Reconstruction in the Sparse Labeled Stochastic Block Model

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

[1]INRIA-ENS

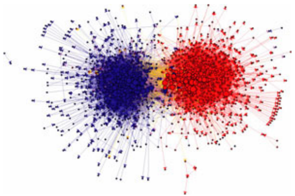
[2]INRIA-Microsoft Research Joint Centre

[3]University of Illinois, Urbana-Champaign

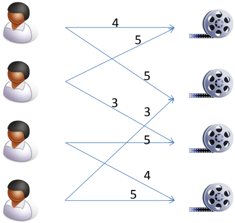
September 10, 2013

Motivation: community detection

Identify underlying communities based on the pairwise interactions represented by graph

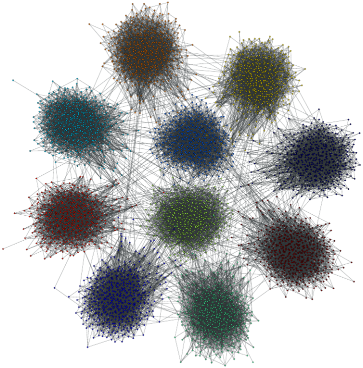


Network of political webblogs [Adamic-Glance '05]



Recommendation system

A popular model: stochastic block model



Graph generated from stochastic block model with $n = 5000$, $r = 10$, $p = 0.999$, $q = 0.001$. Ref. https://projects.skewed.de/graph-tool.

Model we use: sparse labeled stochastic block model

Two important aspects received little attention:

1. Sparse graph: Limited amount of interaction
2. Interaction can be of multiple types: ratings can be 1 to 5

We use sparse labeled stochastic block model:

1. Nodes have bounded average degree: edge probabilities $p, q = O(1/n)$.
2. Edges carry label: label represents interaction type

Model we use: sparse labeled stochastic block model

Two important aspects received little attention:

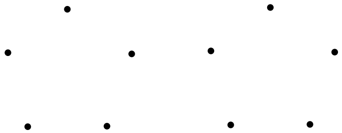
1. Sparse graph: Limited amount of interaction
2. Interaction can be of multiple types: ratings can be 1 to 5

We use sparse labeled stochastic block model:

1. Nodes have bounded average degree: edge probabilities $p, q = O(1/n)$.
2. Edges carry label: label represents interaction type

Sparse labeled stochastic block model: two symmetric communities

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

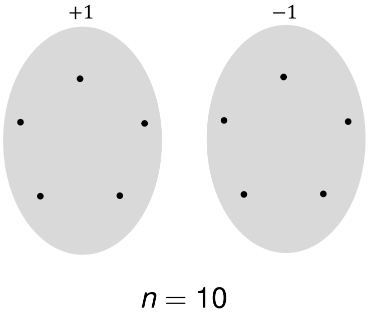


$n = 10$

Sparse labeled stochastic block model: two symmetric communities

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

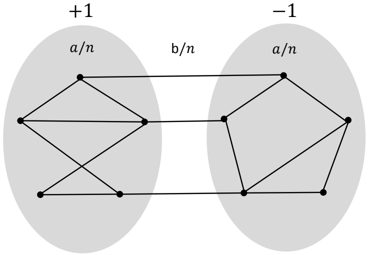
Assign each node to community $+1$ or -1 uniformly at random.



Sparse labeled stochastic block model: two symmetric communities

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

Independently for each pair of nodes:
Draw an edge w.p. a/n if they are in the same community;
w.p. b/n otherwise.

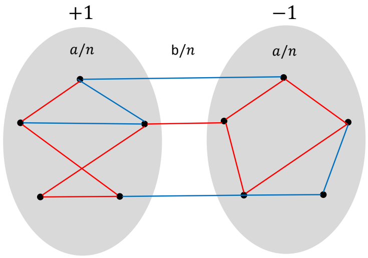


$a = 4, b = 2$

Sparse labeled stochastic block model: two symmetric communities

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

Independently for each edge: Label the edge w.d. μ if the two endpoints are in the same community; w.d. ν otherwise.



