Reconstruction in the Generalized Stochastic Block Model

Marc Lelarge [1] Laurent Massoulié [2] Jiaming Xu [3]

[1]INRIA-ENS

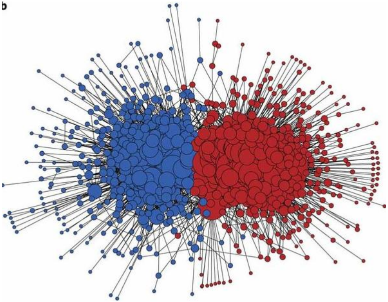
[2]INRIA-Microsoft Research Joint Centre

[3]University of Illinois, Urbana-Champaign

GDR ISIS-Phénix, Nov 25

Motivation

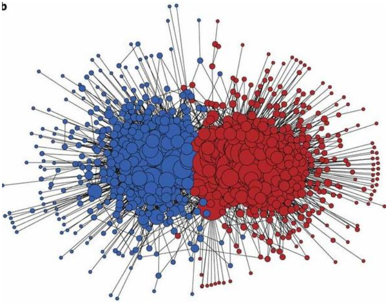
- Community detection in social or biological networks in the sparse regime with a (not too large) average degree.



- Labels to characterize various interaction types, e.g. strong and weak ties in friendship network.

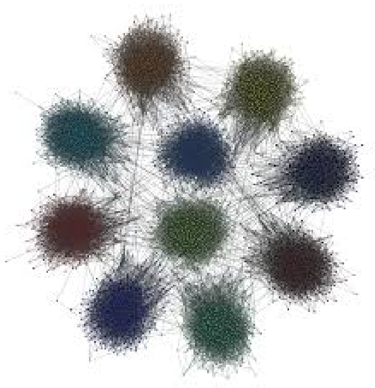
Motivation

- Community detection in social or biological networks in the sparse regime with a (not too large) average degree.



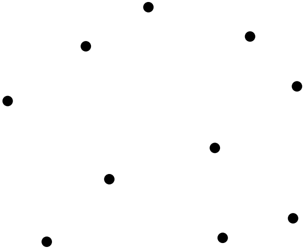
- Labels to characterize various interaction types, e.g. strong and weak ties in friendship network.

A model: the stochastic block model



The sparse stochastic block model

A random graph model on n nodes with two parameters, $a, b \geq 0$.

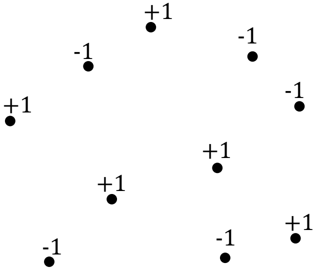


$n = 10$

The sparse stochastic block model

A random graph model on n nodes with two parameters, $a, b \geq 0$.

- Assign each vertex spin $+1$ or -1 uniformly at random.

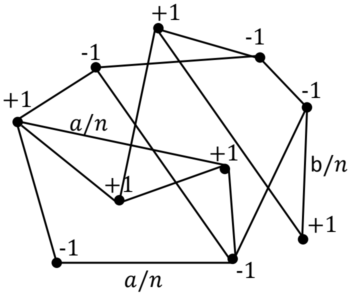


$n = 10$

The sparse stochastic block model

A random graph model on n nodes with two parameters, $a, b \geq 0$.

- Independently for each pair (u, v):
  - if $\sigma_u = \sigma_v$, draw the edge w.p. a/n.
  - if $\sigma_u \neq \sigma_v$, draw the edge w.p. b/n.



$a = 4, b = 2$

Reconstruction problem

- Reconstruct the underlying spin configuration σ based on the observed labeled graph.
- Sparse graph: as $n \to \infty$, the asymptotic degree distribution is Poisson with mean $\frac{a+b}{2}$. With on average, $\frac{a}{2}$ neighbors in the same community and $\frac{b}{2}$ in the other community.
- Isolated nodes render exact reconstruction impossible. Focus on positively correlated reconstruction, i.e., $\hat{\sigma}$ agrees with σ in more than $1/2$ of all its entries.

Reconstruction problem

- Reconstruct the underlying spin configuration σ based on the observed labeled graph.
- Sparse graph: as $n \to \infty$, the asymptotic degree distribution is Poisson with mean $\frac{a+b}{2}$. With on average, $\frac{a}{2}$ neighbors in the same community and $\frac{b}{2}$ in the other community.
- Isolated nodes render exact reconstruction impossible. Focus on positively correlated reconstruction, i.e., $\hat{\sigma}$ agrees with σ in more than $1/2$ of all its entries.

Reconstruction problem

- Reconstruct the underlying spin configuration σ based on the observed labeled graph.
- Sparse graph: as $n \to \infty$, the asymptotic degree distribution is Poisson with mean $\frac{a+b}{2}$. With on average, $\frac{a}{2}$ neighbors in the same community and $\frac{b}{2}$ in the other community.
- Isolated nodes render exact reconstruction impossible. Focus on positively correlated reconstruction, i.e., $\hat{\sigma}$ agrees with σ in more than $1/2$ of all its entries.

