Community detection in stochastic block models via spectral methods

Laurent Massoulié (MSR-Inria Joint Centre, Inria)

based on joint works with:

Dan Tomozei (EPFL), Marc Lelarge (Inria),
Jiaming Xu (UIUC), Charles Bordenave (IMT)
Community Detection

→ Clustering of items based on their observed interactions
→ Typical approach: embedding (e.g., in Euclidean space), then K-means
Application 1: contact recommendation in online social networks

Supporting data: e.g. OSN’s friendship graph

➔ recommend members of user’s implicit community

Variation: NSA’s “co-traveler programme” ➔ spot groups of suspect persons meeting regularly in unusual places
Application 2: item recommendation to users

Supporting data: \{user-item\} matrix
Example: Netflix prize $\rightarrow$ \{user-movie\} ratings matrix

<table>
<thead>
<tr>
<th>User / Movie</th>
<th>$f_1$</th>
<th>$f_2$</th>
<th>…</th>
<th>$f_m$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$u_1$</td>
<td>?</td>
<td>**</td>
<td></td>
<td>***</td>
</tr>
<tr>
<td>$u_2$</td>
<td>***</td>
<td>?</td>
<td></td>
<td>?</td>
</tr>
<tr>
<td>…</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$u_n$</td>
<td>*****</td>
<td>**</td>
<td></td>
<td>**</td>
</tr>
</tbody>
</table>

Use item communities to support recommendation
“users who liked this also liked…”
Application 3: categorizing chemical reactives in biology

Basic data: sets of chemicals jointly involved in specific reaction

More generally

Knowledge graph as generic representation of data
A1 has with B1 interaction of type C1
A2 has with B2 interaction of type C2
...

Outline

• The Stochastic Block Model
  • With labels
  • With general types

• Performance of Spectral Methods
  • “rich signal” case

• The weak signal case: sparse observations
  • Phase transition on detectability
  • Non-backtracking matrix and the Spectral Redemption
The Stochastic Block Model, aka planted partition model [Holland-Laskey-Leinhardt’83]

- \( n \) “nodes” partitioned into \( r \) categories
- Category \( \sigma \): \( \alpha_\sigma \) \( n \) nodes
- Edge between nodes \( u,v \) present with probability \( s \times b_{\sigma(u)\sigma(v)} / n \)
  
  \( s \) : average degree, or signal strength

\[ \rightarrow \text{Observation: adjacency matrix } A = + \text{ noise} \]

**Community Detection**: provide clustering \( \hat{\sigma} \) based on \( A \) that is correlated as much as possible with true planted clustering \( \sigma \), i.e. maximizing

\[
\text{Overlap:} = \text{Max}_\pi \frac{1}{n} \sum_{u=1}^{n} 1_{\pi(\hat{\sigma}(u)) = \sigma(u)} - \text{Max}_{\sigma \in [r]} \alpha_\sigma
\]
The Labeled Stochastic Block Model

- Edges \((u-v)\) labeled by \(L_{uv} \in L\) (finite set)
- Drawn from distribution \(\mu_{\sigma(u)\sigma(v)}\)
- Netflix case: labels 1-5 stars
The SBM with general types [Aldous’81; Lovász’12]

• User type $\sigma(u)$ i.i.d. $\sim P$ in general set (e.g. uniform on [0,1])
• Edge $(u-v)$ present w.p. $b_{\sigma(u)\sigma(v)} s/n$ for “kernel” $b$
  e.g. $b_{x,y} = F(x - y)$

• Edges $(u-v)$ labeled by $L_{uv} \in L$ (finite set)
• Drawn from distribution $\mu_{\sigma(u)\sigma(v)}$

• Technical assumptions: compact type set and continuity of symmetric functions $b$ and $\mu$
Outline

• The Stochastic Block Model
  • With labels
  • With general types

• Performance of Spectral Methods
  • “rich signal” case

• The weak signal case: sparse observations
  • Phase transition on detectability
  • Non-backtracking matrix and the Spectral Redemption
Spectral Clustering

• From Matrix $A$ extract $R$ normed eigenvectors $x_i$ corresponding to $R$ largest eigenvalues $|\lambda_1| \geq \cdots \geq |\lambda_R|$

• Form $R$-dimensional node representatives
  $$y_u = \sqrt{n}(x_i(u))_{i=1\ldots R}$$

