Concentration of random graphs and community detection in networks

Liza Levina

Department of Statistics, University of Michigan

Roman Vershynin (Mathematics, University of California, Irvine)
Can M. Le (Statistics, University of California, Davis)
Outline

- Motivating application: communities in networks
- Concentration results
- Implications for community detection
What are networks?

- Mathematics: a graph (a set of nodes and edges)
- Statistics: data on relationships between entities
What are networks?

- Mathematics: a graph (a set of nodes and edges)
- Statistics: data on relationships between entities

Examples
- Facebook: nodes = users, edges = friend status
- Ecosystems: nodes = animal species; edges = predator/prey relationship
- Transportation: nodes = airports, edges = direct flights
- Political science: nodes = countries, edges = trade, deals, or conflicts
A network with $n$ nodes $\iff$ $n \times n$ adjacency matrix $A$:

$$A_{ij} = 1 \text{(there is an edge from } i \text{ to } j)$$

**Undirected network:** $A = A^T$.

**Self-loops:** an edge from a node to itself ($A_{ii} = 1$), impossible in some contexts, but it is convenient to ignore them.

$$A = \begin{pmatrix}
0 & 1 & 1 & 0 & 0 \\
1 & 0 & 1 & 1 & 1 \\
1 & 1 & 0 & 1 & 1 \\
0 & 1 & 1 & 0 & 1 \\
0 & 1 & 1 & 1 & 0
\end{pmatrix}$$
What are communities?

- Nodes within a community are in some sense more similar to each other than to other nodes.
- Often communities are thought of as tightly knit groups: more links within, fewer links between.
- Other kinds of communities are possible (e.g., genders in a dating network).
- **Community detection**: assigning each node to a community, typically finding a partition of the network.
- Statistical approach: fitting probabilistic models with communities to network data.
The karate club

Zachary (1977)
The dolphin network

Lusseau et al. (2003)
The US politics blogs

Adamic & Glance (2005)
Think of $A$ as a random binary matrix
Independence of edges is commonly assumed
Exchangeable networks: the distribution of $A$ is the same for any permutation of node labels
The information of interest is contained in $P = \mathbb{E}A$
Since we only have a single realization of $A$, we need models with some structure in $\mathbb{E}A$ (e.g. communities)
Need to understand properties of both models and fitting algorithms, under different scaling regimes determined by $\rho_n$,

$$P_n = \rho_n P$$
The Erdös-Rényi (ER) graph $G(n,p)$

- Entries $\{A_{ij} : i \leq j\}$, are i.i.d. Bernoulli variables:
  \[ p = \mathbb{P}(A_{ij} = 1) \]

- Undirected graph: $A^T = A$

- All nodes have the same expected degree,
  \[ d = \mathbb{E}d_i = \mathbb{E}\sum_j A_{ij} = np = n\rho_n \]

- The distribution of node degrees is binomial
The Erdős-Rényi (ER) graph $G(n, p)$

- Entries $\{A_{ij} : i \leq j\}$, are i.i.d. Bernoulli variables:

  $$p = \mathbb{P}(A_{ij} = 1)$$

- Undirected graph: $A^T = A$

- All nodes have the same expected degree,

  $$d = \mathbb{E}d_i = \mathbb{E}\sum_j A_{ij} = np = n\rho_n$$

- The distribution of node degrees is binomial

- This does not happen in real networks
Some properties of the ER graph

- If $d = np < 1$, then the graph has no connected components of size larger than $O(\log n)$ a.s.
- If $d = 1$, the largest connected component (LCC) is of order $O(n^{2/3})$ nodes a.s.
- If $d > 1$ and is a constant, the graph a.s. has one giant component (LCC of size $O(n)$), and no other component has more than $O(\log n)$ nodes.
- If $d < (1 - \varepsilon)\log n$, the graph a.s. contains isolated nodes.
- If $d > (1 + \varepsilon)\log n$, the graph is a.s. connected.
Different asymptotic regimes

- $d \leq 1$: not usually considered.
- $d = \text{const} > 1$: **sparse**. The fraction of isolated nodes does not vanish.
- $d \to \infty$ slower than $\log n$: **semi-sparse**. The graph is a.s. disconnected.
- $d \gtrsim \log n$: **semi-dense**. The graph is a.s. connected.
- $d \sim n$: **dense**.

