Spectral thresholds in the bipartite stochastic block model

Laura Florescu and Will Perkins

NYU and U of Birmingham

September 27, 2016

Stochastic Block Model

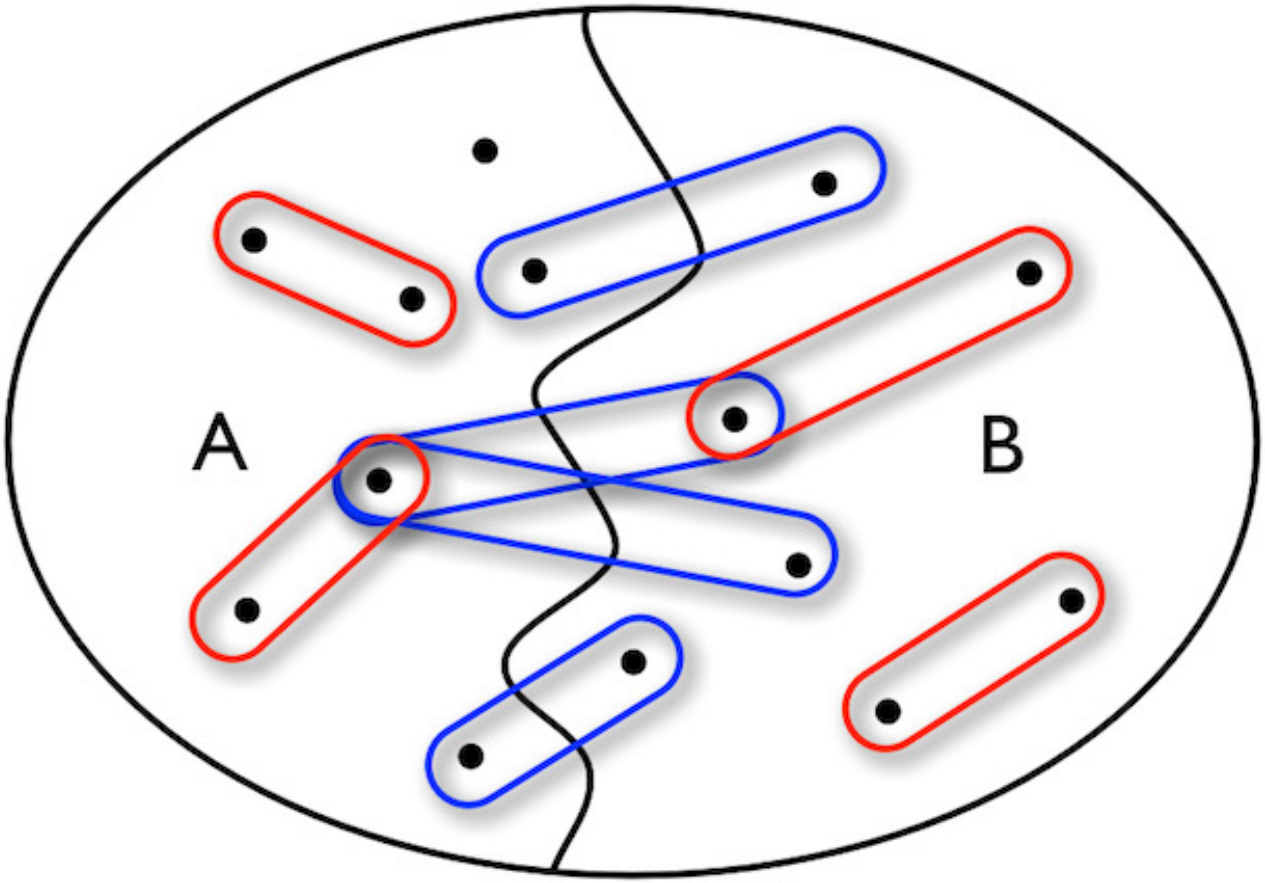
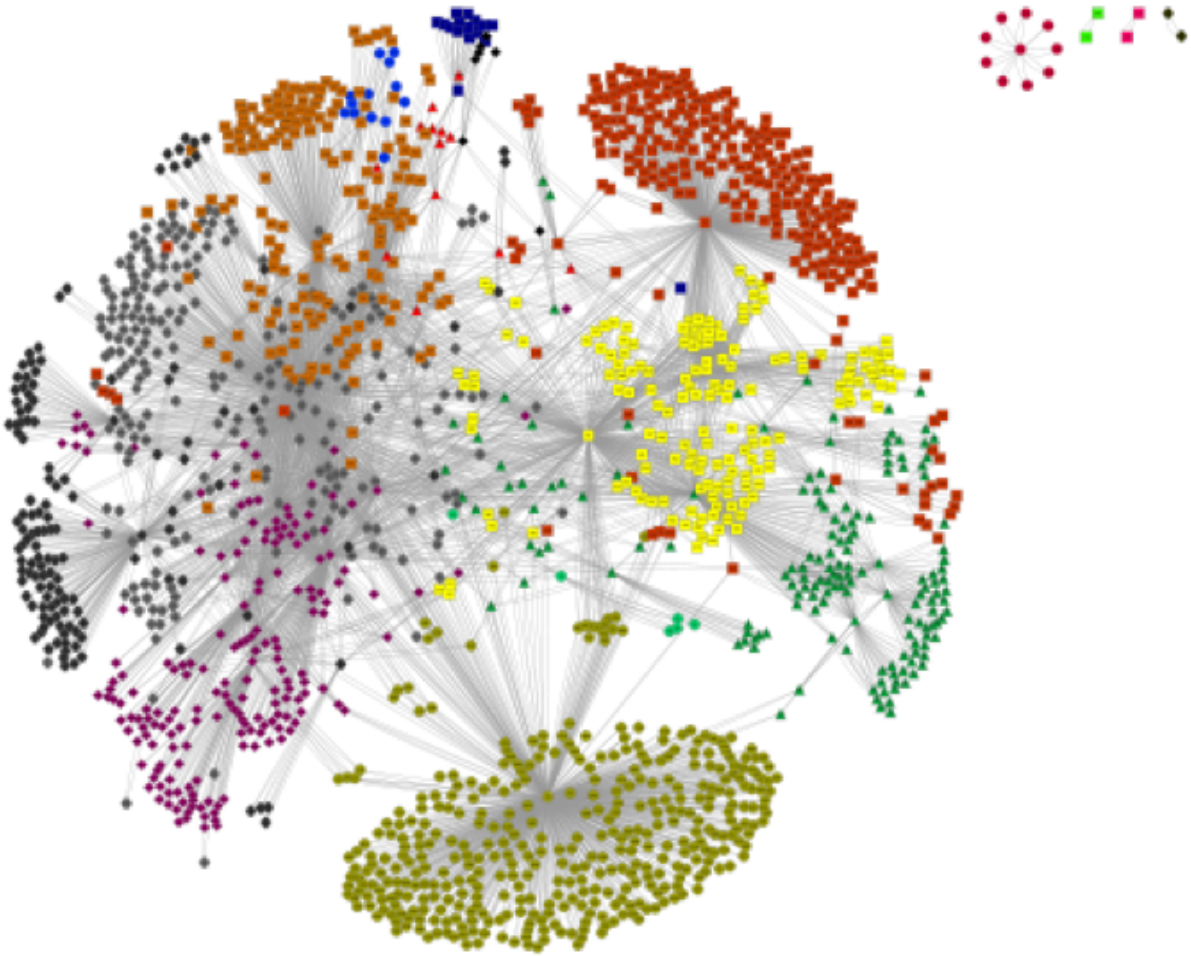