Phase transition

Theorem

If $\tau > 1$, then positively correlated reconstruction is possible.
If $\tau < 1$, then positively correlated reconstruction is impossible.

$$\tau = \frac{(a-b)^2}{2(a+b)}.$$

Conjectured by Decelle, Krzakala, Moore, Zdeborova '11 based on statistical physics arguments.

- Non-reconstruction proved by Mossel, Neeman, Sly '12.
- Reconstruction proved by Massoulié '13 and Mossel, Neeman, Sly '13.

Phase transition

Theorem

If $\tau > 1$, then positively correlated reconstruction is possible.
If $\tau < 1$, then positively correlated reconstruction is impossible.

$$\tau = \frac{(a-b)^2}{2(a+b)}.$$

Conjectured by Decelle, Krzakala, Moore, Zdeborova ’11 based on statistical physics arguments.

- Non-reconstruction proved by Mossel, Neeman, Sly ’12.
- Reconstruction proved by Massoulié ’13 and Mossel, Neeman, Sly ’13.

Phase transition

Theorem

If $\tau > 1$, then positively correlated reconstruction is possible.
If $\tau < 1$, then positively correlated reconstruction is impossible.

$$\tau = \frac{(a-b)^2}{2(a+b)}.$$

Conjectured by Decelle, Krzakala, Moore, Zdeborova '11 based on statistical physics arguments.

- Non-reconstruction proved by Mossel, Neeman, Sly '12.
- Reconstruction proved by Massoulié '13 and Mossel, Neeman, Sly '13.

Efficiency of Spectral Algorithms

Boppana '87, Condon, Karp '01, Carson, Impagliazzo '01, McSherry '01, Kannan, Vempala, Vetta '04...

Proposition

Suppose that for sufficiently large c and c',

$$\frac{(a-b)^2}{a+b} \geq c + c' \frac{a}{a+b} \ln\left(\frac{a+b}{2}\right),$$

then 'trimming+spectral+greedy improvement' outputs a positively correlated partition a.a.s.

Coja-Oghlan '10

What if $a, b \to \infty$?

Proposition

Assume $a \geq \ln^5 n$ and $(a-b)^2 > 164(a+b)$, then the clustering problem is solvable by the simple spectral method.

Lelarge, Massoulié, Xu '13

Lower bound (valid for any a)

Proposition

For $\alpha < 1/2$, define $\delta = \frac{1}{2} - \inf_{\hat{\sigma}} \mathbb{P}(d(\sigma, \hat{\sigma}) > \alpha)$. Then $\delta > 0$ implies

$$\frac{(a-b)^2}{2(a+b)} > 1 - H(\alpha),$$

with $H(x) = -x \log x - (1-x)\log(1-x)$.

Proof: Fano's inequality.

What if $a, b \to \infty$?

Proposition

Assume $a \geq \ln^5 n$ and $(a-b)^2 > 164(a+b)$, then the clustering problem is solvable by the simple spectral method.

Lelarge, Massoulié, Xu '13

Lower bound (valid for any a)

Proposition

For $\alpha < 1/2$, define $\delta = \frac{1}{2} - \inf_{\hat\sigma} \mathbb{P}(d(\sigma, \hat\sigma) > \alpha)$. Then $\delta > 0$ implies

$$\frac{(a-b)^2}{2(a+b)} > 1 - H(\alpha),$$

with $H(x) = -x \log x - (1-x)\log(1-x)$.

Proof: Fano's inequality.

What if $a, b \to \infty$?

Proposition

Assume $a \geq \ln^5 n$ and $(a-b)^2 > 164(a+b)$, then the clustering problem is solvable by the simple spectral method.

Lelarge, Massoulié, Xu '13
Lower bound (valid for any a)

Proposition

For $\alpha < 1/2$, define $\delta = \frac{1}{2} - \inf_{\hat{\sigma}} \mathbb{P}(d(\sigma, \hat{\sigma}) > \alpha)$. Then $\delta > 0$ implies

$$\frac{(a-b)^2}{2(a+b)} > 1 - H(\alpha),$$

with $H(x) = -x \log x - (1-x)\log(1-x)$.

Proof: Fano's inequality.

