The Threshold Value for the Planted Partition Model

Elchanan Mossel

University of California, Berkeley

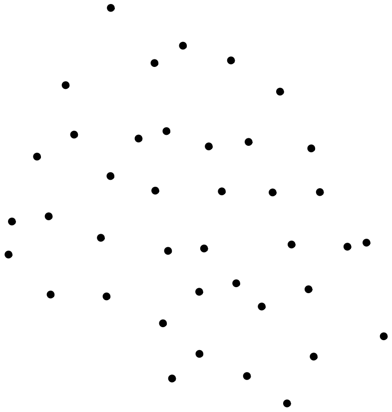
May 31, 2012

Based on a joint work with:

1. Joe Neeman (U.C. Berkeley)
2. Allan Sly (U.C. Berkeley)

The block (aka. planted partition) model

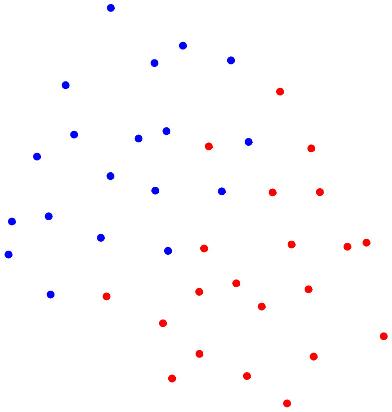
This is a model for a random graph on n nodes. It takes two parameters, $a, b \geq 0$.



The block (aka. planted partition) model

This is a model for a random graph on n nodes. It takes two parameters, $a, b \geq 0$.

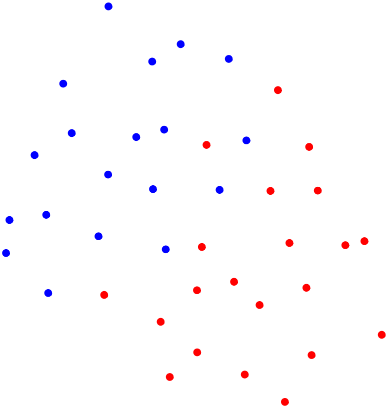
1. Label the vertices $+$ or $-$, uniformly at random.



The block (aka. planted partition) model

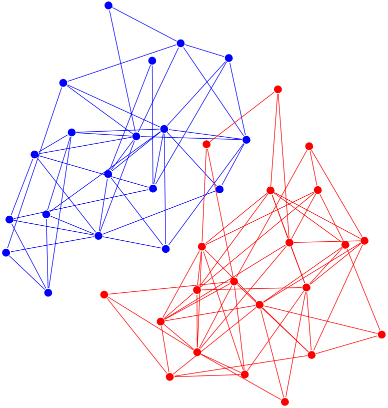
This is a model for a random graph on n nodes. It takes two parameters, $a, b \geq 0$.

1. Label the vertices $+$ or $-$, uniformly at random.
2. Independently for each edge (u, v):



The block (aka. planted partition) model

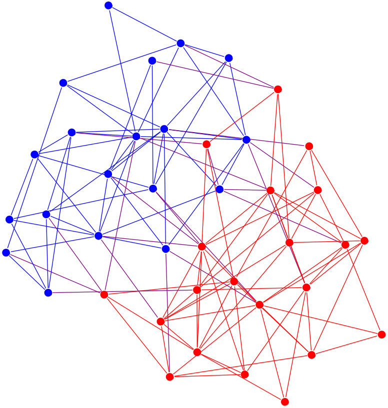
This is a model for a random graph on n nodes. It takes two parameters, $a, b \geq 0$.



1. Label the vertices $+$ or $-$, uniformly at random.
2. Independently for each edge (u, v):
   - if u and v have the same label, include the edge with probability a/n;

The block (aka. planted partition) model

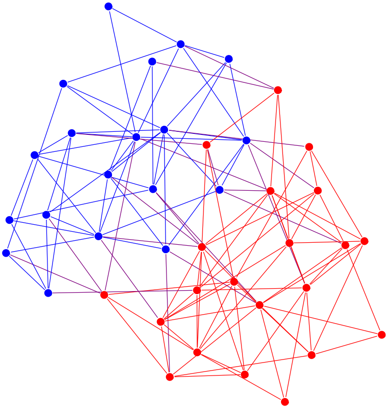
This is a model for a random graph on n nodes. It takes two parameters, $a, b \geq 0$.



1. Label the vertices $+$ or $-$, uniformly at random.
2. Independently for each edge (u, v):
   - if u and v have the same label, include the edge with probability a/n;
   - if u and v have different labels, include the edge with probability b/n.

The block (aka. planted partition) model

This is a model for a random graph on n nodes. It takes two parameters, $a, b \geq 0$.



1. Label the vertices $+$ or $-$, uniformly at random.
2. Independently for each edge (u, v):
   - if u and v have the same label, include the edge with probability a/n;
   - if u and v have different labels, include the edge with probability b/n.

Variations: more than two classes, un-balanced classes, d-regular, ...

Motivation (computer science)

The graph-bisection problem is NP-hard. Could it be easy on average?