• Group nodes $u$ according to proximity of spectral representatives $y_u$

→ Karl Pearson’s PCA:
  “On Lines and Planes of Closest Fit to Systems of Points in Space”, 1901
Illustration for $R=2$

Clustering from SVD

Netflix dataset

SBM with $K=4$
Result for “logarithmic” signal strength $s$

Assume $s=\Omega(\log(n))$ and clusters are distinguishable, i.e.

$$\forall \sigma \neq \sigma' \exists \tau \text{ such that } b_{\sigma\tau} \neq b_{\sigma'\tau}$$

Then spectrum of $A$ consists of

- $R$ eigenvalues $\lambda_i$ of order $\Omega(s)$ ($R \leq K$) and
- $n-R$ eigenvalues $\lambda_i$ of order $O(\sqrt{s})$

Where $R = \operatorname{rank}\{b_{uv}\alpha_v\}_{1 \leq u,v \leq K}$

Node representatives $y_u$ based on top $R$ eigenvectors $x_i$:

- Cluster according to underlying “blocks” except for negligible fraction of nodes
Proof arguments

Control spectral radius of noise matrix
+ perturbation of matrix eigen-elements

\[
A = \text{Block matrix} + \text{random "noise" matrix}
\]

Block matrix
non-zero eigenvalues: \( \Theta(s) \)
Controlling perturbation of eigen-elements of Hermitian matrices

\[ A = \tilde{A} + W \]

Where \( A \) : observed; \( \tilde{A} \) : signal; \( W \) : perturbation

Perturbation of eigenvalues: order \( \lambda_1 \geq \cdots \geq \lambda_n \), \( \bar{\lambda}_1 \geq \cdots \geq \bar{\lambda}_n \)

Then \( |\lambda_i - \bar{\lambda}_i| \leq \rho(W) \), \( i = 1, \ldots, n \) (Weyl)

(a simple consequence of Courant-Fisher’s minimax theorem:  
For Hermitian \( n \times n \) matrix \( H \) :

\[ \lambda_i(H) = \max_{\text{dim}(V)=i} \min_{v \in V, |v|=1} v^* H v \]

\[ \lambda_i(H) = \min_{\text{dim}(V)=n-i+1} \max_{v \in V, |v|=1} v^* H v \]
Controlling perturbation of eigen-elements of Hermitian matrices

\[ A = \tilde{A} + W \]

Where \( A \) : observed; \( \tilde{A} \) : signal; \( W \) : perturbation

Perturbation of eigenvectors: order \( \lambda_1 \geq \cdots \geq \lambda_n \), \( \bar{\lambda}_1 \geq \cdots \geq \bar{\lambda}_n \)

Let \( \Delta := \min_{i,j: \tilde{\lambda}_i \neq \tilde{\lambda}_j} |\tilde{\lambda}_i - \tilde{\lambda}_j| \). If \( \rho(W) < \Delta/2 \)

Then for all \( i \) and for any normed eigenvector \( x_i \leftrightarrow \lambda_i \), there exists \( \bar{x}_i \leftrightarrow \bar{\lambda}_i \) s.t.

\[ x_i \cdot \bar{x}_i \geq \sqrt{1 - \left( \frac{\rho(W)}{\Delta - \rho(W)} \right)^2} \]

( A simple version of [Chandler Davis-Kahane’70] )
Application to SBM

Spectrum of $\bar{A}$ : spectrum of $\{b_{uv}a_v\}_{1 \leq u,v \leq K}$ scaled by $s$, hence $\Delta = \Omega(s)$

Announced result follows if $\rho(W) = o(s)$
Ramanujan graphs
[Lubotzky-Phillips-Sarnak’88]

$s$-regular graphs such that eigenvalues of adjacency matrix verify

$$\lambda := \max_{k:|\lambda_k|< s} |\lambda_k| \leq 2\sqrt{s - 1}$$

The best possible spectral gap [Alon-Boppana’91]

$$\lambda \geq 2\sqrt{s - 1} - O\left(\frac{\sqrt{s}}{\text{diam}(G)}\right)$$
spectral separation properties
“à la Ramanujan”

[Friedman’08] random $s$-regular graph with $s \geq 3$ verifies whp

$$\lambda = 2\sqrt{s - 1} + o(1)$$

[Feige-Ofek’05]: for Erdős-Rényi graph $G(n, s/n)$ and $s = \Omega(\log n)$, then whp $\lambda = O(\sqrt{s})$

