})^2}{g_n(\mathbf{a})} = \sum_{\mathbf{a}} \frac{\left(\mathbb{E}_{\boldsymbol{\sigma}} f_n(\mathbf{a}|\boldsymbol{\sigma})\right)^2}{g_n(\mathbf{a})}$$

$$= \sum_{\mathbf{a}} \frac{\mathbb{E}_{\boldsymbol{\sigma},\tilde{\boldsymbol{\sigma}}}\left(f_n(\mathbf{a}|\boldsymbol{\sigma}) f_n(\mathbf{a}|\tilde{\boldsymbol{\sigma}})\right)}{g_n(\mathbf{a})} \quad (\tilde{\boldsymbol{\sigma}} \text{ is a i.i.d. copy of } \boldsymbol{\sigma}).$$

Let P, Q and $(P+Q)/2$ be the p.m.f.s of $\mathrm{Ber}(p = a/n), \mathrm{Ber}(q = b/n)$, and $\mathrm{Ber}((p+q)/2 = d/n)$.
For example, $P(a) = \mathbb{P}(\mathrm{Ber}(p) = a) = p^a(1-p)^{1-a},\ a \in \{0,1\}$.

$$f_n(\mathbf{a}|\boldsymbol{\sigma}) = \prod_{i<j}\left(P(a_{ij})\mathbf{1}_{\{\sigma_i\sigma_j=1\}} + Q(a_{ij})\mathbf{1}_{\{\sigma_i\sigma_j=-1\}}\right) = \prod_{i<j}\left(\frac{P+Q}{2} + \frac{P-Q}{2}\sigma_i\sigma_j\right),$$

$$g_n(\mathbf{a}) = \prod_{i<j}\left(\frac{P+Q}{2}\right).$$

$$\chi^2(H_1\|H_0)+1=\sum_{\mathbf{a}}\mathbb{E}_{\boldsymbol{\sigma},\tilde{\boldsymbol{\sigma}}}\prod_{i<j}\left(\frac{(\frac{P+Q}{2}+\frac{P-Q}{2}\sigma_i\sigma_j)(\frac{P+Q}{2}+\frac{P-Q}{2}\tilde{\sigma}_i\tilde{\sigma}_j)}{\frac{P+Q}{2}}\right)$$

$$=\mathbb{E}_{\boldsymbol{\sigma},\tilde{\boldsymbol{\sigma}}}\prod_{i<j}\sum_{a_{ij}}\left(\frac{P+Q}{2}+\frac{P-Q}{2}\sigma_i\sigma_j+\frac{P-Q}{2}\tilde{\sigma}_i\tilde{\sigma}_j+\frac{(P-Q)^2}{2(P+Q)}\sigma_i\sigma_j\tilde{\sigma}_i\tilde{\sigma}_j\right)$$

$$\sum_{a_{ij}}\frac{P+Q}{2}(a_{ij})=1,\quad \sum_{a_{ij}}\frac{P-Q}{2}(a_{ij})=0$$

$$\sum_{a_{ij}}\frac{(P-Q)^2}{2(P+Q)}(a_{ij})=\frac{(p-q)^2}{2(p+q)}+\frac{(p-q)^2}{2(2-p+q)}=\frac{\alpha+\epsilon_n}{n},$$

where

$$\alpha:=\frac{(a-b)^2}{2(a+b)}\quad \text{and } 0\le\epsilon_n\le\frac{C'}{n}.$$

$$
\begin{aligned}
\chi^2(H_1 \| H_0) + 1 &= \mathbb{E}_{\boldsymbol{\sigma}, \tilde{\boldsymbol{\sigma}}} \prod_{i<j} \left(1 + \frac{\alpha + \epsilon_n}{n} \sigma_i \sigma_j \tilde{\sigma}_i \tilde{\sigma}_j\right) \\
&\leq \mathbb{E}_{\boldsymbol{\sigma}, \tilde{\boldsymbol{\sigma}}} \exp\left(\frac{\alpha + \epsilon_n}{n} \sum_{i<j} \sigma_i \sigma_j \tilde{\sigma}_i \tilde{\sigma}_j\right) \\
&\leq \mathbb{E}_{\boldsymbol{\sigma}, \tilde{\boldsymbol{\sigma}}} \exp\left(\frac{\alpha + \epsilon_n}{2n} \sum_{i,j} \sigma_i \sigma_j \tilde{\sigma}_i \tilde{\sigma}_j\right) \\
&= \mathbb{E}_{\boldsymbol{\sigma}, \tilde{\boldsymbol{\sigma}}} \exp\left(\frac{\alpha + \epsilon_n}{2n} \langle \boldsymbol{\sigma}, \tilde{\boldsymbol{\sigma}} \rangle^2\right).
\end{aligned}
$$