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

Correlated reconstruction of community assignment

Isolated nodes render the exact reconstruction impossible.

Focus on correlated reconstruction, i.e., agrees with the true community assignment in more than half of all nodes.

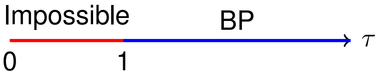
Correlated reconstruction of community assignment

Isolated nodes render the exact reconstruction impossible.

Focus on correlated reconstruction, i.e., agrees with the true community assignment in more than half of all nodes.

Phase transition for correlated reconstruction

[Heimlicher et al. '12] conjectures:

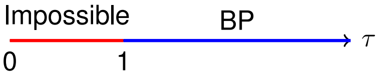


$$\tau = \frac{a+b}{2} \sum_{\ell \in \mathcal{L}} \frac{a\mu(\ell) + b\nu(\ell)}{a+b} \left(\frac{a\mu(\ell) - b\nu(\ell)}{a\mu(\ell) + b\nu(\ell)} \right)^2$$

It generalizes the conjecture in stochastic block model [Decelle et al. '11] [Mossel et al. '12].

Phase transition for correlated reconstruction

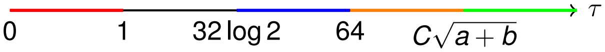
[Heimlicher et al. '12] conjectures:



$$\tau = \frac{a+b}{2} \sum_{\ell \in \mathcal{L}} \frac{a\mu(\ell) + b\nu(\ell)}{a+b} \left(\frac{a\mu(\ell) - b\nu(\ell)}{a\mu(\ell) + b\nu(\ell)} \right)^2$$

It generalizes the conjecture in stochastic block model [Decelle et al. '11] [Mossel et al. '12].

Our contribution



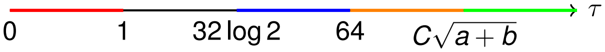
Impossible, MinBisection, Spectral*, Spectral

(A) Prove the impossibility result: correlated reconstruction is impossible if $\tau < 1$.

(B) Prove the achievability result: correlated reconstruction is possible if $\tau > 32 \log 2$ by MinBisection algorithm.

(C) Propose a polynomial-time Spectral algorithm and show it achieves correlated reconstruction if τ is large enough.

Our contribution



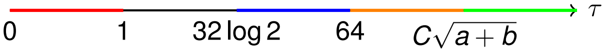
Impossible, MinBisection, Spectral*, Spectral

(A) Prove the impossibility result: correlated reconstruction is impossible if $\tau < 1$.

(B) Prove the achievability result: correlated reconstruction is possible if $\tau > 32 \log 2$ by MinBisection algorithm.

(C) Propose a polynomial-time Spectral algorithm and show it achieves correlated reconstruction if τ is large enough.

Our contribution



Impossible, MinBisection, Spectral*, Spectral

(A) Prove the impossibility result: correlated reconstruction is impossible if $\tau < 1$.

(B) Prove the achievability result: correlated reconstruction is possible if $\tau > 32 \log 2$ by MinBisection algorithm.

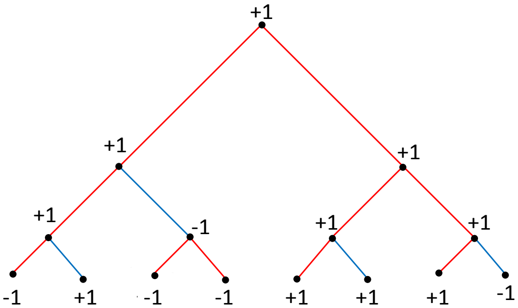
(C) Propose a polynomial-time Spectral algorithm and show it achieves correlated reconstruction if τ is large enough.

(A) Impossibility for correlated reconstruction if $\tau < 1$

First step of proof: correlated reconstruction $\Longrightarrow$ long range correlation: The community memberships of any two nodes randomly chosen are asymptotically correlated conditional on the labeled graph [Mossel et al. '12].