Corollary: for SBM, $\rho(A - \bar{A}) = O(\sqrt{s})$ when $s = \Omega(\log n)$
spectral separation properties “à la Ramanujan”

Consequence: in SBM with \( s = \Omega(\log n) \), whp
\[
\rho(A - \bar{A}) = O(\sqrt{s}) \Rightarrow A’s \text{ leading eigen-elements close to those of } \bar{A}
\]

For \( s = \Theta(1) \), \( \rho(A - \bar{A}) \sim C \sqrt{\frac{\log n}{\log \log n}} \)

→ spectral separation is lost

→ standard spectral method needs fixing to work in low signal regime
Discrepancy between SBM with small K and Netflix

Eigenvalue distributions

SBM with $K=4$  

Netflix (subset)

→ motivates consideration of SBM with general types
SBM with general types

• User types \( \sigma(u) \) i.i.d. \( \sim P \) from general set (e.g. uniform on \([0,1]\))

• Edge \((u-v)\) present w.p. \( b_{\sigma(u)\sigma(v)} \) \( s/n \) for “kernel” \( b \)
  e.g. \( b_{x,y} = F(x - y) \)

• Edges \((u-v)\) labeled by \( L_{uv} \in L \) (finite set)

• Drawn from distribution \( \mu_{\sigma(u)\sigma(v)} \)

\( \rightarrow \) Form matrix \( \{A_{ij}W(L_{ij})\} \) from random projections \( W(l) \) of labels
SBM with general types:
Spectral properties for logarithmic $s$

Define kernel $K(x, y) := \sum_l W(l) b_{xy} \mu_{xy}(l)$ and integral operator $Tf(x) := \int K(x, y)f(y)P(dy)$

$\rightarrow$ spectrum of $s^{-1}\{A_{ij}W(L_{ij})\} \approx$ spectrum of $T$

- Eigenvalue convergence: $s^{-1}\lambda_i^{(n)} \rightarrow \lambda_i$
- Eigenvector convergence: $x_i(u) \rightarrow \phi_i(\sigma_u)$

Associated eigenfunction $\varphi_i(\sigma_u)$

Type of node $u$
SBM with general types:
Spectral properties for logarithmic $s$

→ Flexible model
- power-law spectra (convolution operator + Fourier analysis)

$$b_{x,y} = F(x - y)$$

- better matches to
Netflix data
SBM with general types: estimation for logarithmic $S$

For fixed $R$ form $R$-dimensional node representatives $y_u = \sqrt{n} \left\{ \frac{\lambda_k}{\lambda_1} x_k(u) \right\}_{k=1\ldots R}$

- Embedding allows consistent estimation of label distributions.
- Accuracy controlled by $\varepsilon$ and $\varepsilon_R := \sum_{k>R} \lambda_k^2$.

Use empirical distribution of labels $L(i,k)$ for $k$ in $\varepsilon$-neighborhood of $j$. 

$\Rightarrow$ Embedding allows consistent estimation of label distributions.
Outline

• The Stochastic Block Model
  • With labels
  • With general types
• Performance of Spectral Methods
  • “rich signal” case
• The weak signal case: sparse observations
  • Phase transition on detectability
  • Non-backtracking matrix and the Spectral Redemption
Weak signal strength: $s = \Theta(1)$

- Correct classification of all but negligible fraction of nodes impossible (isolated nodes...)

→ Performance of clustering $\hat{\sigma}$ assessed by overlap metric:

$$\text{Overlap} := \max_{\pi} \frac{1}{n} \sum_{u=1}^{n} 1_{\pi(\hat{\sigma}(u)) = \sigma(u)} - \max_{\sigma \in [r]} \alpha_{\sigma}$$
Positively correlated community detection

Symmetric two-communities scenario: \( E(A) = \begin{pmatrix} \frac{a}{n} & \frac{b}{n} \\ \frac{b}{n} & \frac{a}{n} \end{pmatrix} \)

Conjecture ([Decelle-Krzakala-Moore-Zdeborova 2011]):
• For \( \tau : = \frac{(a-b)^2}{2(a+b)} < 1 \), correlation tends to zero for any \( \hat{\sigma} \)