By CLT, $n^{-1/2} \langle \boldsymbol{\sigma}, \tilde{\boldsymbol{\sigma}} \rangle \xrightarrow{d} Z \sim N(0,1)$. So,

$$
\begin{aligned}
\mathbb{E}_{\boldsymbol{\sigma}, \tilde{\boldsymbol{\sigma}}} \exp\left(\frac{\alpha + \epsilon_n}{2n} \langle \boldsymbol{\sigma}, \tilde{\boldsymbol{\sigma}} \rangle^2\right) &\to \mathbb{E} e^{\frac{\alpha}{2} Z^2} \\
&= \begin{cases} (1-\alpha)^{-1/2} & \text{if } \alpha < 1 \\ +\infty & \text{if } \alpha \geq 1. \end{cases}
\end{aligned}
$$

By hypothesis, $\alpha < 1$. Therefore, $\chi^2(H_1 \| H_0) + 1$ is bounded. □

Estimates for a and b in partial recovery regime $\frac{a-b}{2} > \sqrt{d}$

Enough to estimate $s = \frac{a-b}{2}$ and $d = \frac{a+b}{2}$. Since $d > s > \sqrt{d}$, both $s, d > 1$.

If $G_n \sim G(n, d/n)$ or $G(n, a/n, b/n)$, then $\hat{d}_n = \frac{2\#\text{edges}}{n} \to d$.

A k-cycle: $v_1 \sim v_2 \sim \cdots \sim v_k \sim v_1$ where $v_1, \ldots, v_k$ are distinct. Let X_k be the number of k-cycles (modulo cyclic shifts and orientation).

When $n \to \infty$ and $k \le (\log n)^{1/4}$

$$G_n \sim G(n, d/n): \qquad X_k \overset{d}{\approx} \mathrm{Poi}\left(\frac{d^k}{2k}\right) \quad \approx \frac{d^k}{2k} + O\left(\sqrt{\frac{d^k}{2k}}\right).$$

$$G_n \sim G(n, a/n, b/n): \qquad X_k \overset{d}{\approx} \mathrm{Poi}\left(\frac{d^k + s^k}{2k}\right) \quad \approx \frac{d^k + s^k}{2k} + O\left(\sqrt{\frac{d^k + s^k}{2k}}\right).$$

Suppose $G_n \sim G(n, a/n, b/n)$. If $1 \ll k \le (\log n)^{1/4}$, then

$$\hat{s}_n = (2kX_k - \hat{d}_n^k)^{1/k} \to s.$$

Exploiting the sparseness of G_n, X_k can be evaluated in polynomial time.

First moment calculation

Suppose $G \sim G(n, d/n)$.

$$\mathbb{E}[X_k] = \binom{n}{k} \cdot k! \cdot \frac{1}{2k} \cdot \mathbb{P}(v_1 \sim v_2 \sim \cdots \sim v_k \sim v_1)$$
$$= \binom{n}{k} \cdot k! \cdot \frac{1}{2k} \cdot \left(\frac{d}{n}\right)^k \sim \frac{d^k}{2k}.$$

Suppose $G \sim G(n, a/n, b/n)$. A similar calculation shows

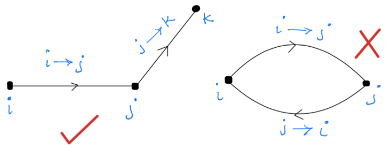
$$\mathbb{E}[X_k] \sim \frac{d^k + s^k}{2k}.$$

$$\mathbb{P}(v_1 \sim v_2 \sim \cdots \sim v_k \sim v_1) = n^{-k}(s^k + d^k).$$

In the partial recovery regime $(a-b)/2 > \sqrt{d}$, the spectral method fails for adjacency matrix. However, the spectral method works for a new matrix called *nonbacktracking matrix*.