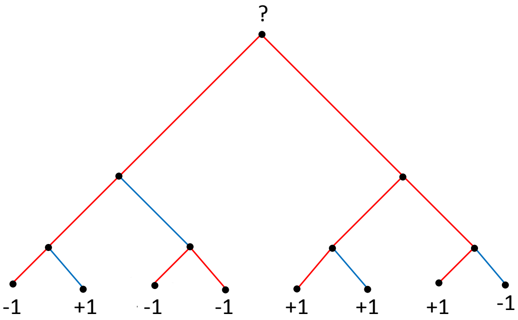
(A) Impossibility for correlated reconstruction if $\tau < 1$

Second step of proof: The small local neighborhood of a node "looks like" a labeled Poisson tree [Mossel et al. '12].



(A) Impossibility for correlated reconstruction if $\tau < 1$

Third step: The community memberships of leaf nodes is uninformative on the community membership of root asymptotically [Heimlicher et al. '12].



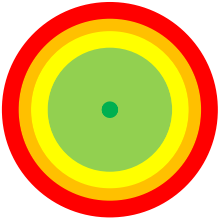
(B) Achievability of correlated reconstruction if $\tau > 32 \log 2$

Consider minimum bisection algorithm: Find two equal-sized parts with the minimum sum of weights of edges between parts with weight function $w(\ell) = \frac{a\mu(\ell) - b\nu(\ell)}{a\mu(\ell) + b\nu(\ell)}$.

Theorem

solutions of minimum bisection are correlated with the true community assignment if $\tau > 32 \log 2$ and $a \sum_\ell \mu(\ell) w^2(\ell), b \sum_\ell \nu(\ell) w^2(\ell) > 8 \ln 2$.

Proof uses the Chernoff bound and the weight function $w(\ell)$ is chosen optimally.



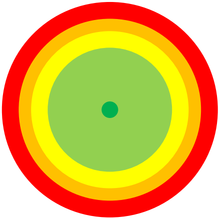
(B) Achievability of correlated reconstruction if $\tau > 32\log 2$

Consider minimum bisection algorithm: Find two equal-sized parts with the minimum sum of weights of edges between parts with weight function $w(\ell) = \frac{a\mu(\ell) - b\nu(\ell)}{a\mu(\ell) + b\nu(\ell)}$.

Theorem

solutions of minimum bisection are correlated with the true community assignment if $\tau > 32\log 2$ and $a\sum_\ell \mu(\ell) w^2(\ell), b\sum_\ell \nu(\ell) w^2(\ell) > 8\ln 2$.

Proof uses the Chernoff bound and the weight function $w(\ell)$ is chosen optimally.



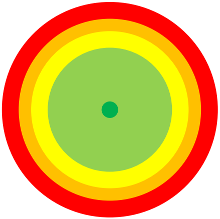
(B) Achievability of correlated reconstruction if $\tau > 32 \log 2$

Consider minimum bisection algorithm: Find two equal-sized parts with the minimum sum of weights of edges between parts with weight function $w(\ell) = \frac{a\mu(\ell) - b\nu(\ell)}{a\mu(\ell) + b\nu(\ell)}$.

Theorem

solutions of minimum bisection are correlated with the true community assignment if $\tau > 32 \log 2$ and $a \sum_\ell \mu(\ell) w^2(\ell), b \sum_\ell \nu(\ell) w^2(\ell) > 8 \ln 2$.

Proof uses the Chernoff bound and the weight function $w(\ell)$ is chosen optimally.



(C) Polynomial-time reconstruction for τ large enough

A polynomial-time algorithm exploiting the spectrum of weighted adjacency matrix W with weight function $w(\ell) = \frac{a\mu(\ell) - b\nu(\ell)}{a\mu(\ell) + b\nu(\ell)}$

- The expectation of W conditional on the true community assignment σ satisfies:

$$\mathbb{E}[W|\sigma] = \frac{a-b}{2n}\mathbf{1}\mathbf{1}^\top + \frac{\tau}{n}\sigma\sigma^\top.$$

- If $W \approx \mathbb{E}[W|\sigma]$, then the top left singular vector of $W - \frac{a-b}{2n}\mathbf{1}\mathbf{1}^\top$ approximates σ.
- Due to nodes with high degrees $\Omega(\frac{\log n}{\log \log n})$, W is not concentrated around its conditional mean.