Spectral analysis

Assume that $a > \ln^5 n$, and $a - b \approx \sqrt{a+b}$ so that $a \sim b$.

$$A = \frac{a+b}{2}\frac{\mathbf{1}}{\sqrt{n}}\frac{\mathbf{1}^T}{\sqrt{n}} + \frac{a-b}{2}\frac{\sigma}{\sqrt{n}}\frac{\sigma^T}{\sqrt{n}} + A - \mathbb{E}[A]$$

$\frac{a+b}{2}$ is the mean degree and degrees in the graph are very concentrated in the regime $a > \ln^5 n$. We can construct

$$A - \frac{a+b}{2n}J = \frac{a-b}{2}\frac{\sigma}{\sqrt{n}}\frac{\sigma^T}{\sqrt{n}} + A - \mathbb{E}[A]$$

Spectral analysis

Assume that $a > \ln^5 n$, and $a - b \approx \sqrt{a+b}$ so that $a \sim b$.

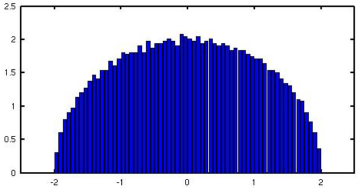
$$A = \frac{a+b}{2}\frac{\mathbf{1}}{\sqrt{n}}\frac{\mathbf{1}^T}{\sqrt{n}} + \frac{a-b}{2}\frac{\sigma}{\sqrt{n}}\frac{\sigma^T}{\sqrt{n}} + A - \mathbb{E}[A]$$

$\frac{a+b}{2}$ is the mean degree and degrees in the graph are very concentrated in the regime $a > \ln^5 n$. We can construct

$$A - \frac{a+b}{2n}J = \frac{a-b}{2}\frac{\sigma}{\sqrt{n}}\frac{\sigma^T}{\sqrt{n}} + A - \mathbb{E}[A]$$

Spectrum of the noise matrix

The matrix $A - \mathbb{E}[A]$ is a symmetric random matrix with independent centered entries having variance $\sim \frac{a}{n}$.
To have convergence to the Wigner semicircle law, we need to normalize the variance to $\frac{1}{n}$.



$$ESD\left(\frac{A - \mathbb{E}[A]}{\sqrt{a}}\right) \to \mu_{sc}(x) = \begin{cases} \frac{1}{2\pi}\sqrt{4 - x^2}, & \text{if } |x| \leq 2; \\ 0, & \text{otherwise.} \end{cases}$$

Naive spectral analysis

To sum up, we can construct:

$$\begin{aligned} M &= \frac{1}{\sqrt{a}}\left(A - \frac{a+b}{2n}J\right) \\ &= \theta\frac{\sigma}{\sqrt{n}}\frac{\sigma}{\sqrt{n}}^T + \frac{A - \mathbb{E}[A]}{\sqrt{a}}, \end{aligned}$$

with $\theta = \frac{a-b}{2\sqrt{a}}$.
We should be able to detect signal as soon as

$$\theta > 2 \Leftrightarrow \frac{(a-b)^2}{2(a+b)} > 4$$

Naive spectral analysis

To sum up, we can construct:

$$\begin{aligned} M &= \frac{1}{\sqrt{a}}\left(A - \frac{a+b}{2n}J\right) \\ &= \theta\frac{\sigma}{\sqrt{n}}\frac{\sigma^T}{\sqrt{n}} + \frac{A - \mathbb{E}[A]}{\sqrt{a}}, \end{aligned}$$

with $\theta = \frac{a-b}{2\sqrt{a}}$.
We should be able to detect signal as soon as

$$\theta > 2 \Leftrightarrow \frac{(a-b)^2}{2(a+b)} > 4$$

Some hope

A lower bound on the spectral radius of $M = \theta \frac{\sigma}{\sqrt{n}} \frac{\sigma^T}{\sqrt{n}} + W$:

$$\lambda_1(M) = \sup_{\|x\|=1} \|Mx\| \geq \|M \frac{\sigma}{\sqrt{n}}\|$$

But

$$\begin{aligned} \|M \frac{\sigma}{\sqrt{n}}\|^2 &= \theta^2 + \|W \frac{\sigma}{\sqrt{n}}\|^2 + 2\langle W, \frac{\sigma}{\sqrt{n}} \rangle \\ &\approx \theta^2 + \frac{1}{n} \sum_{i,j} W_{ij}^2 \\ &\approx \theta^2 + 1. \end{aligned}$$

As a result, we get

$$\lambda_1(M) > 2 \Leftrightarrow \theta > 1 \Leftrightarrow \frac{(a-b)^2}{2(a+b)} > 1.$$

Some hope

A lower bound on the spectral radius of $M = \theta \frac{\sigma}{\sqrt{n}} \frac{\sigma^T}{\sqrt{n}} + W$:

$$\lambda_1(M) = \sup_{\|x\|=1} \|Mx\| \geq \|M \frac{\sigma}{\sqrt{n}}\|$$

But

$$\begin{aligned} \|M \frac{\sigma}{\sqrt{n}}\|^2 &= \theta^2 + \|W \frac{\sigma}{\sqrt{n}}\|^2 + 2\langle W, \frac{\sigma}{\sqrt{n}} \rangle \\ &\approx \theta^2 + \frac{1}{n} \sum_{i,j} W_{ij}^2 \\ &\approx \theta^2 + 1. \end{aligned}$$