Let $G = (V, E)$ be undirected graph. For each $(i, j) \in E$, form two directed edges $i \to j$ and $j \to i$. The non-backtracking matrix B is a 2|E| × 2|E| matrix such that

$$B_{i\to j, k\to l} = \begin{cases} 1 & \text{if } j = k, i \neq l \\ 0 & \text{otherwise.} \end{cases}$$



$(A^r)_{i,j} = \#$ walks of $r+1$ vertices starting from i and ending at j.

$(B^r)_{i\to j, k\to l} = \#$ non-backtracking walks of $r+1$ directed edges starting from $i \to j$ and ending at $k \to l$.

Some properties of non-backtracking matrix

- B is not symmetric. So, its eigenvalues are complex-valued in general.

  Perron-Frobenius theorem. $\lambda_1 \geq |\lambda_2| \geq \cdots \geq |\lambda_{2m}| \qquad (m = \#$ edges).

- Spectrum of B is given by Ihara-Bass-Hashimoto identity:

$$\det(I - zB) = (1 - z^2)^{|E|-|V|} \det(I - zA + z^2(D - I)),$$

  where $D = \mathrm{diag}(\deg(1), \ldots, \deg(n))$ is the diagonal degree matrix.

- B has $2(m-n)$ eigenvalues ± 1 (non-informative). The rest of the $2n$ eigenvalues are informative.

- If the graph is d-regular, then $D = dI$. Then

$$\mathrm{eig}(B) = \{\pm 1\} \cup \{\lambda : \lambda^2 - \lambda\mu + (d-1) = 0, \mu \in \mathrm{eig}(A)\}.$$

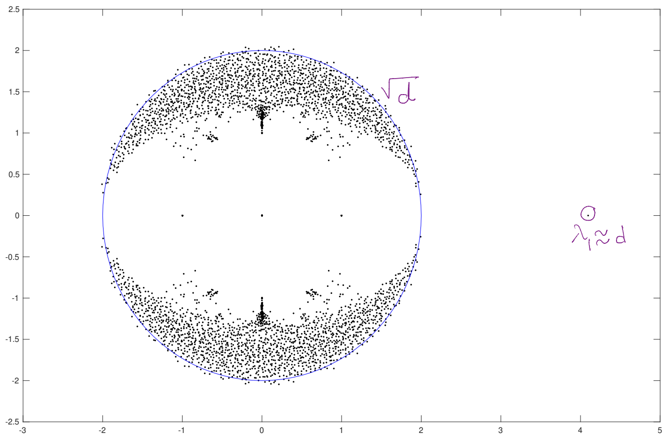
Theorem

Let $d > 1$. The following events happen with high probabilities.

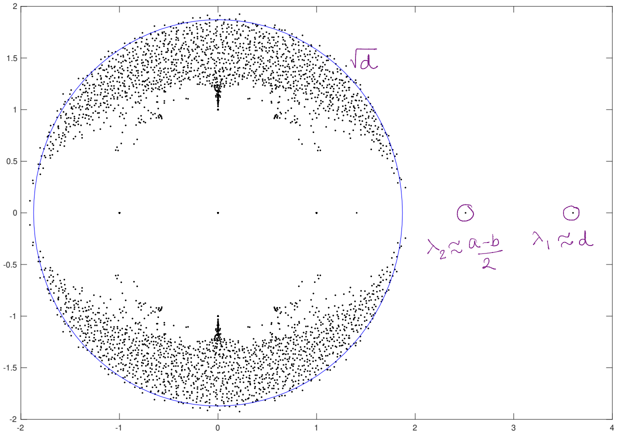
(a) $G(n, d/n)$: B has a single eigenvalue close to d. The remaining eigenvalues are within disk $\{z : |z| \leq \sqrt{d} + \epsilon\}$.

(b) SBM $G(n, a/n, b/n)$ with $\frac{a-b}{2} > \sqrt{d}$: B has two eigenvalues close to d and $\frac{a-b}{2}$. The remaining eigenvalues are within disk $\{z : |z| \leq \sqrt{d} + \epsilon\}$.