Generally, real world networks are relatively sparse.
The stochastic block model (SBM)

Holland, Laskey, Leinhardt (1983)

- Node labels \( \{c_i, i = 1, \ldots, n\} \) are independent multinomial with
  \[
  \mathbb{P}(c_i = k) = \pi_k, \ k \in \{1, \ldots, K\}
  \]

- Given the label vector \( c \), \( \{A_{ij}, i \leq j\} \) are independent Bernoulli with
  \[
  \mathbb{P}(A_{ij} = 1 | c) = P_{ij} = B_{ci,cj}
  \]
  where \( B \) is a \( K \times K \) symmetric matrix

- \( A \) is observed, \( c \) is not observed

Estimate \( c, B, \pi \) from \( A \)
Spectral properties of the SBM

- Let $z_i$ be a vector of length $K$ with the $k$-th coordinate given by
  \[ z_{ik} = \mathbf{1}(c_i = k) \]

- Arrange all $n$ vectors $z_i$ as rows in a $n \times K$ matrix $Z$; then
  \[ \mathbb{E}A = P = ZBZ^T \]
A simple example

- The **balanced planted partition model**: \( K = 2, n_1 = n_2 = n/2, \)

\[
B = \begin{bmatrix}
p & q \\
q & p \\
\end{bmatrix}
\]

- The eigenvalues and eigenvectors of \( P \) are

\[
\lambda_1 \propto p + q, \quad u_1 \propto \begin{pmatrix}
1 \\
\vdots \\
1 \\
1 \\
\end{pmatrix} \\
\lambda_2 \propto p - q, \quad u_2 \propto \begin{pmatrix}
+1 \\
\vdots \\
+1 \\
-1 \\
-1 \\
\end{pmatrix}
\]

- The **second eigenvector** contains community information
- In general, communities are encoded in the first \( K \) eigenvectors
Spectral clustering on the adjacency matrix

- Arrange the $K$ leading eigenvectors of $A$ in a $n \times K$ matrix $U$.
- Run the $K$-means clustering algorithm on the rows of $U$. 

Liza Levina (University of Michigan) 
Concentration of random graphs
Spectral clustering on the adjacency matrix

- Arrange the $K$ leading eigenvectors of $A$ in a $n \times K$ matrix $U$
- Run the $K$-means clustering algorithm on the rows of $U$

The $K$-means algorithm

0. Initialize $K$ cluster centers with randomly selected points
1. Assign each data point to the nearest center
2. Update centers to new clusters’ means
3. Alternate steps 1 and 2 until convergence
Spectral clustering on the graph Laplacian

- Graph Laplacian can be defined in multiple ways, e.g. through diffusion or random walks on the graph.
- The symmetric normalized Laplacian is given by $I - D^{-1/2}AD^{-1/2}$, where $D = \text{diag}(d_1, \ldots, d_n)$.
- For spectral clustering, can equivalently look at $L = D^{-1/2}AD^{-1/2}$.

In theory, spectral clustering on both $L$ and $A$ give correct results under the SBM.

In practice, degrees are often more heterogeneous than the SBM allows; then spectral clustering on $L$ tends to work better than on $A$. 
When does spectral clustering work?

1. **Model**: communities are encoded in the eigenvectors of $E_A$

2. **Concentration**: $A$ concentrates around $E_A$ in spectral norm, i.e., $||A - E_A||$ not too large

3. **Davis-Kahan theorem**: if $A$ concentrates around $E_A$ in matrix spectral norm, then communities can be recovered from the eigenvectors of $A$, because

$$\text{Eigenstructure}(A) \approx \text{Eigenstructure}(E_A)$$
The challenge of concentration: sparsity

- Understanding concentration of $A$ and $L$ is key to understanding behavior of spectral clustering.
- Spectral clustering tends to fail if the network is sparse.
- Below some level of density, every method will fail.
- **Key questions**
The challenge of concentration: sparsity

- Understanding concentration of $A$ and $L$ is key to understanding behavior of spectral clustering
- Spectral clustering tends to fail if the network is sparse
- Below some level of density, every method will fail

Key questions

When does $A$ concentrate?
The challenge of concentration: sparsity

- Understanding concentration of $A$ and $L$ is key to understanding behavior of spectral clustering.
- Spectral clustering tends to fail if the network is sparse.
- Below some level of density, every method will fail.
- **Key questions**
  - When does $A$ concentrate?
  - If it does not, can we fix it?
Back to the simple example

SBM with $K = 2$, $n_1 = n_2$, 

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

Key quantities:

- **Expected node degree** $d = n(p + q)/2$: the larger $d$ is, the easier the community detection problem

- **Out-in-ratio** $\beta = q/p$: the smaller $\beta$ is, the easier the community detection problem

- **Number of nodes** $n$: the larger $n$ is, the easier the community detection problem ($K$ is fixed!)