Proven by [Mossel-Neeman-Sly 2012]

• For \( \tau > 1 \), positive correlation can be achieved

Proven by [LM 2013] and [Mossel-Neeman-Sly 2013]


Spectral methods based on non-backtracking matrix can achieve positively correlated detection
Non-backtracking matrix $B$ of graph $G = (V, E)$

Defined on oriented edges $\overrightarrow{uv}$ for $(u, v) \in E$:

$$B_{\overrightarrow{uv}, xy} = \begin{cases} 1 & v = x \land u \neq y \\ 0 & \text{otherwise} \end{cases}$$

$\Rightarrow$ Asymmetric, such that $B^k_{ef} = \text{number of non-backtracking paths on } G \text{ of length } k+1 \text{ starting at } e \text{ and ending at } f$
Non-backtracking matrix $B$ of graph $G = (V, E)$

Defined on oriented edges $u\overrightarrow{v}$ for $(u, v) \in E : B_{u\overrightarrow{v}, xy} = 1_{v=x} 1_{u \neq y}$

Ihara-Bass formula [Ihara 66]

for $A$: adjacency matrix, $D$: diagonal matrix of node degrees,

$$
(1 - u^2)^{n-m} \det(I - uB) = \det(I - uA + u^2(D - I))
$$

Hence $s$-regular graph Ramanujan iff eigenvalues $\lambda_k$ of $B$ verify

$$
\lambda' := \max_{k:|\lambda_k| < \lambda_1} |\lambda_k| \leq \sqrt{\lambda_1}
$$

A definition that extends to non-regular graphs

([Stark-Terras 96]'s graph theory Riemann hypothesis)
Notations and assumptions

Let $\mu_1, \ldots, \mu_r$, $|\mu_1| \geq \ldots \geq |\mu_r|$ be the leading eigenvalues of $E(A)$ and $x_i$ corresponding eigenvectors in $R^n$

Let $y_i$: $y_i(e) = x_i(e_1), y_i \in R^{n(n-1)}$: lift of eigenvector $x_i$ to $R^{n(n-1)}$

In symmetric two-community case, $\mu$ ‘s: eigenvalues of

$$
\begin{bmatrix}
a/2 & b/2 \\
b/2 & a/2
\end{bmatrix}
$$

$\rightarrow \mu_1 = \frac{a+b}{2}, \mu_2 = \frac{a-b}{2}$ (hence $\tau = \frac{\mu_2^2}{\mu_1}$)

Assume constant per-community average degree: $\forall j \in [r], \sum_{i=1}^r \alpha_i b_{ij} \equiv \mu_1$
Notations and assumptions

Call eigenvalue $\mu_i$ “visible” if $|\mu_i| > \sqrt{\mu_1}$

Let $r_0: |\mu_{r_0}| > \sqrt{\mu_1} \geq |\mu_{r_0+1}|$: index of smallest “visible” eigenvalue

Kesten-Stigum condition:

$r_0 > 1$, i.e. there is some $i^* > 1$ such that $\mu_{i^*}$ is “visible”

In symmetric two-community case, Kesten-Stigum condition:

$\mu_2^2 > \mu_1$ or equivalently, $\tau = \frac{\mu_2^2}{\mu_1} > 1$
Main result: spectrum of non-backtracking matrix for stochastic block model

With high probability as $n \to \infty$, eigenvalues $\lambda_i$ of $B$ satisfy

\[
\begin{align*}
    i \leq r_0 & \Rightarrow |\lambda_i - \mu_i| = o(1) \\
    i > r_0 & \Rightarrow |\lambda_i| \leq \sqrt{\mu_1} + o(1)
\end{align*}
\]

For $i \leq r_0$, if $\mu_i$ has multiplicity 1, corresponding eigenvector:

asymptotically parallel to $B^k \left( B^k \right)' y_i$ for $k \sim \log n$
Illustration for 2-community symmetric Stochastic block model

\[ n = 500, \quad a = 7, \quad b = 1, \quad \mu_1 = 4, \quad \mu_2 = 3. \]
Corollary 1: proof of Spectral Redemption Conjecture

Assume $r_0 > 1$, equal community sizes $\alpha_i \equiv \frac{1}{r}$ and $\mu_{i^*}$ has multiplicity 1 for some $i^*: 1 < i^* \leq r_0$