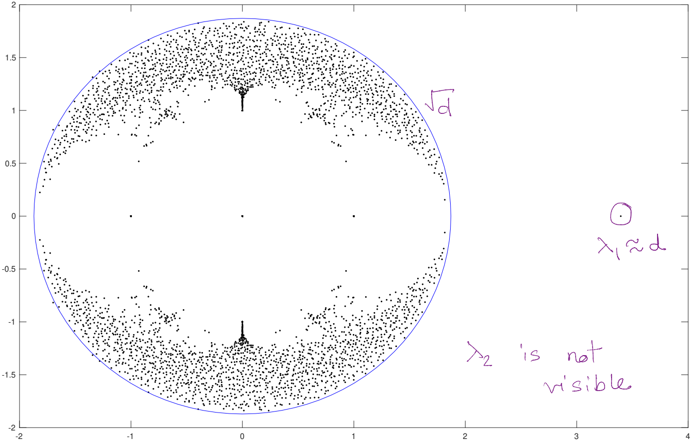
Eigenvalues of NB matrix of $G(n, d/n)$ with $n = 2000$ and $d = 4$



Eigenvalues of NB matrix of $G(n, a/n, b/n)$ with $n = 2000$ and $a = 6, b = 1$



Eigenvalues of NB matrix of $G(n, a/n, b/n)$ with $n = 2000$ and $a = 4, b = 3$



Spectral Redemption

Consider SBM $G(n, a/n, b/n)$ with $\frac{a-b}{2} > \sqrt{d}$. Let ξ be the eigenvector of B corresponding to eigenvalue $\lambda_2 \approx \frac{a-b}{2}$. Define

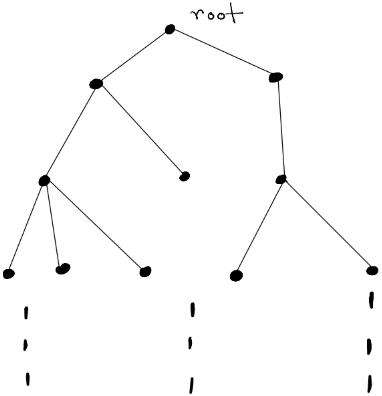
$$\hat{\sigma}_v := \mathrm{sgn}\Big(\sum_{u:u\sim v} \xi_{u\to v} \Big)$$

Theorem

There exists $c > 0$ such that with high probability,

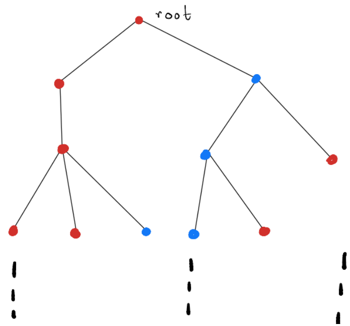
$$\frac{1}{n}\sum_{v} \mathbf{1}(\hat{\sigma}_v = \sigma_v) \geq \frac{1}{2} + c \quad \textit{for large } n.$$

The local neighborhood of a random vertex of $G(n, d/n)$ looks like a Galton-Watson tree where each vertex has an independent $\mathrm{Poi}(d)$ many children.



The local neighborhood of a random vertex of $G(n, a/n, b/n)$ looks like a multi-type Galton-Watson tree.

- The root is red or blue with probability $1/2$.
- Recursively, each vertex gives birth to a $\mathrm{Poi}(a/2)$ vertices of the same color and a $\mathrm{Poi}(b/2)$ vertices of the different color (red or blue).



- Alternate Description. Generate a Galton-Watson tree with $\mathrm{Poi}(d)$ offspring distribution. The root is red or blue with probability $1/2$. The color each children is same as its parent with probability $\frac{a}{a+b}$ and opposite with probability $\frac{b}{a+b}$, independent of other individuals.

Kesten-Stigum threshold

For a vertex v of the tree, let us define

$$\sigma_v = \begin{cases} +1 & \text{if } v \text{ is red} \\ -1 & \text{if } v \text{ is blue} \end{cases}$$

and

$$\mathrm{Maj}_r = \mathrm{sgn}\Big(\sum_{\mathrm{d(root,v)=r}} \sigma_v\Big),$$

i.e., if $\mathrm{Maj}_r = 1$ if the majority of the vertices at depth r are red and $\mathrm{Maj}_r = -1$ otherwise.