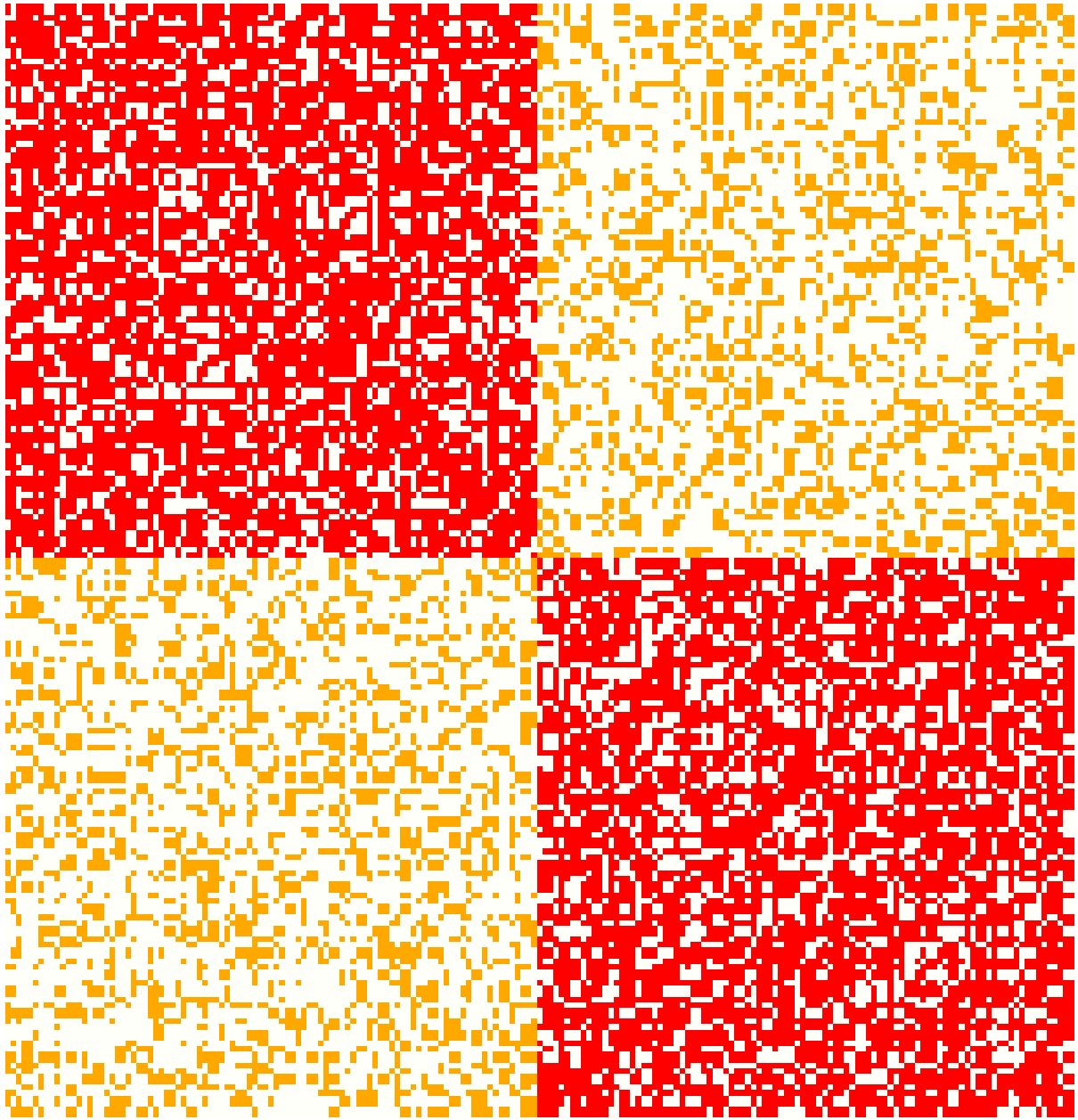
Figure: Red edges added with $\mathbb{P} = p$ and blue edges with $\mathbb{P} = q$.

Community detection

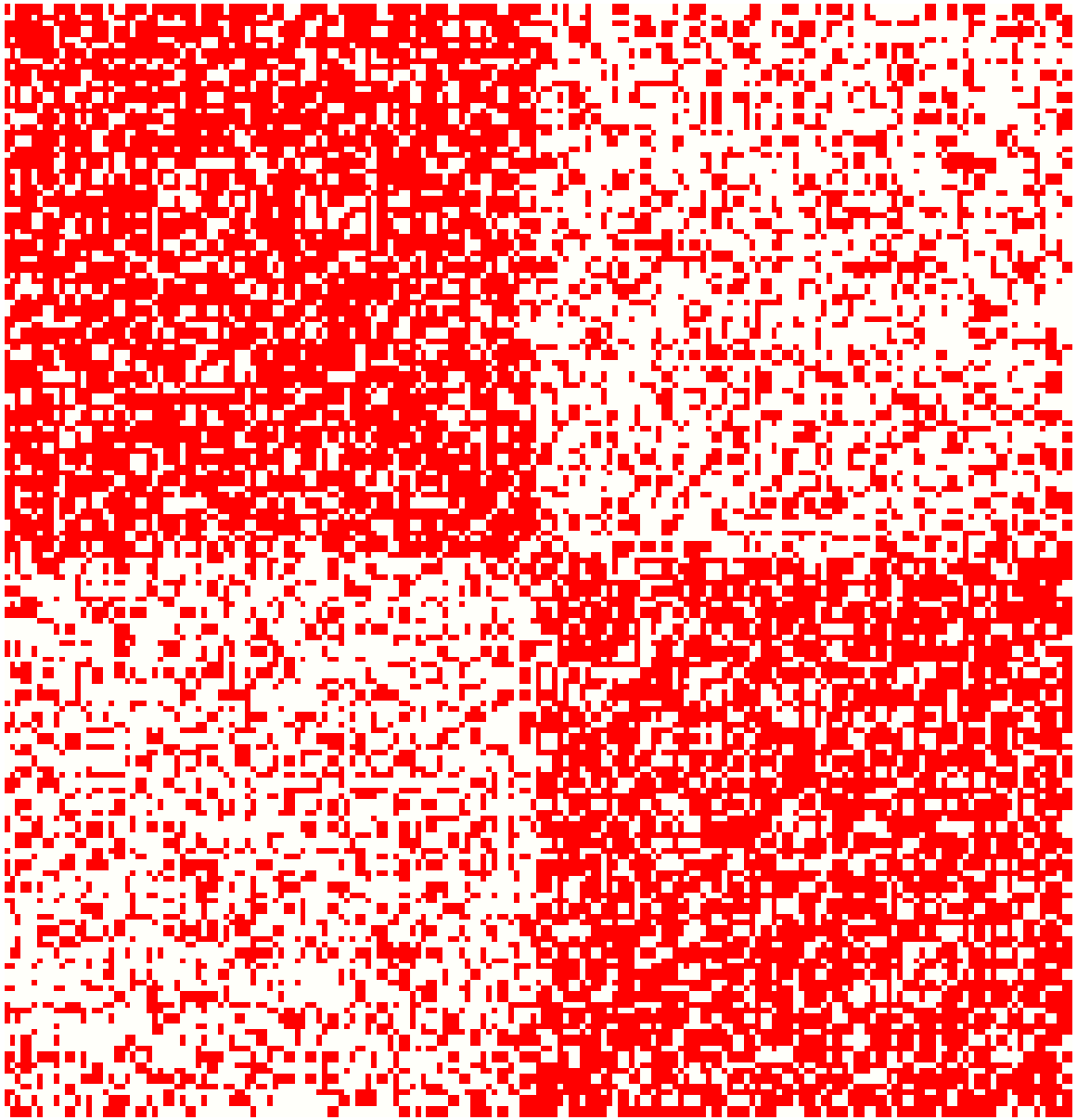


Goal: Detect communities in networks.