Fix $\beta = 0.05$, vary $d$ and $n$
Networks \((n = 800, \beta = 0.05)\)
Adjacency matrices

\[ d = 1 \quad d = 2 \quad d = 3 \quad d = 5 \quad d = 7 \quad d = 14 \]
Adjacency matrices

\[ d = 1 \quad d = 2 \quad d = 3 \quad d = 5 \quad d = 7 \quad d = 14 \]
Adjacency matrices

d = 1

d = 2

d = 3

d = 5

d = 7

d = 14

nz = 822

nz = 1674

nz = 2426

nz = 4074

nz = 5462

nz = 11092
Spectral clustering fails on sparse graphs. Why?

\[ n = 80, \ d = 2 \]
The leading eigenvectors of the Laplacian

\( n = 80, \ d = 2 \)
What about the largest connected component?

\[ n = 80, \ d = 2 \]
Restrict to the largest connected component?

\[ n = 800, \ d = 3, \ |LCC| = 737 \]
Leading eigenvectors of the LCC

$n = 800, \ d = 3, \ |\text{LCC}| = 737$
Where is the problem now?
Recall key requirement: \(\|A - \mathbb{E}A\|\) not too large (relative to \(\|\mathbb{E}A\|\))

A lot of work has been done: see the proceedings paper for history

Inhomogeneous Erdös-Rényi model \(G(n, [P_{ij}])\) (IERM): edges are formed independently with possibly different probabilities \(P_{ij}\)

IERM is the most general model with independent edges; includes ER, SBM, and many other network models.
Dense networks concentrate

(with high probability)

- For the ER graph $G(n, p)$, with $d = np$, Bai-Yin law implies
  \[ \|A - \mathbb{E}A\| \leq 2\sqrt{d} \]

- For the IERM $G(n, (P_{ij}))$, let $d = \max_i \sum_j p_{ij}$. Then $A$ concentrates:
  \[ \|A - \mathbb{E}A\| \lesssim \sqrt{d} + \sqrt{\log n} \]
  \[ \|\mathbb{E}A\| \sim d \]

- In particular, if $d \gtrsim \log n$, then optimal concentration holds:
  \[ \|A - \mathbb{E}A\| \lesssim \sqrt{d} \]
Sparse networks do not concentrate

Even for the ER graph $G(n, p)$, if $d = np \ll \log n$, then

$$\|A - EA\| \gg d = \|EA\|$$

This happens because w.h.p. there is a node with a large degree, $O(\log n / \log \log n)$

If $d = O(1)$ then $\|EA\| \sim O(1)$, while

$$\|A - EA\| \gtrsim \sqrt{\log n}$$
For $A$, the problem is caused by high degree nodes

In theory, removing all nodes with degrees greater than, for example, $2\bar{d}$, makes $A$ concentrate.

In practice, removing edges from an already sparse network is a terrible idea.

Could instead change weights of existing edges to keep degrees bounded.

Spectral clustering on the Laplacian tends to work better.
Proof idea

1. A concentrates on $N_0$ in the weaker $\ell_\infty \to \ell_2$ norm with no regularization
2. Upgrade to concentration in spectral norm using Grothendieck factorization theorem
3. Show $A$ is well behaved on $R_0$ and $C_0$ because of their small number of rows/columns
4. Apply the same trick to the remainder and repeat to obtain concentration on the whole network

(a) First step

(b) Iterations

(c) Final
Concentration of the Laplacian

\[ L = L(A) = D^{-1/2}AD^{-1/2} \]

- Just like $A$, $L(A)$ concentrates when $d \gtrsim \log n$, and fails to concentrate if $d \ll \log n$.
- Isolated nodes and "dangling trees" cause problems for $L(A)$, and when $d \ll \log n$, the probability of an isolated node is non-vanishing.
- Conversely, $A$ fails to concentrate because of high-degree nodes, but the end result is the same.
Multiple methods proposed; we focus on *Amini et al (2013)*

Replace $A$ with

$$
A' = A + \rho \frac{\bar{d}}{n} 11^T, \quad \rho \in (0, 1)
$$

where $\bar{d}$ is the average node degree

Apply the usual spectral clustering to $L(A')$

Intuition: adding a weak connection between every pair of nodes removes small components
Concentration of the regularized Laplacian

\[ A_\tau = A + \frac{\tau}{n} 11^T \]