As a result, we get

$$\lambda_1(M) > 2 \Leftrightarrow \theta > 1 \Leftrightarrow \frac{(a-b)^2}{2(a+b)} > 1.$$

Some hope

A lower bound on the spectral radius of $M = \theta \frac{\sigma}{\sqrt{n}} \frac{\sigma^T}{\sqrt{n}} + W$:

$$\lambda_1(M) = \sup_{\|x\|=1} \|Mx\| \geq \|M \frac{\sigma}{\sqrt{n}}\|$$

But

$$\begin{aligned} \|M \frac{\sigma}{\sqrt{n}}\|^2 &= \theta^2 + \|W \frac{\sigma}{\sqrt{n}}\|^2 + 2\langle W, \frac{\sigma}{\sqrt{n}} \rangle \\ &\approx \theta^2 + \frac{1}{n} \sum_{i,j} W_{ij}^2 \\ &\approx \theta^2 + 1. \end{aligned}$$

As a result, we get

$$\lambda_1(M) > 2 \Leftrightarrow \theta > 1 \Leftrightarrow \frac{(a-b)^2}{2(a+b)} > 1.$$

Baik, Ben Arous, Péché phase transition

Rank one perturbation of a Wigner matrix:

$$\lambda_1(\theta\sigma\sigma^T + W) \overset{a.s}{\to} \begin{cases} \theta + \frac{1}{\theta} & \text{if } \theta > 1, \\ 2 & \text{otherwise.} \end{cases}$$

Let $\tilde{\sigma}$ be the eigenvector associated with $\lambda_1(\theta uu^T + W)$, then

$$|\langle \tilde{\sigma}, \sigma \rangle|^2 \overset{a.s}{\to} \begin{cases} 1 - \frac{1}{\theta^2} & \text{if } \theta > 1, \\ 0 & \text{otherwise.} \end{cases}$$

Baik, Ben Arous, Péché '05

Rigorous proof of the phase transition for $a \geq \ln^5 n$

Proposition

Assume $a \geq \ln^5 n$. Then the clustering problem is solvable by the simple spectral method, provided

$$\frac{(a-b)^2}{2(a+b)} > 1.$$

Lelarge '13

Proof: control the spectral norm thanks to Vu '05 and adapt the argument in Benaych-Georges, Nadakuditi '11.

In agreement with Nadakuditi, Newman '12.

Rigorous proof of the phase transition for $a \geq \ln^5 n$

Proposition

Assume $a \geq \ln^5 n$. Then the clustering problem is solvable by the simple spectral method, provided

$$\frac{(a-b)^2}{2(a+b)} > 1.$$

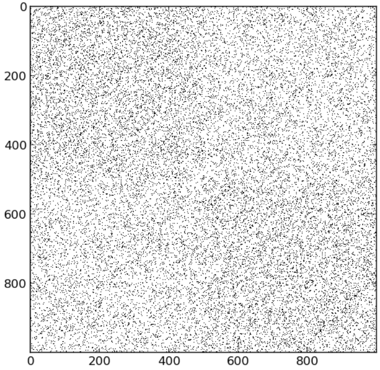
Lelarge '13

Proof: control the spectral norm thanks to Vu '05 and adapt the argument in Benaych-Georges, Nadakuditi '11.

In agreement with Nadakuditi, Newman '12.

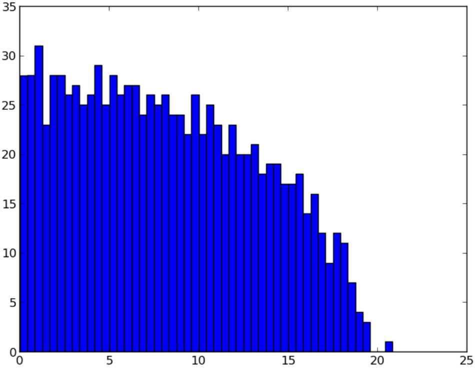
Spectral Algorithm

Original adjacency matrix with 2 communities. $a = 120$, $b = 92$, $\theta = \frac{a-b}{\sqrt{2(a+b)}} = 1.46385...$



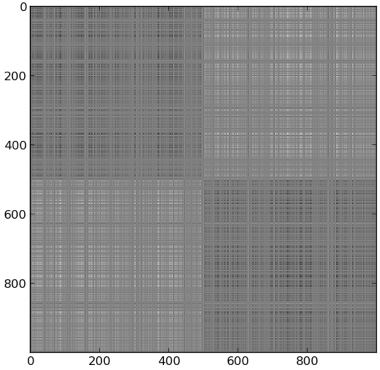
Spectral Algorithm

Spectrum of the original adjacency matrix. $a = 120$, $b = 92$,
$\theta = \frac{a-b}{\sqrt{2(a+b)}} = 1.46385...$