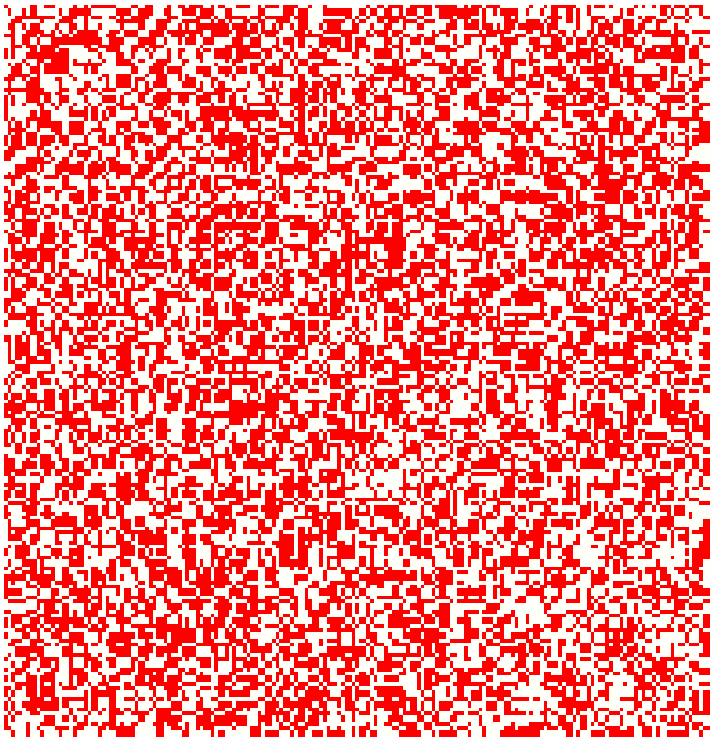
Stochastic Block Model



Entries are not colored



Nor ordered



Problem: Detect/estimate the partition

Stochastic Block Model

First introduced by Holland, Laskey, Leinhardt in 1983.
Motivation: discover communities in large networks.

Theorem (Boppana, Dyer/Frieze, Snijders/Nowicki, Condon/Karp, McSherry, Bickel/Chen, etc)

There are efficient algorithms for exactly recovering the true colors, provided that $|p - q|$ is large enough as $n \to \infty$.

Bipartite Stochastic Block Model

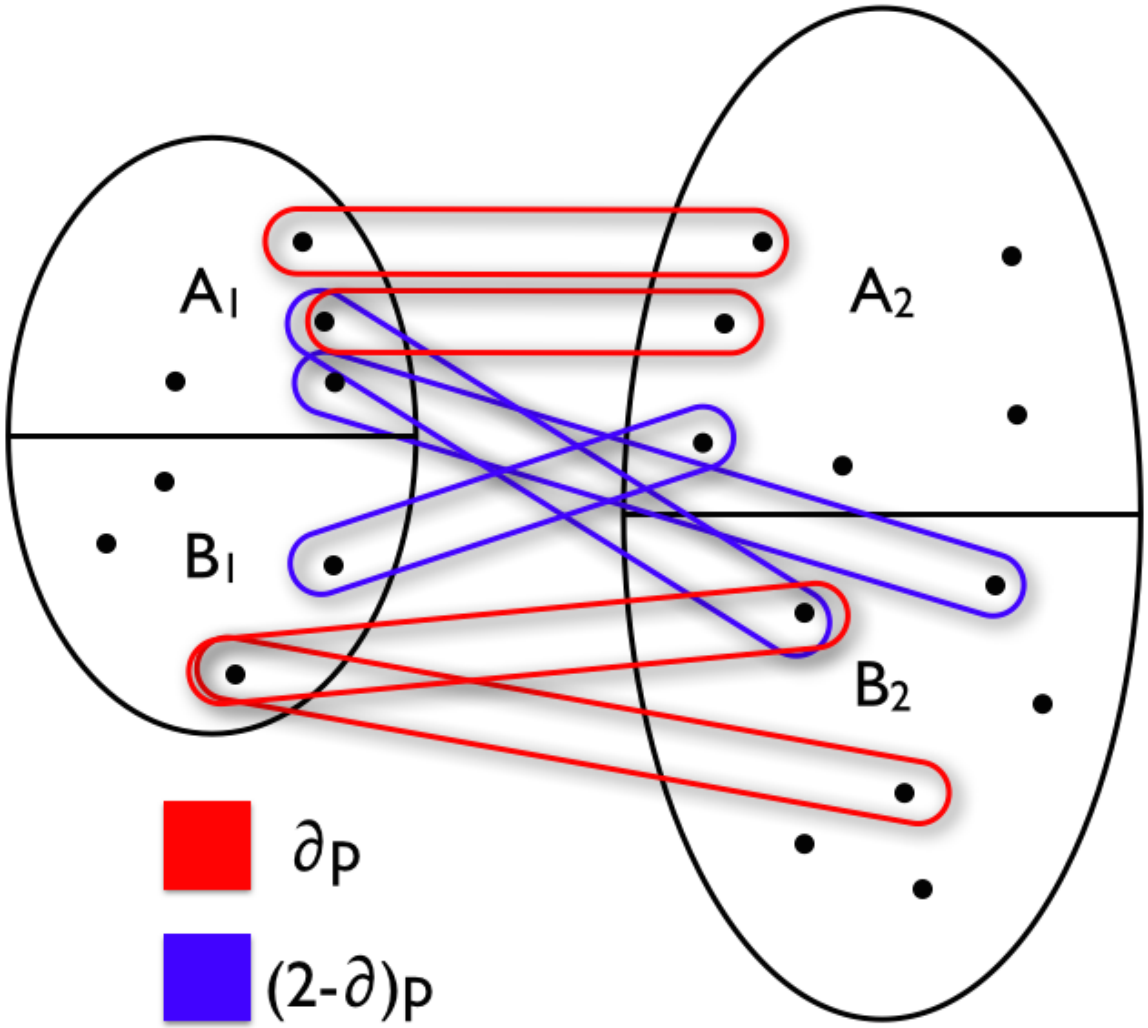


Figure: Bipartite stochastic model on V_1 and V_2. Red edges added with $\mathbb{P} = \delta p(n_1, n_2)$ and blue edges with $\mathbb{P} = (2 - \delta)p(n_1, n_2)$.

