Extreme eigenvalues of Erdős-Rényi random graphs

Florent Benaych-Georges

j.w.w. Charles Bordenave and Antti Knowles

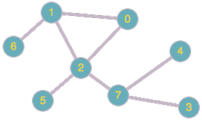
MAP5, Université Paris Descartes

May 18, 2018
IPAM – UCLA

Inhomogeneous Erdős-Rényi random graph

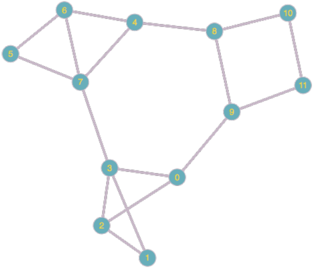
Random graph G :

- vertices $\{1, \ldots, n\}$
- edges $\mathbb{1}_{\{i,j\} \in G} \sim \text{Bernoulli}(p_{ij})$, independent.



Examples :

- homogeneous Erdős-Rényi graph : $p_{ij} = p$.
- Stochastic Block Model. Partition vertices into communities : $\{1, \ldots, n\} = \bigsqcup_\alpha N_\alpha$, p_{ij} depends only on the communities $N_\alpha \ni i$ and $N_\beta \ni j$.



In this talk :

- $A :=$ adjacency matrix of G
- Eigenvalues of A : $\lambda_1(A) \geq \cdots \geq \lambda_n(A)$
- $n \gg 1$, $p_{ij} = p_{ij}(n)$

Example 1 : homogeneous graph ($p_{ij} = d/n$ for all i, j)

Theorem [Krivelevich, Sudakov ; 2003 + Vu ; 2007].

$$\lambda_1(A) \sim \begin{cases} (\log n)^{1/2} & \text{if } d \ll (\log n)^{1/2} \\ d & \text{if } d \gg (\log n)^{1/2} \end{cases}$$

$$\lambda_2(A) \sim \begin{cases} (\log n)^{1/2} & \text{if } d \ll (\log n)^{1/2} \\ 2\sqrt{d} & \text{if } d \gg (\log n)^4 \end{cases}$$

(up to $\log \log n$ factors)

$\hookrightarrow$ What about $\lambda_2(A)$ if $(\log n)^{1/2} \ll d \ll (\log n)^4$? (question related to the **spectral gap**)

Conjecture : transition at $d \sim \log n$ (*graph connectivity threshold*)

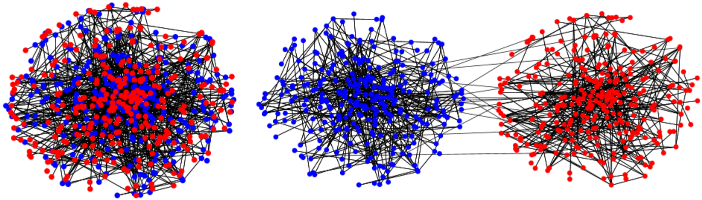
$A = A - \mathbb{E}A + \mathbb{E}A$ hence by Weyl's interlacing inequality :

$$\cdots \lambda_3(A) \leq \lambda_2(A - \mathbb{E}A) \leq \lambda_2(A) \leq \lambda_1(A - \mathbb{E}A) \leq \lambda_1(A)$$

Example 2 : Stochastic Block Model (SBM)

Partition vertices into communities : $\{1, \ldots, n\} = \bigsqcup_\alpha N_\alpha$, p_{ij} depends only on the communities $N_\alpha \ni i$ and $N_\beta \ni j$.

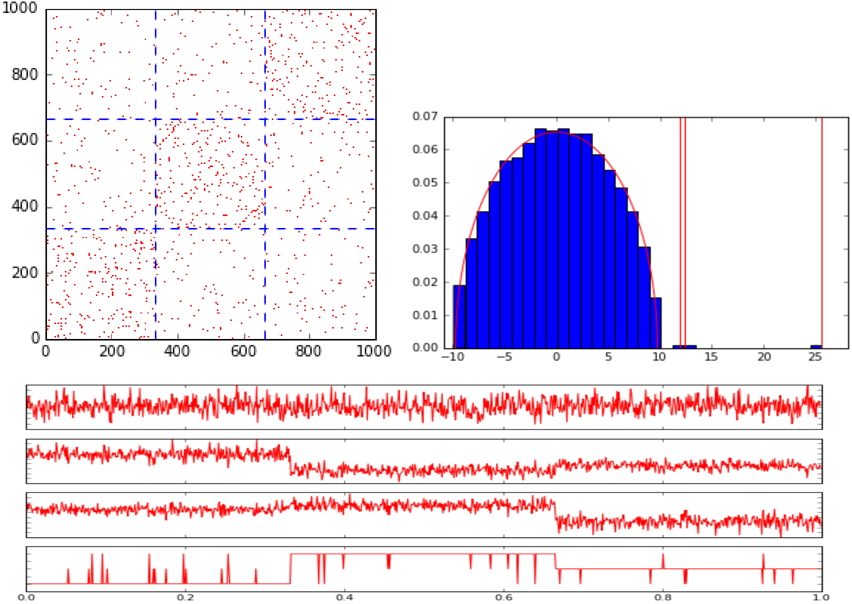
Question : how to recover the communities from the graph observation ?