(C) Polynomial-time reconstruction for τ large enough

A polynomial-time algorithm exploiting the spectrum of weighted adjacency matrix W with weight function $w(\ell) = \frac{a\mu(\ell) - b\nu(\ell)}{a\mu(\ell) + b\nu(\ell)}$

- The expectation of W conditional on the true community assignment σ satisfies:

$$\mathbb{E}[W|\sigma] = \frac{a-b}{2n}\mathbf{1}\mathbf{1}^\top + \frac{\tau}{n}\sigma\sigma^\top.$$

- If $W \approx \mathbb{E}[W|\sigma]$, then the top left singular vector of $W - \frac{a-b}{2n}\mathbf{1}\mathbf{1}^\top$ approximates σ.
- Due to nodes with high degrees $\Omega(\frac{\log n}{\log \log n})$, W is not concentrated around its conditional mean.

(C) Polynomial-time reconstruction for τ large enough

A polynomial-time algorithm exploiting the spectrum of weighted adjacency matrix W with weight function $w(\ell) = \frac{a\mu(\ell) - b\nu(\ell)}{a\mu(\ell) + b\nu(\ell)}$

- The expectation of W conditional on the true community assignment σ satisfies:

$$\mathbb{E}[W|\sigma] = \frac{a-b}{2n}\mathbf{1}\mathbf{1}^\top + \frac{\tau}{n}\sigma\sigma^\top.$$

- If $W \approx \mathbb{E}[W|\sigma]$, then the top left singular vector of $W - \frac{a-b}{2n}\mathbf{1}\mathbf{1}^\top$ approximates σ.
- Due to nodes with high degrees $\Omega(\frac{\log n}{\log \log n})$, W is not concentrated around its conditional mean.

(C) Polynomial-time reconstruction for τ large enough

Spectral-Reconstruction algorithm: Input a, b, μ, ν, W; Output community assignment.

1. Remove nodes with degree greater than $\frac{3}{2}(a+b)$ and assign random community to these nodes.
2. Define W' by setting to zero the rows and columns of W corresponding to the nodes removed.
3. Let $\hat{x}$ denote the left singular vector associated with the largest singular values of $W' - \frac{a-b}{n}\mathbf{1}\mathbf{1}^\top$. Output $\mathrm{sign}(\hat{x})$ as the community assignment for remaining nodes.

(C) Polynomial-time reconstruction for τ large enough

Spectral-Reconstruction algorithm: Input a, b, μ, ν, W; Output community assignment.

1. Remove nodes with degree greater than $\frac{3}{2}(a+b)$ and assign random community to these nodes.
2. Define W' by setting to zero the rows and columns of W corresponding to the nodes removed.
3. Let $\hat{x}$ denote the left singular vector associated with the largest singular values of $W' - \frac{a-b}{n}\mathbf{1}\mathbf{1}^\top$. Output sign($\hat{x}$) as the community assignment for remaining nodes.

(C) Polynomial-time reconstruction for τ large enough

Spectral-Reconstruction algorithm: Input a, b, μ, ν, W; Output community assignment.

1. Remove nodes with degree greater than $\frac{3}{2}(a+b)$ and assign random community to these nodes.
2. Define W' by setting to zero the rows and columns of W corresponding to the nodes removed.
3. Let $\hat{x}$ denote the left singular vector associated with the largest singular values of $W' - \frac{a-b}{n}\mathbf{1}\mathbf{1}^\top$. Output $\mathrm{sign}(\hat{x})$ as the community assignment for remaining nodes.

(C) Polynomial-time reconstruction for τ large enough

Theorem

Spectral-Reconstruction algorithm outputs a community assignment correlated with the true one if $\tau > C\sqrt{a+b}$.

Proof:

- The spectral norm of $W' - \mathbb{E}[W|\sigma]$ is $O(\sqrt{a+b})$. Ref. [Feige-Ofek '05], [Coja-Oghlan '10].
- The L_2 distance between $\sigma/\sqrt{n}$ and $\hat{x}$ is bounded by $O(\|W' - \mathbb{E}[W|\sigma]\|/\tau)$.