SBM

- Goal: get the planted assignment σ (on V_1 for bipartite stochastic model)

SBM

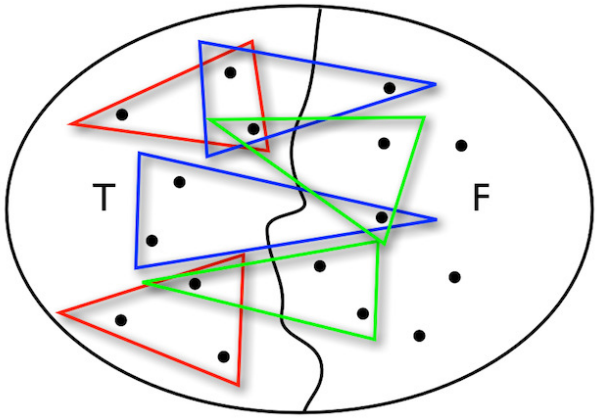
- Goal: get the planted assignment σ (on V_1 for bipartite stochastic model)
- **Detection**: compute v that agrees with σ on $1/2 + \epsilon$ fraction of vertices

SBM

- Goal: get the planted assignment σ (on V_1 for bipartite stochastic model)
- **Detection**: compute v that agrees with σ on $1/2 + \epsilon$ fraction of vertices
- **Recovery**: compute v that agrees with σ on $1 - o(1)$ fraction of vertices

Background

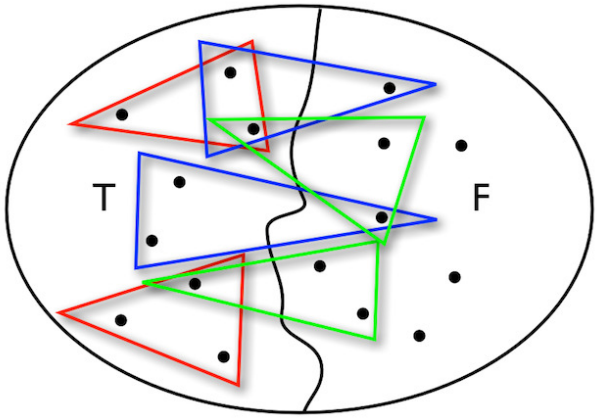
Intermediate step in recovering solutions in planted problems [Feldman, Perkins, Vempala '14].



- planted constraint satisfaction problems (CSP)

Background

Intermediate step in recovering solutions in planted problems [Feldman, Perkins, Vempala '14].



- planted constraint satisfaction problems (CSP)
- Reducing planted problems on n variables will give vertex sets of size $n_1 = n$, $n_2 = n^{k-1}$. $(n_2 \approx n^2)$

Unified Planted k-CSP model

Unified Planted k-CSP model

Definition (Feldman-Perkins-Vempala'14)

Given a planting distribution $Q : \{\pm 1\}^k \mapsto [0,1]$,

Unified Planted k-CSP model

Definition (Feldman-Perkins-Vempala'14)

Given a planting distribution $Q : \{\pm 1\}^k \mapsto [0,1]$,
and an assignment $\sigma \in \{\pm 1\}^n$,

Unified Planted k-CSP model

Definition (Feldman-Perkins-Vempala'14)

Given a planting distribution $Q : \{\pm 1\}^k \mapsto [0,1]$,
and an assignment $\sigma \in \{\pm 1\}^n$,
define the random constraint satisfaction problem $F_{Q,\sigma}(n,m)$

Unified Planted k-CSP model

Definition (Feldman-Perkins-Vempala'14)

Given a planting distribution $Q : \{\pm 1\}^k \mapsto [0,1]$,
and an assignment $\sigma \in \{\pm 1\}^n$,
define the random constraint satisfaction problem $F_{Q,\sigma}(n,m)$
by drawing m k-clauses from $\mathcal{C}_k$ (the set of all k-tuples) independently according to

$$Q_\sigma(C) = \frac{Q(\sigma(C))}{\sum_{C' \in \mathcal{C}_k} Q(\sigma(C'))}$$

Unified Planted k-CSP model

Definition (Feldman-Perkins-Vempala'14)

Given a planting distribution $Q : \{\pm 1\}^k \mapsto [0,1]$,
and an assignment $\sigma \in \{\pm 1\}^n$,
define the random constraint satisfaction problem $F_{Q,\sigma}(n,m)$
by drawing m k-clauses from $\mathcal{C}_k$ (the set of all k-tuples) independently according to

$$Q_\sigma(C) = \frac{Q(\sigma(C))}{\sum_{C' \in \mathcal{C}_k} Q(\sigma(C'))}$$

where $\sigma(C)$ is the vector of values that σ assigns to the k-tuple of literals comprising C.

Planted random k-SAT and Goldreich PRG

Planted random k-SAT: Form a truth assignment ϕ of literals, then select each clause independently from the k-tuples of literals where at least one literal is set to 1 by ϕ.

Planted random k-SAT and Goldreich PRG

Planted random k-SAT: Form a truth assignment ϕ of literals, then select each clause independently from the k-tuples of literals where at least one literal is set to 1 by ϕ.

Goldreich PRG: also add a 0/1, depending on a predicate evaluated on literals. (cryptography)

Planted random k-SAT and Goldreich PRG

Planted random k-SAT: Form a truth assignment ϕ of literals, then select each clause independently from the k-tuples of literals where at least one literal is set to 1 by ϕ.

Goldreich PRG: also add a 0/1, depending on a predicate evaluated on literals. (cryptography)

Feldman, Perkins, Vempala '14 gave a reduction of above and others to the BSBM.

Information theory threshold

When $p = a/n$ and $q = b/n$

Theorem (Mossel, Neeman, Sly, 2012)

There is a test to distinguish the partition that succeeds with high probability if and only if $a + b > 2$ and

$$(a - b)^2 > 2(a + b).$$

Proves conjecture of [Decelle, Krzakala, Moore, Zdeborova '13].

Computational threshold

- Dyer, Frieze 1989 $p = na > q = nb$ fixed
- Condon, Karp 2001 $a - b \gg n^{1/2}$
- McSherry 2001 $a - b \gg \sqrt{b \log n}$
- Coja-Oghlan 2010 $a - b \gg \sqrt{b}$

Computational threshold

- Dyer, Frieze 1989 $p = na > q = nb$ fixed
- Condon, Karp 2001 $a - b \gg n^{1/2}$
- McSherry 2001 $a - b \gg \sqrt{b \log n}$
- Coja-Oghlan 2010 $a - b \gg \sqrt{b}$
- Massoulié 2013 and Mossel, Neeman, Sly 2013 - detection possible and efficient

$$(a - b)^2 > 2(a + b).$$

Computational threshold

- Dyer, Frieze 1989 $p = na > q = nb$ fixed
- Condon, Karp 2001 $a - b \gg n^{1/2}$
- McSherry 2001 $a - b \gg \sqrt{b \log n}$
- Coja-Oghlan 2010 $a - b \gg \sqrt{b}$
- Massoulié 2013 and Mossel, Neeman, Sly 2013 - detection possible and efficient

$$(a - b)^2 > 2(a + b).$$

Ingenious spectral methods

Previous work

- MNS Idea: nbhd of vertex in $G(n, a/n, b/n)$ looks like a random labelled tree, where each child gives birth to Pois(a) vertices of same type, Pois(b) vertices of different type

Previous work

- MNS Idea: nbhd of vertex in $G(n, a/n, b/n)$ looks like a random labelled tree, where each child gives birth to Pois(a) vertices of same type, Pois(b) vertices of different type

- show that conditioned on the labels of the bdry of the tree, the label of root is asymp indep of the rest of graph

Binary symmetric broadcast model

T : Galton-Watson tree with mean offspring distribution mean b.