Spectral clustering algorithm :

- Spectral decomposition of A : $\lambda_1 \geq \lambda_2 \cdots$, $\overrightarrow{V_1}$, $\overrightarrow{V_2}, \ldots$
- $\lambda_1, \ldots, \lambda_r$: **isolated eigenvalues**
- Apply k-means or EM to the rows of the $n \times r$ matrix $\left[\overrightarrow{V_1} \quad \cdots \quad \overrightarrow{V_r}\right]$

SBM $n = 1000$, 3 classes (3.5% of misclustered vertices)



When and why does spectral clustering work well ?

Spectral clustering algorithm :

- Spectral decomposition of A : $\lambda_1 \geq \lambda_2 \cdots, \overrightarrow{V_1}, \overrightarrow{V_2}, \ldots$
- $\lambda_1, \ldots, \lambda_r$: **isolated eigenvalues**
- Apply k-means or EM to the rows of the $n \times r$ matrix $\left[\overrightarrow{V_1} \quad \cdots \quad \overrightarrow{V_r}\right]$

Usually, **spectral clustering works well** when $\|A - \mathbb{E}A\| \ll \|\mathbb{E}A\|$

Explanation :

a) Eigenvectors of $\mathbb{E}A$ is where the information about communities lies

b) $A = \mathbb{E}A + (A - \mathbb{E}A)$

c) By perturbation theory (Davis-Kahan theorem),

$$\|A - \mathbb{E}A\| \ll \|\mathbb{E}A\| \text{ and } \lambda_i(A) \text{ isolated } \implies \overrightarrow{V_i(A)} \approx \overrightarrow{V_i(\mathbb{E}A)}$$

(or use BBP if $\|A - \mathbb{E}A\| \leq c\|\mathbb{E}A\|$ for c of order one, large enough)

Main results

Conjecture in example 1 (homogeneous graph) and spectral clustering efficiency assessment in example 2 (SBM) both lead to the study of the largest eigenvalues of $A - \mathbb{E}A$.

Hypothesis : G is an inhomogeneous Erdős-Rényi random graph with :

- constant mean degree d (*i.e.* for all i, $\sum_j p_{ij} = d$),
- $\max_{i,j} p_{i,j} \leq n^{-1+\kappa}$, for some $\kappa > 0$ small.

Theorem 1. [BBK ; 2017] W.h.p.,

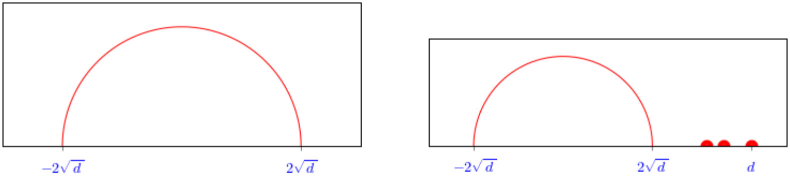
$$\frac{\|A-\mathbb{E}A\|}{\sqrt{d}} \le 2 + C\frac{\eta}{\sqrt{1 \vee \log \eta}}, \qquad \text{with} \qquad \eta := \sqrt{\frac{\log n}{d}}.$$

Bound $\frac{\|A-\mathbb{E}A\|}{\sqrt{d}} \le 2 + O(\eta^{2/3})$ *also in a recent preprint by Latała, van Handel and Youssef.*

Consequence : If $d \gg \log n$, then empirical spectral distributions

$$\frac{1}{n}\sum_{i=1}^{n} \delta_{\lambda_i}$$

of $A - \mathbb{E}A$ and A look like :