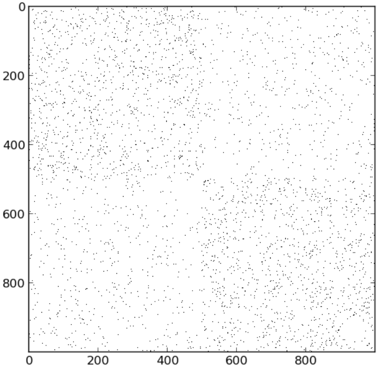
Spectral Algorithm

Rank-1 approximation of the adjacency matrix. $a = 120$, $b = 92$, $\theta = \frac{a-b}{\sqrt{2(a+b)}} = 1.46385...$



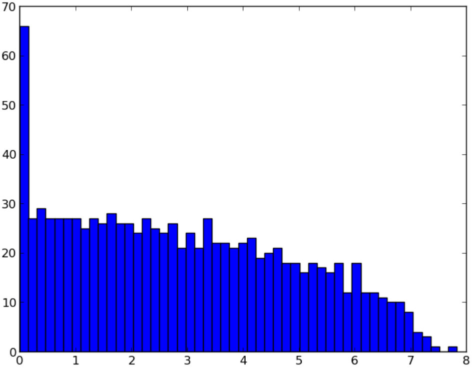
Spectral Algorithm: low degree

Original adjacency matrix with 2 communities. $a = 20$, $b = 9$, $\theta = \frac{a-b}{\sqrt{2(a+b)}} = 1.44437...$



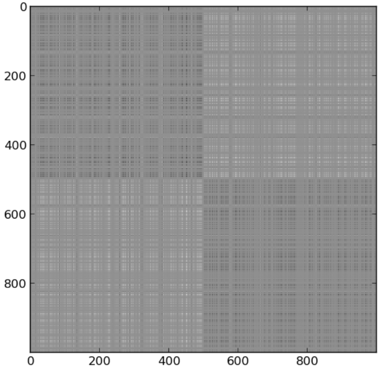
Spectral Algorithm: low degree

Spectrum of the original adjacency matrix (after trimming).
$a = 20$, $b = 9$, $\theta = \frac{a-b}{\sqrt{2(a+b)}} = 1.44437...$



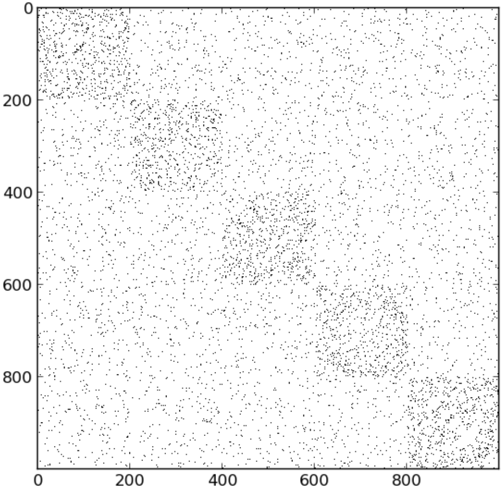
Spectral Algorithm: low degree

Rank-1 approximation of the adjacency matrix. $a = 20$, $b = 9$, $\theta = \frac{a-b}{\sqrt{2(a+b)}} = 1.44437...$



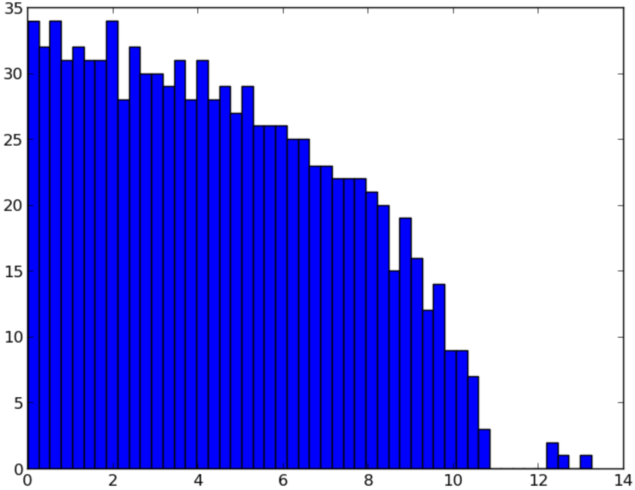
Spectral Algorithm: more communities

Original adjacency matrix with 5 communities.



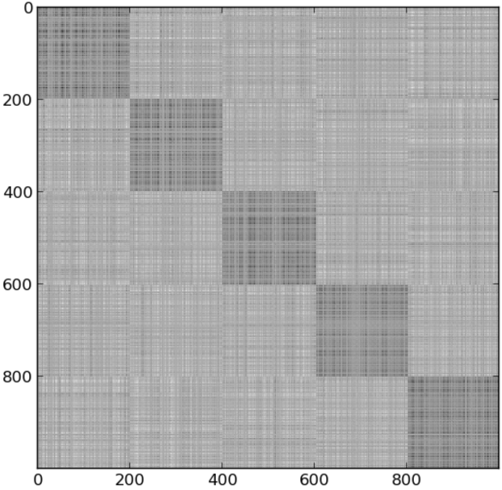
Spectral Algorithm: more communities

Spectrum of the original adjacency matrix.