Binary symmetric broadcast model

T : Galton-Watson tree with mean offspring distribution mean b.

Root R labeled uniformly $+1/-1$, each child takes parent's label with $\mathbb{P} = 1 - \eta$ and opposite label with $\mathbb{P} = \eta$.

Binary symmetric broadcast model

T : Galton-Watson tree with mean offspring distribution mean b.

Root R labeled uniformly $+1/-1$, each child takes parent's label with $\mathbb{P} = 1 - \eta$ and opposite label with $\mathbb{P} = \eta$.

Goal: reconstruct value of R from labels at level n.

Binary symmetric broadcast model

T : Galton-Watson tree with mean offspring distribution mean b.

Root R labeled uniformly $+1/-1$, each child takes parent's label with $\mathbb{P} = 1 - \eta$ and opposite label with $\mathbb{P} = \eta$.

Goal: reconstruct value of R from labels at level n.

Theorem (Evans, Kenyon, Peres, Schulman '00)

Probability of correct reconstruction of value of R tends to $\frac{1}{2}$ as $n \to \infty$ if

$$(1 - 2\eta)^2 \leq p_c(T),$$

where $p_c(T)$ is the critical probability for percolation on T.

Binary symmetric broadcast model

T : Galton-Watson tree with mean offspring distribution mean b.

Root R labeled uniformly $+1/-1$, each child takes parent's label with $\mathbb{P} = 1 - \eta$ and opposite label with $\mathbb{P} = \eta$.

Goal: reconstruct value of R from labels at level n.

Theorem (Evans, Kenyon, Peres, Schulman '00)

Probability of correct reconstruction of value of R tends to $\frac{1}{2}$ as $n \to \infty$ if

$$(1 - 2\eta)^2 \leq p_c(T),$$

where $p_c(T)$ is the critical probability for percolation on T.

Can think of $p_c(T)$ as the edge density at which the tree is connected.

Binary symmetric broadcast model

T : Galton-Watson tree with mean offspring distribution mean b.

Root R labeled uniformly $+1/-1$, each child takes parent's label with $\mathbb{P} = 1 - \eta$ and opposite label with $\mathbb{P} = \eta$.

Goal: reconstruct value of R from labels at level n.

Theorem (Evans, Kenyon, Peres, Schulman '00)

Probability of correct reconstruction of value of R tends to $\frac{1}{2}$ as $n \to \infty$ if

$$(1 - 2\eta)^2 \leq p_c(T),$$

where $p_c(T)$ is the critical probability for percolation on T.

Can think of $p_c(T)$ as the edge density at which the tree is connected. trees with offspring distribution Pois$(\frac{a+b}{2})$ and take $1 - \eta = \frac{a}{a+b}$.

Binary symmetric broadcast model

T : Galton-Watson tree with mean offspring distribution mean b.

Root R labeled uniformly $+1/-1$, each child takes parent's label with $\mathbb{P} = 1 - \eta$ and opposite label with $\mathbb{P} = \eta$.

Goal: reconstruct value of R from labels at level n.

Theorem (Evans, Kenyon, Peres, Schulman '00)

Probability of correct reconstruction of value of R tends to $\frac{1}{2}$ as $n \to \infty$ if

$$(1 - 2\eta)^2 \leq p_c(T),$$

where $p_c(T)$ is the critical probability for percolation on T.

Can think of $p_c(T)$ as the edge density at which the tree is connected.
trees with offspring distribution Pois$(\frac{a+b}{2})$ and take $1 - \eta = \frac{a}{a+b}$.
Then threshold reduces to $(a - b)^2 \leq 2(a + b)$.

Previous work - Spectral methods

- Applying some classical results to bipartite model using spectrum with $p = O(1/n_1)$ recovers partition

Previous work - Spectral methods

- Applying some classical results to bipartite model using spectrum with $p = O(1/n_1)$ recovers partition
- typical analysis of spectral algos: 2nd singular value $>$ spectral norm of noise matrix $M - \mathbb{E}M$;

Previous work - Spectral methods

- Applying some classical results to bipartite model using spectrum with $p = O(1/n_1)$ recovers partition
- typical analysis of spectral algos: 2nd singular value $>$ spectral norm of noise matrix $M - \mathbb{E}M$;
- here $\lambda_2(\mathbb{E}M) = \tilde{\Theta}(p\sqrt{n_1 n_2})$, norm of noise $||M - \mathbb{E}M|| = \tilde{\Theta}(\sqrt{pn_2})$.

Previous work - Spectral methods

- Applying some classical results to bipartite model using spectrum with $p = O(1/n_1)$ recovers partition
- typical analysis of spectral algos: 2nd singular value $>$ spectral norm of noise matrix $M - \mathbb{E}M$;
- here $\lambda_2(\mathbb{E}M) = \tilde{\Theta}(p\sqrt{n_1 n_2})$, norm of noise $||M - \mathbb{E}M|| = \tilde{\Theta}(\sqrt{pn_2})$.
- Feldman, Perkins, Vempala '14: subsampled power iteration recovers partition whp with $p = \tilde{O}((n_1 n_2)^{-1/2})$

Questions

1. Here $\lambda_2 < ||M - \mathbb{E}M||$. Is SVD doomed for $p \ll 1/n_1$?

Questions

1. Here $\lambda_2 < ||M - \mathbb{E}M||$. Is SVD doomed for $p \ll 1/n_1$?

2. What is the optimal threshold for detection in BSBM?

Our results - sharp reconstruction/impossibility

Theorem

On the other hand, if $n_2 \geq n_1$ and

$$p \leq \frac{1}{(\delta - 1)^2 \sqrt{n_1 n_2}},$$

then no algorithm can detect the partition.

Our results - sharp reconstruction/impossibility

Theorem

On the other hand, if $n_2 \geq n_1$ and

$$p \leq \frac{1}{(\delta - 1)^2 \sqrt{n_1 n_2}},$$

then no algorithm can detect the partition.

Idea: Couple to a broadcast model on a **multi-type** Galton Watson tree. Show that conditioned on the labels of a $\log n$ bdry of the tree, the label of root is asymp indep of the rest of graph.

Our results - sharp reconstruction/impossibility

Theorem

Let $n_2 \gg n_1$. Then there is a polynomial-time algorithm that detects the partition $V_1 = A_1 \cup B_1$ if

$$p > \frac{1+\epsilon}{(\delta-1)^2\sqrt{n_1 n_2}}$$

for any fixed $\epsilon > 0$.

Our results - sharp reconstruction/impossibility

Theorem

Let $n_2 \gg n_1$. Then there is a polynomial-time algorithm that detects the partition $V_1 = A_1 \cup B_1$ if

$$p > \frac{1+\epsilon}{(\delta-1)^2\sqrt{n_1 n_2}}$$

for any fixed $\epsilon > 0$.

Idea: reduce to SBM on graph on V_1 induced by paths of length 2 in bipartite graph.