(C) Polynomial-time reconstruction for τ large enough

Theorem

Spectral-Reconstruction algorithm outputs a community assignment correlated with the true one if $\tau > C\sqrt{a+b}$.

Proof:

- The spectral norm of $W' - \mathbb{E}[W|\sigma]$ is $O(\sqrt{a+b})$. Ref. [Feige-Ofek '05], [Coja-Oghlan '10].
- The L_2 distance between $\sigma/\sqrt{n}$ and $\hat{x}$ is bounded by $O(\|W' - \mathbb{E}[W|\sigma]\|/\tau)$.

(C) Polynomial-time reconstruction for τ large enough

Theorem

Spectral-Reconstruction algorithm outputs a community assignment correlated with the true one if $\tau > C\sqrt{a+b}$.

Proof:

- The spectral norm of $W' - \mathbb{E}[W|\sigma]$ is $O(\sqrt{a+b})$. Ref. [Feige-Ofek '05], [Coja-Oghlan '10].
- The L_2 distance between $\sigma/\sqrt{n}$ and $\hat{x}$ is bounded by $O(\|W' - \mathbb{E}[W|\sigma]\|/\tau)$.

(C) Polynomial-time reconstruction for $\tau > 64$

Better performance guarantee if $\max\{a, b\} = \Omega(\log^6 n)$.

Theorem

Spectral-Reconstruction algorithm outputs a community assignment correlated with the true one if $\tau > 64$ and $\max\{a, b\} = \Omega(\log^6 n)$.

Proof: $\|W - \mathbb{E}[W|\sigma]\| \leq 2\sqrt{2\tau}$. Ref. [Vu '07].

(C) Polynomial-time reconstruction for $\tau > 64$

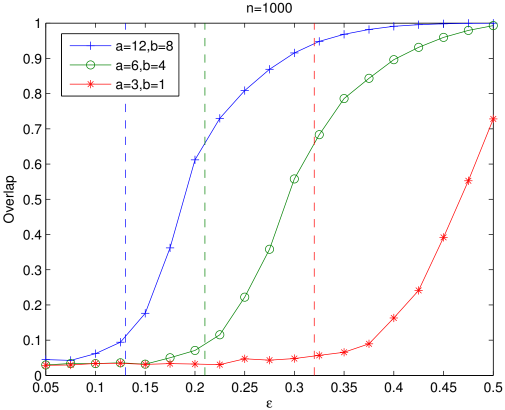
Better performance guarantee if $\max\{a, b\} = \Omega(\log^6 n)$.

Theorem

Spectral-Reconstruction algorithm outputs a community assignment correlated with the true one if $\tau > 64$ and $\max\{a, b\} = \Omega(\log^6 n)$.

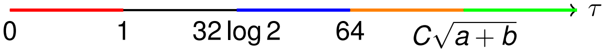
Proof: $\|W - \mathbb{E}[W|\sigma]\| \leq 2\sqrt{2\tau}$. Ref. [Vu '07].

Simulation of Spectral-Reconstruction algorithm



Two labels: r and b with $\mu(\mathrm{r}) = 0.5 + \epsilon$ and $\nu(\mathrm{r}) = 0.5 - \epsilon$. The threshold $\tau = 1$ is depicted as a vertical dash line.

Summary of our results



Impossible, Open,MinBisection, Spectral*, Spectral

Future work

- Apply sparse labeled stochastic block model into real data
- Convergence of belief propagation
- Reconstruction algorithms approaching the reconstruction threshold

Future work

- Apply sparse labeled stochastic block model into real data
- Convergence of belief propagation
- Reconstruction algorithms approaching the reconstruction threshold

Future work

- Apply sparse labeled stochastic block model into real data
- Convergence of belief propagation
- Reconstruction algorithms approaching the reconstruction threshold

Future work

- Apply sparse labeled stochastic block model into real data
- Convergence of belief propagation
- Reconstruction algorithms approaching the reconstruction threshold