Then positive correlation obtained by following procedure

1) Extract eigenvector $z_{i^*}$ of $B$
2) Form signs $s(e) = \text{sgn}(z_{i^*}(e) - \tau)$
3) To each node $u$ assign $s(u) = s(e)$ for edge $e$ picked uniformly among graph edges with $e_1 = u$
4) Assign community $\hat{\sigma}(u)$ at random according to distribution $\pi(s(u))$
Illustration for Erdős-Rényi graph with $n = 500, \alpha = 4$
Corollary 2: Erdős-Rényi graphs are nearly Ramanujan

For Erdős-Rényi graph $G \left( \frac{\alpha}{n}, n \right)$, spectrum $\{\lambda_i\}$ of associated non-backtracking matrix $B$ satisfies with high probability as $n \to \infty$

$$\max_{k} \colon |\lambda_k| < \lambda_1 \mid \lambda_k \mid \leq \sqrt{\lambda_1} + o(1)$$

$\rightarrow$ An approximate version of

$$\max_{k} \colon |\lambda_k| < \lambda_1 \mid \lambda_k \mid \leq \sqrt{\lambda_1}$$

Property generalizes notion of Ramanujan graph to non-regular case
[Ihara’66, Stark-Terras’96]

Hence: Corollary 2 a « non-regular » version of Friedman’s result for regular graphs:

most random graphs (with given number of edges) approximately Ramanujan according to extended definition
Proof strategy

Key step: for $k \sim \log n$ prove matrix expansion

$$B^k = \sum_{i=1}^{r} \tilde{\lambda}_i^k v_i w'_i + R_k$$

where $\tilde{\lambda}_i \sim \mu_i$ (near-eigenvalue), $v_i \propto B^k (B^k)' y_i$ (near-right eigenvector), $w_i \propto (B^k)' y_i$ (near-left eigenvector), $R_k$ a perturbation

and with high probability

\[ i \neq j \Rightarrow v'_i v_j, v'_i w_j, w'_i w_j = o(1), \] (near-orthonormal system)

\[ v'_i w_i > c > 0 \] (low condition number)

\[ \rho(R_k) = \tilde{O}\left(\sqrt{\mu_1^k}\right) \] (negligible perturbation)

→ Yields the result, after leveraging Bauer-Fike theorem

(a tool to control impact of additive perturbation on eigenvalues of matrix not necessarily symmetric)
Proof elements (perturbations negligible)

To show $\rho(R_k) = \tilde{O}\left(\sqrt{\mu_1^k}\right)$, establish « small norm in complement »:

$$\sup_{|x|=1, x'w_i=0, i=1,\ldots,r} \|B^k x\| = \tilde{O}\left(\sqrt{\mu_1^k}\right)$$

Challenge: directions $w_i$ random and correlated with $B^k$
Remaining mysteries about SBM’s (1)

Conjectured “phase diagram” for more than 2 blocks
(assuming fixed inter-community parameter b)

- Detection “easy” (spectral methods or BP) when $a = \Theta(r)$
  - Above Kesten-Stigum threshold
- Detection infeasible when $a = \Theta(\sqrt{r \log(r)})$
- Detection hard but feasible (how? In polynomial time?) for $r = 4, 5$

Intra-community parameter $a$ vs. Number of communities $r$
Remaining mysteries about SBM’s (2)

Clique detection problem: add a size-$K$ clique to random graph with edge-probability $\frac{1}{2}$

i.e. a 2-block SBM with unbalanced block sizes:

$\rightarrow$ for $K = \Omega \left( \sqrt{n} \right)$ clique easily detectable (e.g. inspection of node degrees)

$\rightarrow$ are there polynomial-time algorithms for smaller yet large $K$? (e.g. $K = \Theta \left( \frac{3}{\sqrt{n}} \right)$)

A notoriously hard problem (“planted clique detection” recently proposed as a new benchmark of algorithmic hardness)
Conclusions and Outlook

- Variations of basic spectral methods still to be invented: interesting mathematics and practical relevance

- Detection in SBM = rich playground for analysis of computational complexity with methods of statistical physics

- Computationally efficient methods for “hard” cases (planted clique, intermediate phase for multiple communities)?

- Non-regular Ramanujan graphs: theory still in its infancy (proper analogue of Alon-Boppana’s theorem missing)
Thanks!