Proof sketch

- Reduce to a graph G' by replacing each path of length 2 from V_1 to V_2 back to V_1 with a single edge between the endpoints in V_1.

Proof sketch

- Reduce to a graph G' by replacing each path of length 2 from V_1 to V_2 back to V_1 with a single edge between the endpoints in V_1.

Proof sketch

- Reduce to a graph G' by replacing each path of length 2 from V_1 to V_2 back to V_1 with a single edge between the endpoints in V_1.
- $|E'| = \frac{(1+\epsilon)^2 n_1}{(\delta-1)^4}(1 + o(1))$

Proof sketch

- Reduce to a graph G' by replacing each path of length 2 from V_1 to V_2 back to V_1 with a single edge between the endpoints in V_1.
- $|E'| = \frac{(1+\epsilon)^2 n_1}{(\delta-1)^4}(1 + o(1))$

Proof sketch

- Reduce to a graph G' by replacing each path of length 2 from V_1 to V_2 back to V_1 with a single edge between the endpoints in V_1.
- $|E'| = \frac{(1+\epsilon)^2 n_1}{(\delta-1)^4}(1+o(1))$
- Now we can compute $p_a = \mathbb{P}[e = (u,v) | \sigma(u) = \sigma(v)]$ and $p_b = \mathbb{P}[e = (u,v) | \sigma(u) \neq \sigma(v)]$

Proof sketch

- Reduce to a graph G' by replacing each path of length 2 from V_1 to V_2 back to V_1 with a single edge between the endpoints in V_1.
- $|E'| = \frac{(1+\epsilon)^2 n_1}{(\delta-1)^4}(1+o(1))$
- Now we can compute $p_a = \mathbb{P}[e = (u,v) | \sigma(u) = \sigma(v)]$ and $p_b = \mathbb{P}[e = (u,v) | \sigma(u) \neq \sigma(v)]$

Proof sketch

- Reduce to a graph G' by replacing each path of length 2 from V_1 to V_2 back to V_1 with a single edge between the endpoints in V_1.
- $|E'| = \frac{(1+\epsilon)^2 n_1}{(\delta-1)^4}(1+o(1))$
- Now we can compute $p_a = \mathbb{P}[e=(u,v)|\sigma(u)=\sigma(v)]$ and $p_b = \mathbb{P}[e=(u,v)|\sigma(u)\neq\sigma(v)]$
- Now compute a and b accordingly:

$$a = \frac{(1+\epsilon)(2-2\delta+\delta^2)}{(\delta-1)^4}(1+o(1)) \quad b = \frac{(1+\epsilon)(2\delta-\delta^2)}{(\delta-1)^4}(1+o(1))$$

Proof sketch

- Reduce to a graph G' by replacing each path of length 2 from V_1 to V_2 back to V_1 with a single edge between the endpoints in V_1.
- $|E'| = \frac{(1+\epsilon)^2 n_1}{(\delta-1)^4}(1+o(1))$
- Now we can compute $p_a = \mathbb{P}[e=(u,v)|\sigma(u)=\sigma(v)]$ and $p_b = \mathbb{P}[e=(u,v)|\sigma(u)\neq\sigma(v)]$
- Now compute a and b accordingly:

$$a = \frac{(1+\epsilon)(2-2\delta+\delta^2)}{(\delta-1)^4}(1+o(1)) \quad b = \frac{(1+\epsilon)(2\delta-\delta^2)}{(\delta-1)^4}(1+o(1))$$

Proof sketch

- Reduce to a graph G' by replacing each path of length 2 from V_1 to V_2 back to V_1 with a single edge between the endpoints in V_1.
- $|E'| = \frac{(1+\epsilon)^2 n_1}{(\delta-1)^4}(1+o(1))$
- Now we can compute $p_a = \mathbb{P}[e = (u,v)|\sigma(u) = \sigma(v)]$ and $p_b = \mathbb{P}[e = (u,v)|\sigma(u) \neq \sigma(v)]$
- Now compute a and b accordingly:

$$a = \frac{(1+\epsilon)(2-2\delta+\delta^2)}{(\delta-1)^4}(1+o(1)) \quad b = \frac{(1+\epsilon)(2\delta-\delta^2)}{(\delta-1)^4}(1+o(1))$$

- Apply criterion $(a-b)^2 \geq (1+\epsilon)2(a+b)$.

Implications for planted k-SAT

- detection in the block model exhibits a sharp threshold at

$$m^* = \Theta(n^{r/2}) \text{ hyperedges/clauses}$$

Implications for planted k-SAT

- detection in the block model exhibits a sharp threshold at

$$m^* = \Theta(n^{r/2}) \text{ hyperedges/clauses}$$

Definition

The distribution complexity r of the planting distribution Q is the smallest $r > 0$ so that Q is an $(r-1)$-wise independent distribution on $\{\pm\}^k$ but not r-wise independent.

Spectral algorithms

- Standard SVD: Compute left singular vector of M (adjacency matrix) corresponding to 2nd singular value, round signs to get v; compare σ and v

Spectral algorithms

- Standard SVD: Compute left singular vector of M (adjacency matrix) corresponding to 2nd singular value, round signs to get v; compare σ and v
- Diagonal deletion SVD: Set diagonal entries of MM^T to 0, compute second eigenvector, round signs to get v; compare σ and v

Our results - spectral

Theorem

Let $n_2 \gg n_1$, with $n_1 \to \infty$. Then

1. If $p_D > (n_1 n_2)^{-1/2}$, then whp the diagonal deletion SVD algorithm recovers the partition $V_1 = A_1 \cup B_1$.

Our results - spectral

Theorem

Let $n_2 \gg n_1$, with $n_1 \to \infty$. Then

1. If $p_D > (n_1 n_2)^{-1/2}$, then whp the diagonal deletion SVD algorithm recovers the partition $V_1 = A_1 \cup B_1$.

2. If $p_V > n_1^{-2/3} n_2^{-1/3}$, then whp the standard SVD algorithm recovers the partition.

Our results - spectral

Theorem

Let $n_2 \gg n_1$, with $n_1 \to \infty$. Then

1. If $p_D > (n_1 n_2)^{-1/2}$, then whp the diagonal deletion SVD algorithm recovers the partition $V_1 = A_1 \cup B_1$.