(a) **Centered matrix** $A-\mathbb{E}A$ **:** semicircle law with **no outlier**

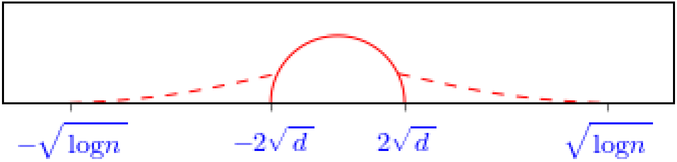
(b) **Adjacency matrix** A **:** semicircle law with **outliers** at the positions of the eigenvalues of (usual) order d of $\mathbb{E}A$

Theorem 2. [BBK ; 2017] Define the **ordered degrees** D_i through $D_1^\downarrow \geq D_2^\downarrow \geq \cdots \geq D_n^\downarrow$.

For $d \ll \log n$, for any $k \leq n^{0.99}$, we have

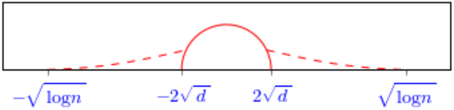
$$\lambda_k(A - \mathbb{E}A) \ \sim \ \sqrt{D_k^\downarrow} \ \sim \ \sqrt{\frac{\log(n/k)}{\log((\log n)/d)}}$$

Consequence :

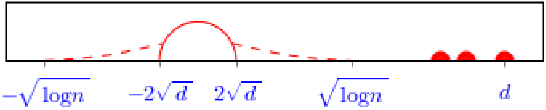


Centered matrix $A - \mathbb{E}A$ **:** spectrum distributed according to the semicircle law, plus $n^{0.99}$ eigenvalues in a cloud up to $\sqrt{\log n}$

Consequence for the **adjacency matrix** A when $d \ll \log n$:



$d \ll \sqrt{\log n}$: no outlier



$d \gg \sqrt{\log n}$: outliers at the positions of the eigenvalues of (usual) order d of $\mathbb{E}A$

Corollary 1. For homogeneous Erdős-Rényi graphs, the transition mentioned above is actually at $d \sim \log n$.

Corollary 2. In the SBM, when the non zero eigenvalues of $\mathbb{E}A$ have order d and have gaps with order d, spectral clustering works well if and only if $d \gg \sqrt{\log n}$.

Interpretation of the case $d \ll \log n$

$$\forall k \leq n^{0.99}, \quad \lambda_k(A - \mathbb{E}A) \ \sim \ \sqrt{\log(n/k)}$$

Consequence : asymptotic "density" of eigenvalues of $A/\sqrt{\log n}$ at $x \in (0, 1)$:

$$2\,(\log n)\,n^{1-x^2}\,x$$

Previously, in random matrix theory, two types of behaviour have been observed for extreme eigenvalues (out of finite sets of outliers at deterministic positions) :

(a) convergence to the edge of the limit support, usually with Tracy-Widom fluctuation (e.g. Wigner matrices) ;

(b) convergence (after rescaling) to a Poisson point process (e.g. heavy-tailed random matrices).

Our theorem 2 implies that neither is true here when $d \ll \log n$. In fact, there is no deterministic sequence α_n such that the point process

$$\{\alpha_n \lambda_k(A) : 2 \leq k \leq n^{0.99}\}$$

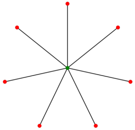
converges to a nondegenerate process.

Proof of the $d \ll \log n$ case

We want to prove that for $k \leq n^{0.99}$, $\lambda_k(A - \mathbb{E}A) \sim \sqrt{D_k^{\downarrow}}$, with $D_k^{\downarrow}$ the k-the largest degree.

Remark. Spectrum of a star-graph with degree D :

$$-\sqrt{D}, \sqrt{D}, 0, 0, \ldots.$$



Lemma. $D_k^{\downarrow} \sim \frac{\log(n/k)}{\log((\log n)/d)}$.

Lemma. Let G' be the graph of the $n^{0.99}$ largest stars, where we have removed all edges joining centers of stars in 1 or 2 steps. Then w.h.p.,

- the stars present in G' are disjoint and have degrees $D_k^{\downarrow} - O(1)$
- the matrix $A - \mathbb{E}A$ rewrites

$$A - \mathbb{E}A = \mathrm{Adj}(G') + (\text{matrix with norm} \ll \sqrt{\frac{\log n}{\log \log n}})$$

(Le, Levina, Vershynin).