Spectral Algorithm: more communities

Rank-4 approximation of the adjacency matrix.



Extension: r symmetric communities

Proposition

Assume $a \geq \ln^5 n$ and $r \geq 2$ symmetric communities. Then the clustering problem is solvable by the simple spectral method, provided

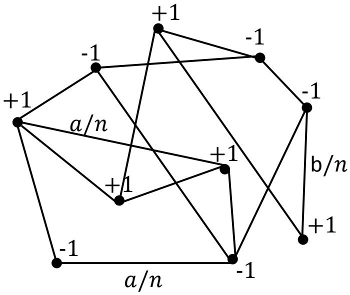
$$\frac{(a-b)^2}{r(a+(r-1)b)} > 1.$$

Lelarge '13

The sparse labeled stochastic block model

A random graph model on n nodes with two parameters, $a, b \geq 0$ and two discrete prob. distributions, μ, ν.

- Independently for each pair (u, v):
  - if $\sigma_u = \sigma_v$, draw the edge w.p. a/n.
  - if $\sigma_u \neq \sigma_v$, draw the edge w.p. b/n.

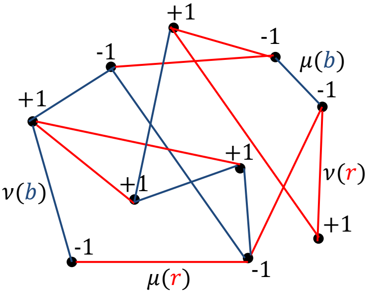


$a = 4, b = 2$

The sparse labeled stochastic block model

A random graph model on n nodes with two parameters, $a, b \geq 0$ and two discrete prob. distributions, μ, ν.

- Independently for each edge (u, v):
  - if $\sigma_u = \sigma_v$, label the edge with $L_{uv} \sim \mu$.
  - if $\sigma_u \neq \sigma_v$, label the edge with $L_{uv} \sim \nu$.



$$\mu(r) = 0.6, \mu(b) = 0.4$$
$$\nu(r) = 0.4, \nu(b) = 0.6$$

How to use labels?

- Maximum log likelihood estimation:

$$\max_{\sigma} \sum_{(u,v)\in E(G)} \sigma_u \sigma_v \log \frac{a\mu(L_{uv})}{b\nu(L_{uv})}$$

$$\text{s.t.} \sum_{u} \sigma_u = 0,\ \sigma_u \in \{-1, 1\}$$

- Minimum bisection with edge weights $w(\ell) = \log \frac{a\mu(\ell)}{b\nu(\ell)}$.
- Minimum bisection is NP-hard. Let's try some statistical physics!

How to use labels?

- Maximum log likelihood estimation:

$$\max_{\sigma} \sum_{(u,v)\in E(G)} \sigma_u \sigma_v \log \frac{a\mu(L_{uv})}{b\nu(L_{uv})}$$

$$\text{s.t.} \sum_u \sigma_u = 0, \ \sigma_u \in \{-1, 1\}$$

- Minimum bisection with edge weights $w(\ell) = \log \frac{a\mu(\ell)}{b\nu(\ell)}$.
- Minimum bisection is NP-hard. Let's try some statistical physics!

How to use labels?

- Maximum log likelihood estimation:

$$\max_{\sigma} \sum_{(u,v)\in E(G)} \sigma_u \sigma_v \log \frac{a\mu(L_{uv})}{b\nu(L_{uv})}$$

$$\text{s.t.} \sum_u \sigma_u = 0,\ \sigma_u \in \{-1, 1\}$$

- Minimum bisection with edge weights $w(\ell) = \log \frac{a\mu(\ell)}{b\nu(\ell)}$.
- Minimum bisection is NP-hard. Let's try some statistical physics!

Phase transition (with labels)

Conjecture

If $\tau_L > 1$, then positively correlated reconstruction is possible.
If $\tau_L < 1$, then positively correlated reconstruction is impossible.

$$\tau_L = \frac{1}{2} \sum_{\ell \in \mathcal{L}} \frac{(a\mu(\ell) - b\nu(\ell))^2}{a\mu(\ell) + b\nu(\ell)}.$$

Heimlicher, Lelarge, Massoulié '12

- Generalize the result for (standard) stochastic block model and $\tau_L \geq \tau$.
- τ_L comes from the local stability analysis of a fixed point of Belief Propagation.