Fact

- If $\frac{a-b}{2} > \sqrt{d}$ then there exists $c > 0$

$$\mathbb{P}(\sigma_{\text{root}} = \mathrm{Maj}_r) \geq \frac{1}{2} + c, \qquad \text{for large } r.$$

- If $\frac{a-b}{2} \leq \sqrt{d}$ then

$$\lim_{r\to\infty} \mathbb{P}(\sigma_{\text{root}} = \mathrm{Maj}_r) = \frac{1}{2}.$$

Approximation of second eigenvector of B assuming $s > \sqrt{d}$

Let $s = \frac{a-b}{2}$. Define

$$\xi^{(r)}_{u\to v} = s^{-r} \sum_{d(u\to v, x\to y)=r} \sigma_y.$$

- We will show that $\xi^{(r)}$ is an approximate eigenvector of B with approximate eigenvalue $s = \frac{a-b}{2}$ for large r.
- From multi-type Galton-Watson approximation and Kesten-Stigum bound, for a random vertex v and $u \sim v$,

$$\mathbb{P}(\sigma_v = \mathrm{sgn}(\xi^{(r)}_{u\to v})) \geq \frac{1}{2} + c \quad \text{for large } r,$$

  which implies that

$$\mathbb{P}(\sigma_v = \mathrm{sgn}(\sum_{u:u\sim v} \xi^{(r)}_{u\to v})) \geq \frac{1}{2} + c' \quad \text{for large } r.$$

$$(B\xi^{(r)})_{u\to v} = s^{-r} \sum_{d(u\to v, x\to y)=r+1} \sigma_y = s \cdot \xi^{(r+1)}_{u\to v}$$

or,

$$B\xi^{(r)} = s \cdot \xi^{(r+1)}.$$

$$\xi^{(r)}_{u\to v} - \xi^{(r+1)}_{u\to v} = s^{-r} \sum_{d(u\to v, x\to y)=r} \underbrace{\Big(\sigma_y - s^{-1} \sum_{z\sim y, z\neq x} \sigma_z\Big)}_{=:V_y}$$

There are d^r many terms in the sum on average.

Given the spins of the vertices at depth r from v, the random variables V_y's are mean zero and of constant variance. Therefore,

$$\mathbb{E}\Big(\xi^{(r)}_{u\to v} - \xi^{(r+1)}_{u\to v}\Big)^2 \leq Cs^{-2r} d^r \approx 0,$$

under the assumption that $s > \sqrt{d}$ and r is large.

So, $\xi^{(r)} \approx_{r\to\infty} \xi^{(\infty)}$ and $B\xi^{(\infty)} \approx s\xi^{(\infty)}$.

Almost all eigenvalues satisfy $|\lambda| \leq \sqrt{d}$

Let $\lambda_1, \ldots, \lambda_{2m}$ be the eigenvalues of B.

For any $k \geq 1$

$$\begin{aligned}\frac{1}{2m}\sum_i |\lambda_i|^{2k} &= \frac{1}{2m}\sum_i |\lambda_i^k|^2 \leq \frac{1}{2m}\mathrm{tr}(B^k(B^k)^T) \\ &= \frac{1}{2m}\sum_{u\to v, x\to y}(B^k)_{u\to v, x\to y}(B^k)^T_{x\to y, u\to v} \\ &= \frac{1}{2m}\sum_{u\to v, x\to y}(B^k)_{u\to v, x\to y}(B^k)_{y\to x, v\to u}.\end{aligned}$$

In the last line, we used the fact $B^T_{u\to v, x\to y} = B_{v\to u, y\to x}$. Consequently, $(B^k)^T_{u\to v, x\to y} = B^k_{v\to u, y\to x}$.

Recall $(B^k)_{u\to v,x\to y}$ counts the number of non-backtracking walks of involving $k+1$ edges from $u \to v$ to $x \to y$.

If the local neighborhood of v is a tree, then there can be at most one such path $u \to v$ to $x \to y$.

$$\sum_{x\to y} (B^k)_{u\to v,x\to y}(B^k)_{y\to x,v\to u}$$

$$= \#x \to y \text{ that are distance } k \text{ from } u \to v \approx d^k.$$

So, with high probability,

$$\frac{1}{2m}\sum_i |\lambda_i|^{2k} \le d^k.$$

This implies that all but a vanishing proportion of the eigenvalues are confined within the disk of radius $\sqrt{d}$.

References

- R Vershynin (2018). High Dimensional Probability : An Introduction with Applications in Data Science.
- C Bordenave, M Lelarge, L Massoulié (2015): Non-backtracking spectrum of random graphs: community detection and non-regular Ramanujan graphs.
- F Krzakala, C Moore, E Mossel, J Neeman, A Sly, L Zdeborova, and P Zhang (2013): Spectral redemption in clustering sparse networks.
- E Mossel, J Neeman, A Sly (2012). Stochastic block models and reconstruction.
- Y. Wu and J. Xu (2018): Statistical problems with planted structures: Information-theoretical and computational limits.