- Under IERM, as long as \( \tau \sim n \max_{ij} p_{ij} \), w.h.p.,

\[ \|L(A_\tau) - L(EA_\tau)\| = O\left(\frac{1}{\sqrt{d}}\right) \]

- Need to separately bound \( \|L(EA_\tau) - L(EA)\| \), but these are non-random matrices.
Back to example: regularized spectral clustering

\[ n = 800, \ d = 3, \ |CC| = 737 \]

A dangling tree breaks concentration
Back to example: regularized spectral clustering

\[ n = 800, \quad d = 3, \quad |CC| = 737 \]

Spectral clustering (\( \rho = 0 \))
Back to example: regularized spectral clustering

\[ n = 800, \ d = 3, \ |CC| = 737 \]

\[ \rho = 0.1 \]
Back to example: regularized spectral clustering

\[ n = 800, \ d = 3, \ |CC| = 737 \]

\[ \rho = 0.25 \]
Back to example: regularized spectral clustering

\( n = 800, \ d = 3, \ |CC| = 737 \)

\( \rho = 0.4 \)
Back to example: regularized spectral clustering

$n = 800, \ d = 3, \ |CC| = 737$

$\rho = 1$
This mostly follows an excellent review by *Abbe (2018)*

- Let $\hat{c}$ be the estimated node labels, and $c$ true labels
- Define the disagreement between $\hat{c}$ and $c$ by

$$D(\hat{c}, c) = \min_{\sigma} \frac{1}{n} \sum_{i=1}^{n} 1(c_i \neq \sigma(\hat{c}_i))$$

where the minimum is taken over all permutations of labels $1, \ldots, K$

- This is simply the fraction of “misclassified” nodes
Modes of recovery

- **Detectability**, or partial recovery: a constant fraction of vertices can be recovered (more than by random guessing)

\[
P \left( D(\hat{c}, c) < \frac{1}{K} - \varepsilon \right) \to 1
\]

- **Weak consistency**, or almost exact recovery: vanishing fraction of errors

\[
D(\hat{c}, c) \xrightarrow{P} 0
\]

- **Strong consistency**, or exact recovery: almost surely no errors

\[
P(D(\hat{c}, c) = 0) \to 1
\]
Results for symmetric SBM

- Symmetric SBM\((n, K, p, q)\): \(B_{kk} = p\) for all \(k\), and \(B_{kl} = q\) for all \(k \neq l\)
- Communities of equal size; the unbalanced case is substantially harder
- Most results are now available for general \(K\), but easier to state and understand for \(K = 2\)
Results for symmetric SBM with $K = 2$

- **Detectability** for $\text{SBM}(n, 2, \frac{a}{n}, \frac{b}{n})$ holds if and only if
  \[
  \frac{(a - b)^2}{2(a + b)} > 1
  \]

- **Weak consistency** for $\text{SBM}(n, 2, \frac{a}{n}, \frac{b}{n})$ is achievable, and with a polynomial time algorithm, if and only if
  \[
  \frac{(a - b)^2}{2(a + b)} \rightarrow \infty
  \]

- **Strong consistency** for $\text{SBM}(n, 2, \frac{\log n}{n}, \frac{\log n}{n})$ is achievable, and with a polynomial time algorithm, if
  \[
  |\sqrt{a} - \sqrt{b}| > \sqrt{2}
  \]
  and impossible if
  \[
  |\sqrt{a} - \sqrt{b}| < \sqrt{2}
  \]
Results for spectral clustering

- **Sparse** regime, $d = O(1)$: regularized spectral clustering can detect communities up to a fraction $\varepsilon$ as long as $\frac{(a-b)^2}{2(a+b)} > C\varepsilon$, i.e. optimal up to a constant.

- **Semi-sparse** regime, $d \to \infty$: weak consistency achieved for $SBM(n, 2, \frac{a}{n}, \frac{b}{n})$ at the optimal condition.

- **Semi-dense** regime, $d \sim \log n$: strong consistency achieved for $SBM(n, 2, \frac{a \log n}{n}, \frac{b \log n}{n})$ at the optimal condition.

- **Dense** regime, $d \sim n$: trivial.
Even for SBM, some problems remain open: growing $K$, unbalanced community sizes

SBM is too simplistic for many real networks; for more sophisticated models, not many results available

One of the biggest challenges: getting rid of independent edges, e.g. to allow transitivity

Many promising directions for future work
Thank you