Phase transition (with labels)

Conjecture

If $\tau_L > 1$, then positively correlated reconstruction is possible.
If $\tau_L < 1$, then positively correlated reconstruction is impossible.

$$\tau_L = \frac{1}{2} \sum_{\ell \in \mathcal{L}} \frac{(a\mu(\ell) - b\nu(\ell))^2}{a\mu(\ell) + b\nu(\ell)}.$$

Heimlicher, Lelarge, Massoulié '12

- Generalize the result for (standard) stochastic block model and $\tau_L \geq \tau$.
- τ_L comes from the local stability analysis of a fixed point of Belief Propagation.

Phase transition (with labels)

Conjecture

If $\tau_L > 1$, then positively correlated reconstruction is possible.
If $\tau_L < 1$, then positively correlated reconstruction is impossible.

$$\tau_L = \frac{1}{2} \sum_{\ell \in \mathcal{L}} \frac{(a\mu(\ell) - b\nu(\ell))^2}{a\mu(\ell) + b\nu(\ell)}.$$

Heimlicher, Lelarge, Massoulié '12

- Generalize the result for (standard) stochastic block model and $\tau_L \geq \tau$.
- τ_L comes from the local stability analysis of a fixed point of Belief Propagation.

Non-reconstruction

Theorem

If $\tau_L < 1$, then for any fixed vertices u and v, conditional on the spin of v, the spin of u is asymptotically uniformly distributed.

Lelarge, Massoulié, Xu, 13

- It further implies that it is impossible to reconstruct a positively correlated partition.
- Proof: similar to Mossel, Neeman, Sly '12, uses local tree argument, conditional independence property and the Ising spin model on labeled tree.

Non-reconstruction

Theorem

If $\tau_L < 1$, then for any fixed vertices u and v, conditional on the spin of v, the spin of u is asymptotically uniformly distributed.

Lelarge, Massoulié, Xu, 13

- It further implies that it is impossible to reconstruct a positively correlated partition.
- Proof: similar to Mossel, Neeman, Sly '12, uses local tree argument, conditional independence property and the Ising spin model on labeled tree.

Spectral method with labels

- A is the weighted adjacency matrix: $A_{uv} = 1((u,v) \in E(G))w(L_{uv})$.
- Spectral method as a relaxation of the minimum bisection:

$$\max \sum_{(u,v)} \sigma_u A_{uv} \sigma_v$$
$$\text{s.t. } \sum_i \sigma_i = 0, \; \|\sigma\|_2 = 1.$$

- Perturbed low-rank matrix A

$$\mathbb{E}[A|\sigma] = \frac{a\overline{\mu} + b\overline{\nu}}{2n}\mathbf{1}\mathbf{1}^\top + \frac{a\overline{\mu} - b\overline{\nu}}{2n}\sigma\sigma^\top.$$

- Curse from vertices of high degrees $\Omega(\frac{\log n}{\log \log n})$.

Spectral method with labels

- A is the weighted adjacency matrix: $A_{uv} = 1((u,v) \in E(G))w(L_{uv})$.
- Spectral method as a relaxation of the minimum bisection:

$$\max \sum_{(u,v)} \sigma_u A_{uv} \sigma_v$$

$$\text{s.t. } \sum_i \sigma_i = 0, \ \|\sigma\|_2 = 1.$$

- Perturbed low-rank matrix A

$$\mathbb{E}[A|\sigma] = \frac{a\overline{\mu} + b\overline{\nu}}{2n}\mathbf{1}\mathbf{1}^\top + \frac{a\overline{\mu} - b\overline{\nu}}{2n}\sigma\sigma^\top.$$

- Curse from vertices of high degrees $\Omega(\frac{\log n}{\log \log n})$.

Spectral method with labels

- A is the weighted adjacency matrix: $A_{uv} = 1((u,v) \in E(G))w(L_{uv})$.
- Spectral method as a relaxation of the minimum bisection:

$$\max \sum_{(u,v)} \sigma_u A_{uv} \sigma_v$$
$$\text{s.t. } \sum_i \sigma_i = 0, \ \|\sigma\|_2 = 1.$$

- Perturbed low-rank matrix A

$$\mathbb{E}[A|\sigma] = \frac{a\overline{\mu} + b\overline{\nu}}{2n}\mathbf{1}\mathbf{1}^\top + \frac{a\overline{\mu} - b\overline{\nu}}{2n}\sigma\sigma^\top.$$

- Curse from vertices of high degrees $\Omega(\frac{\log n}{\log\log n})$.

Spectral method with labels

- A is the weighted adjacency matrix: $A_{uv} = 1((u,v) \in E(G))w(L_{uv})$.
- Spectral method as a relaxation of the minimum bisection:

$$\max \sum_{(u,v)} \sigma_u A_{uv} \sigma_v$$

$$\text{s.t. } \sum_i \sigma_i = 0, \ \|\sigma\|_2 = 1.$$

- Perturbed low-rank matrix A

$$\mathbb{E}[A|\sigma] = \frac{a\overline{\mu} + b\overline{\nu}}{2n}\mathbf{1}\mathbf{1}^\top + \frac{a\overline{\mu} - b\overline{\nu}}{2n}\sigma\sigma^\top.$$

- Curse from vertices of high degrees $\Omega(\frac{\log n}{\log\log n})$.