2. If $p_V > n_1^{-2/3} n_2^{-1/3}$, then whp the standard SVD algorithm recovers the partition.

Our results - spectral

Theorem

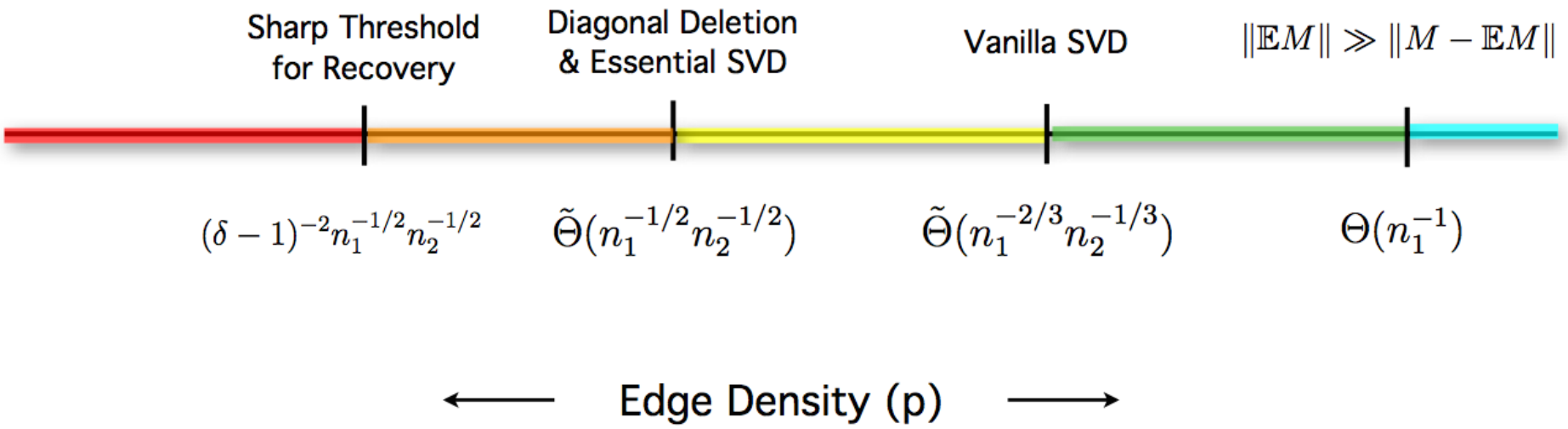
Let $n_2 \gg n_1$, with $n_1 \to \infty$. Then

1. If $p_D > (n_1 n_2)^{-1/2}$, then whp the diagonal deletion SVD algorithm recovers the partition $V_1 = A_1 \cup B_1$.

2. If $p_V > n_1^{-2/3} n_2^{-1/3}$, then whp the standard SVD algorithm recovers the partition.

When $n_2 = n^2$, $p_D \approx n^{-3/2}$, $p_V \approx n^{-4/3}$.

Timeline


Sharp Threshold for Recovery
Diagonal Deletion & Essential SVD
Vanilla SVD
$\|\mathbb{E}M\| \gg \|M - \mathbb{E}M\|$
$(\delta - 1)^{-2} n_1^{-1/2} n_2^{-1/2}$
$\tilde{\Theta}(n_1^{-1/2} n_2^{-1/2})$
$\tilde{\Theta}(n_1^{-2/3} n_2^{-1/3})$
$\Theta(n_1^{-1})$
Edge Density (p)

Our results

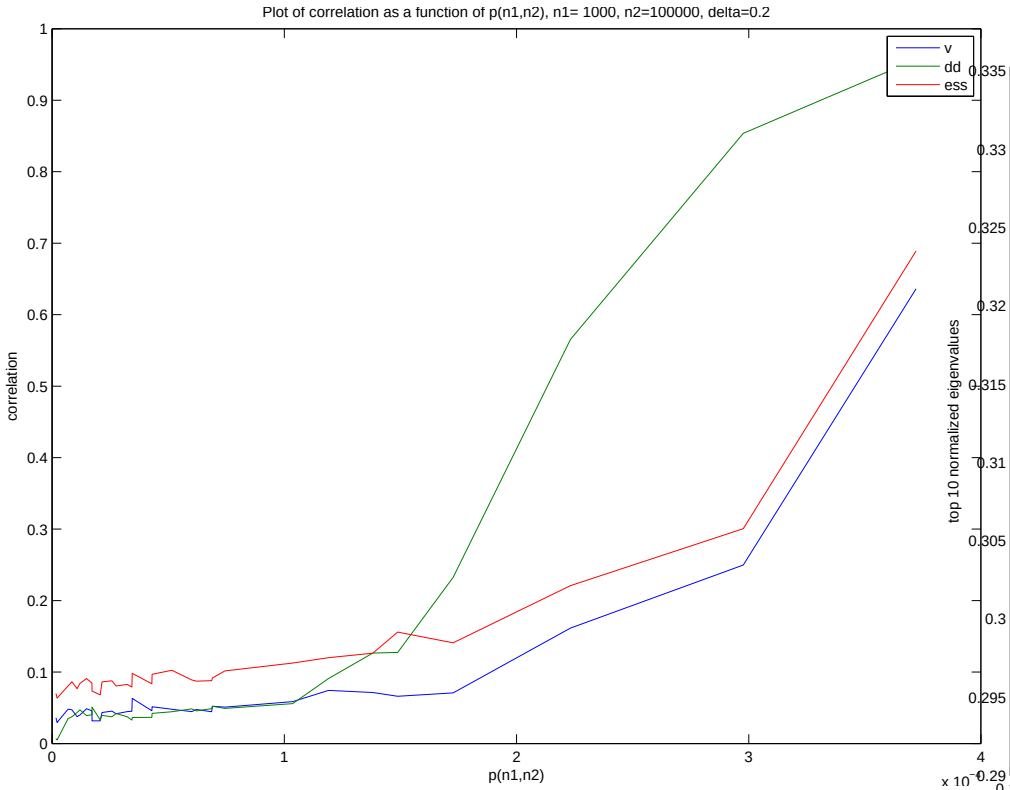


Figure: Correlations of computed vectors with planted vector

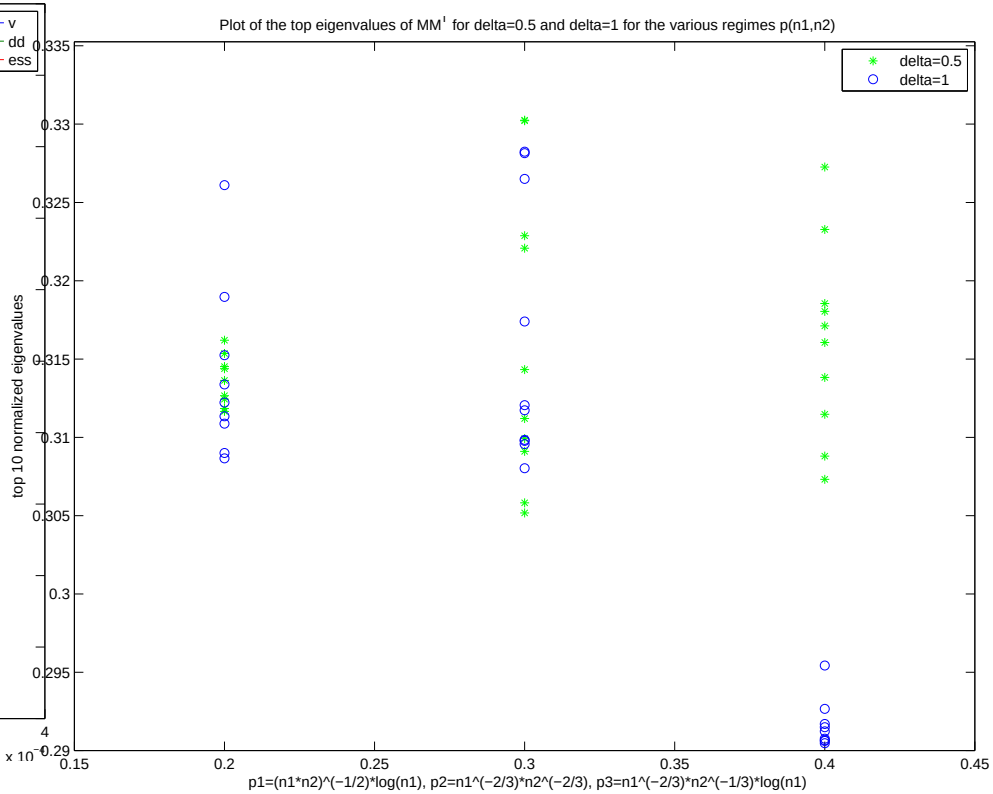


Figure: Eigenvalue separation

Thresholds origins

- DiagD: $B = MM^T - D_V$, SVD: $B' = B + D_V$

Thresholds origins

- DiagD: $B = MM^T - D_V$, SVD: $B' = B + D_V$
- σ: partition, $e_2(B)$: second largest eigenvector of B, D_V: degrees.