Then use perturbation inequalities to prove $\lambda_k(A - \mathbb{E}A) \sim \sqrt{D_k^{\downarrow}}$.

Proof of the $d \gg \log n$ case

Goal : estimate $\|H\|$ with $H := d^{-1/2}(A - \mathbb{E}A)$.

Let $\lambda = (\lambda_1, \dots, \lambda_n)$ be the eigenvalues of H.

Goal : $\|\lambda\|_{\ell^\infty}$. Much easier to estimate $\|\lambda\|_{\ell^p}$ for large (even) p.

Elementary fact : $\|\lambda\|_{\ell^p}$ is close to $\|\lambda\|_{\ell^\infty}$ if $p \gg \log n$.

Thus, we have to estimate $\|\lambda\|_{\ell^p}$ for $p \gg \log(n)$. More precisely,

$$\mathbb{E}\|\lambda\|_{\ell^p}^p = \mathbb{E}\,\mathrm{Tr}\,H^p = \mathbb{E} \sum_{1 \leq i_1, \dots, i_p \leq n} H_{i_1 i_2} H_{i_2 i_3} \cdots H_{i_p i_1}\,.$$

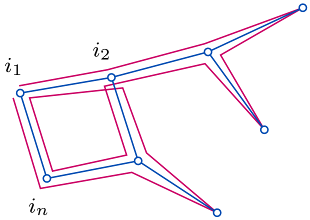
Right-hand side analyzed using that entries of H are independent and have mean zero (Füredi-Komlós-approach).

Use graphs to encode terms arising from $\sum_{i_1,\ldots,i_p} \mathbb{E}(H_{i_1 i_2} H_{i_2 i_3} \cdots H_{i_p i_1})$.

Vertices $= \{i_1, i_2, \ldots, i_p\}$.

Edges $= \{\{i_k, i_{k+1}\} : k = 1, \ldots, p\}$.

Each non zero term : a walk $i_1, i_2, \ldots, i_p$ of length p on a graph, such that each edge is visited at least twice.

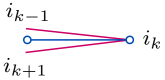


Works well when $d \gg (\log n)^4$, but a fundamental problem arises otherwise : proliferation of subtrees, which leads to very complicated combinatorics.

Nonbacktracking matrix

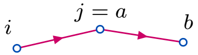
To make combinatorics manageable, we try to kill all subtrees.

Key observation : each leaf of the graph gives rise to a backtracking piece of the walk : $i_{k-1} = i_{k+1}$.



To H we associate its nonbacktracking matrix $B = (B_{ef})_{e,f \in \{1,\dots,n\}^2}$ indexed by directed edges :

$$B_{(i,j)(a,b)} := H_{ab} \, \mathbb{1}_{j=a} \, \mathbb{1}_{i \neq b} \,.$$



Note that B is $n^2 \times n^2$ and non-Hermitian.

$\mathbb{E} \operatorname{Tr} B^p (B^*)^p$ can be written in terms of walks on graphs with no subtrees : *nonbacktracking walks*.

Estimates of $\mathbb{E}\,\mathrm{Tr}\,B^p(B^*)^p$ (made possible by the non-backtracking structure) give :

Proposition 1 [BBK]. W.h.p.,

$$\rho(B) \;:=\; \text{(Spectral Radius of } B\text{)} \;\leq\; 1 + \frac{C}{\sqrt{d}}.$$

Proposition 2 [BBK, Ihara-Bass type formula]. We have

$$\|H\| \leq \|H\|_{2\to\infty}\, f\left(\frac{\rho(B)}{\|H\|_{2\to\infty}}\right) + 7\|H\|_{1\to\infty}\,,$$

for $f(x) := 2 \cdot \mathbb{1}_{x\leq 1} + \left(x + \frac{1}{x}\right) \cdot \mathbb{1}_{x\geq 1}$ and

$$\|H\|_{2\to\infty} := \max_i \sqrt{\sum_j |H_{ij}|^2}\,, \qquad \|H\|_{1\to\infty} := \max_{i,j}|H_{ij}|\,.$$

Lemma. If $d \gg \log(n)$, then $\|H\|_{2\to\infty} \approx 1$ (*Bennett's inequality)*

Consequence :
As $\|H\|_{1\to\infty} \ll 1$, $\|H\| \;\lesssim\; 2$ and Theorem 1 follows.