Spectral Algorithm: rigorous results

Remove nodes of degree greater than $\frac{3}{2}\frac{a+b}{2}$.
'Optimal' weight function:

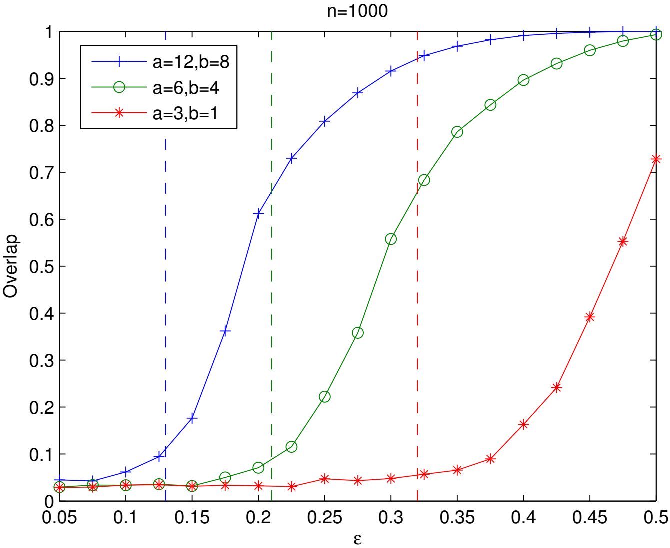
$$w(\ell) = \frac{a\mu(\ell) - b\nu(\ell)}{a\mu(\ell) + b\nu(\ell)}$$

Theorem

If $\tau_L > C\sqrt{a+b}$, then w.h.p. the spectral algorithm gives a positively correlated partition.

- Proof: spectrum of truncated ER random graph, extension of Feige Ofek '05

Spectral Algorithm: empirical results



Rigorous results for $a \geq \ln^5 n$

Proposition

Assume $a \geq \ln^5 n$ and $r \geq 2$ symmetric communities. Then the clustering problem is solvable by the simple spectral method, provided

$$\frac{1}{r}\sum_{\ell}\frac{(a\mu(\ell)-b\nu(\ell))^2}{a\mu(\ell)+(r-1)b\nu(\ell)} > 1$$

Lelarge '13

Proved using more Random Matrix Theory.

Extensions

- Some results for models with latent space allowing to relax the low-rank assumption and overlapping communities. If the signal strength is at least $\log n$, then consistent estimation of the edge label distribution is possible.
- For the planted clique problem, clique of size larger than $\sqrt{n}$ are detectable by a simple spectral algorithm. Deshpande Montanari '13 message passing algorithm works for sizes $\sqrt{n/e} = 0.60653...\sqrt{n}$. However cliques of size $2\log_2 n$ can be found by exhaustive search...

Extensions

- Some results for models with latent space allowing to relax the low-rank assumption and overlapping communities. If the signal strength is at least $\log n$, then consistent estimation of the edge label distribution is possible.
- For the planted clique problem, clique of size larger than $\sqrt{n}$ are detectable by a simple spectral algorithm. Deshpande Montanari '13 message passing algorithm works for sizes $\sqrt{n/e} = 0.60653...\sqrt{n}$.
  However cliques of size $2\log_2 n$ can be found by exhaustive search...

Summary

- For the labeled (not too sparse) stochastic block model, there is a phase transition between an impossible regime and an easy regime where the simple spectral algorithm is successful.
- How well does the spectral algorithm performs in term of 'overlap'? What if parameters are unknown?
- Is there a computational threshold for $r \geq 5$?

THANK YOU!

Summary

- For the labeled (not too sparse) stochastic block model, there is a phase transition between an impossible regime and an easy regime where the simple spectral algorithm is successful.
- How well does the spectral algorithm performs in term of 'overlap'? What if parameters are unknown?
- Is there a computational threshold for $r \geq 5$?

THANK YOU!

Summary

- For the labeled (not too sparse) stochastic block model, there is a phase transition between an impossible regime and an easy regime where the simple spectral algorithm is successful.
- How well does the spectral algorithm performs in term of 'overlap'? What if parameters are unknown?
- Is there a computational threshold for $r \geq 5$?

THANK YOU!

Summary

- For the labeled (not too sparse) stochastic block model, there is a phase transition between an impossible regime and an easy regime where the simple spectral algorithm is successful.
- How well does the spectral algorithm performs in term of 'overlap'? What if parameters are unknown?
- Is there a computational threshold for $r \geq 5$?

THANK YOU!