Thresholds origins

- DiagD: $B = MM^T - D_V$, SVD: $B' = B + D_V$
- σ: partition, $e_2(B)$: second largest eigenvector of B, D_V: degrees.

- DiagD: $\sin(B, \mathbb{E}B) \leq \frac{C\|B-\mathbb{E}B\|}{\lambda_2}$ SVD: $\sin(B', \mathbb{E}B') \leq \frac{C\|B-\mathbb{E}B\|+\|D_V-\mathbb{E}D_V\|}{\lambda_2}$
  by Sin Theta Theorem - sin of angle between eigenvector spaces $\leq$ norm/eigenvalue gap

Thresholds origins

- DiagD: $B = MM^T - D_V$, SVD: $B' = B + D_V$
- σ: partition, $e_2(B)$: second largest eigenvector of B, D_V: degrees.

- DiagD: $\sin(B, \mathbb{E}B) \leq \frac{C\|B-\mathbb{E}B\|}{\lambda_2}$ SVD: $\sin(B', \mathbb{E}B') \leq \frac{C\|B-\mathbb{E}B\|+\|D_V-\mathbb{E}D_V\|}{\lambda_2}$
  by Sin Theta Theorem - sin of angle between eigenvector spaces $\leq$ norm/eigenvalue gap

- $\leq C\frac{n_1^{1/2}n_2^{1/2}p}{(\delta-1)^2 n_1 n_2 p^2}$; $\leq C\frac{n_1^{1/2}n_2^{1/2}p+(C\sqrt{n_2 p \log n_1})}{(\delta-1)^2 n_1 n_2 p^2}$
  (2nd λ asymptotics)

Thresholds origins

- DiagD: $B = MM^T - D_V$, SVD: $B' = B + D_V$
- σ: partition, $e_2(B)$: second largest eigenvector of B, D_V: degrees.

- DiagD: $\sin(B, \mathbb{E}B) \leq \frac{C\|B-\mathbb{E}B\|}{\lambda_2}$ SVD: $\sin(B', \mathbb{E}B') \leq \frac{C\|B-\mathbb{E}B\|+\|D_V-\mathbb{E}D_V\|}{\lambda_2}$
  by Sin Theta Theorem - sin of angle between eigenvector spaces $\leq$ norm/eigenvalue gap

- $\leq C\frac{n_1^{1/2}n_2^{1/2}p}{(\delta-1)^2 n_1 n_2 p^2}$; $\leq C\frac{n_1^{1/2}n_2^{1/2}p+(C\sqrt{n_2 p \log n_1})}{(\delta-1)^2 n_1 n_2 p^2}$
  (2nd λ asymptotics)

- $= O\left(\frac{1}{\log n_1}\right)$; $= O\left(\frac{1}{\log n_1}\right)$

Thresholds origins

- DiagD: $B = MM^T - D_V$, SVD: $B' = B + D_V$
- σ: partition, $e_2(B)$: second largest eigenvector of B, D_V: degrees.

- DiagD: $\sin(B, \mathbb{E}B) \leq \frac{C\|B-\mathbb{E}B\|}{\lambda_2}$ SVD: $\sin(B', \mathbb{E}B') \leq \frac{C\|B-\mathbb{E}B\|+\|D_V-\mathbb{E}D_V\|}{\lambda_2}$
  by Sin Theta Theorem - sin of angle between eigenvector spaces $\leq$ norm/eigenvalue gap

- $\leq C\frac{n_1^{1/2}n_2^{1/2}p}{(\delta-1)^2 n_1 n_2 p^2}$; $\leq C\frac{n_1^{1/2}n_2^{1/2}p+(C\sqrt{n_2 p \log n_1})}{(\delta-1)^2 n_1 n_2 p^2}$
  (2nd λ asymptotics)

- $= O\left(\frac{1}{\log n_1}\right)$; $= O\left(\frac{1}{\log n_1}\right)$

- $\|e_2(B) - \sigma/\sqrt{n_1}\| = O(\log^{-1} n_1)$ (by special case of Sin Theta Theorem).

Thresholds origins

- DiagD: $B = MM^T - D_V$, SVD: $B' = B + D_V$
- σ: partition, $e_2(B)$: second largest eigenvector of B, D_V: degrees.
- DiagD: $\sin(B, \mathbb{E}B) \leq \frac{C\|B-\mathbb{E}B\|}{\lambda_2}$ SVD: $\sin(B', \mathbb{E}B') \leq \frac{C\|B-\mathbb{E}B\|+\|D_V-\mathbb{E}D_V\|}{\lambda_2}$
  by Sin Theta Theorem - sin of angle between eigenvector spaces $\leq$ norm/eigenvalue gap
- $\leq C\frac{n_1^{1/2}n_2^{1/2}p}{(\delta-1)^2 n_1 n_2 p^2}$; $\leq C\frac{n_1^{1/2}n_2^{1/2}p+(C\sqrt{n_2 p \log n_1})}{(\delta-1)^2 n_1 n_2 p^2}$
  (2nd λ asymptotics)
- $= O\left(\frac{1}{\log n_1}\right)$; $= O\left(\frac{1}{\log n_1}\right)$
- $\|e_2(B) - \sigma/\sqrt{n_1}\| = O(\log^{-1} n_1)$ (by special case of Sin Theta Theorem).
- Conclude by rounding signs of $e_2(B)$.

Conclusions

Theorem

- Can efficiently detect partition in BSBM if $p > \frac{1+\epsilon}{(\delta-1)^2\sqrt{n_1 n_2}}$
- Cannot detect if $p \leq \frac{1}{(\delta-1)^2\sqrt{n_1 n_2}}$

Conclusions

Theorem

- Can efficiently detect partition in BSBM if $p > \frac{1+\epsilon}{(\delta-1)^2\sqrt{n_1 n_2}}$
- Cannot detect if $p \leq \frac{1}{(\delta-1)^2\sqrt{n_1 n_2}}$

- spectral method still works if $\lambda_2 \leq$ norm of noise matrix

Conclusions

Theorem

- Can efficiently detect partition in BSBM if $p > \frac{1+\epsilon}{(\delta-1)^2\sqrt{n_1 n_2}}$
- Cannot detect if $p \leq \frac{1}{(\delta-1)^2\sqrt{n_1 n_2}}$

- spectral method still works if $\lambda_2 \leq$ norm of noise matrix
- modifying adjacency matrix improves recovery significantly

Open problems

- apply Diagonal Deletion type of algorithm for improvement over SVD in other problems?

Open problems

- apply Diagonal Deletion type of algorithm for improvement over SVD in other problems?
- sharper detection thresholds for planted k-SAT?

Thank you!