Motivation (computer science)

The graph-bisection problem is NP-hard. Could it be easy on average? Yes:

Bui et al '84	min-cut	$a \sim c,\ b = O(n^{-2/(a+b+1)})$
Dyer-Frieze '89	vertex degree	$a - b = \Omega(n)$
Boppana '87	spectral	$\frac{a-b}{\sqrt{a+b}} = \Omega(\log n)$
Juels '96 Carson-Impagliazzo '01	hill-climbing hill-climbing	$a - b = \Omega(n)$ $\frac{a-b}{\sqrt{a+b}} = \Omega(\log n)$
Jerrum-Sorkin '98	Metropolis	$a - b = \Omega(n^{5/6+\epsilon})$
Condon-Karp '01	greedy	$a - b = \Omega(n^{1/2+\epsilon})$
McSherry '01	correlation	$\frac{a-b}{\sqrt{a+b}} = \Omega(\log n)$
Coja-Oghlan '10	spectral	$\frac{a-b}{\sqrt{a+b}} = \Omega(1)$

Motivation (computer science)

The graph-bisection problem is NP-hard. Could it be easy on average? Yes:

Bui et al '84	min-cut	$a \sim c,\ b = O(n^{-2/(a+b+1)})$
Dyer-Frieze '89	vertex degree	$a - b = \Omega(n)$
Boppana '87	spectral	$\frac{a-b}{\sqrt{a+b}} = \Omega(\log n)$
Juels '96 Carson-Impagliazzo '01	hill-climbing hill-climbing	$a - b = \Omega(n)$ $\frac{a-b}{\sqrt{a+b}} = \Omega(\log n)$
Jerrum-Sorkin '98	Metropolis	$a - b = \Omega(n^{5/6+\epsilon})$
Condon-Karp '01	greedy	$a - b = \Omega(n^{1/2+\epsilon})$
McSherry '01	correlation	$\frac{a-b}{\sqrt{a+b}} = \Omega(\log n)$
Coja-Oghlan '10	spectral	$\frac{a-b}{\sqrt{a+b}} = \Omega(1)$

Motivation (statistics, physics and social computing)

Clustering network data. This model was introduced in statistics by Holland-Laskey-Leinhardt '83. Both parameter estimation and label recovery are studied.

A basic model of communities (see Lancichinetti and Fortunato)

Motivation (statistics, physics and social computing)

Clustering network data. This model was introduced in statistics by Holland-Laskey-Leinhardt '83. Both parameter estimation and label recovery are studied.
A basic model of communities (see Lancichinetti and Fortunato)

Snijders-Nowicki '97	ML, EM, etc.	$a - b = \Omega(n)$
Bickel-Chen '09	G-N modularity	$\frac{a-b}{\sqrt{a+b}} = \Omega(\log n)$
Chatterjee-Rohe-Yu '10	spectral	$a - b = \Omega(n)$
Choi-Wolfe-Airoldi '10	ML	$a + b = \Omega(\log^3 n)$, $\frac{a-b}{\sqrt{a+b}} = \Omega(1)$

Many, many more algorithms without performance guarantees (survey in Lancichinetti-Fortunato '09).

A Quiz

- **Q**: Given the $+/-$ status of the vertices what is the distribution of edges?

A Quiz

- **Q**: Given the $+/-$ status of the vertices what is the distribution of edges?
- **A:** $(+,+)$ with probability a/n and $(+,-)$ with probability b/n.

A Quiz

- **Q**: Given the $+/-$ status of the vertices what is the distribution of edges?
- **A:** $(+,+)$ with probability a/n and $(+,-)$ with probability b/n.
- **Q:** Given the graph G what is the distribution of the $+,-$ on the vertices?

A Quiz

- **Q**: Given the $+/-$ status of the vertices what is the distribution of edges?
- **A:** $(+,+)$ with probability a/n and $(+,-)$ with probability b/n.
- **Q:** Given the graph G what is the distribution of the $+,-$ on the vertices?
- **A:** It's P

$$P(\sigma) = Z^{-1} a^{|\{(u,v) \in G : \sigma(u) = \sigma(v)\}|} b^{|\{(u,v) \in G : \sigma(u) \neq \sigma(v)\}|}$$

A Quiz

- **Q**: Given the $+/-$ status of the vertices what is the distribution of edges?
- **A:** $(+,+)$ with probability a/n and $(+,-)$ with probability b/n.
- **Q:** Given the graph G what is the distribution of the $+,-$ on the vertices?
- **A:** It's P

$$P(\sigma) = Z^{-1} a^{|\{(u,v) \in G : \sigma(u) = \sigma(v)\}|} b^{|\{(u,v) \in G : \sigma(u) \neq \sigma(v)\}|}$$

- Also known as the Ising model on G!

A Quiz

- **Q**: Given the $+/-$ status of the vertices what is the distribution of edges?
- **A:** $(+,+)$ with probability a/n and $(+,-)$ with probability b/n.
- **Q:** Given the graph G what is the distribution of the $+,-$ on the vertices?
- **A:** It's P

$$P(\sigma) = Z^{-1} a^{|\{(u,v) \in G : \sigma(u) = \sigma(v)\}|} b^{|\{(u,v) \in G : \sigma(u) \neq \sigma(v)\}|}$$

- Also known as the Ising model on G!
- Is this correct?

A Quiz

- **Q**: Given the $+/-$ status of the vertices what is the distribution of edges?
- **A:** $(+,+)$ with probability a/n and $(+,-)$ with probability b/n.
- **Q:** Given the graph G what is the distribution of the $+,-$ on the vertices?
- **A:** It's P

$$P(\sigma) = Z^{-1} a^{|\{(u,v)\in G:\sigma(u)=\sigma(v)\}|} b^{|\{(u,v)\in G:\sigma(u)\neq\sigma(v)\}|}$$

- Also known as the Ising model on G!
- Is this correct?
- Almost - it's actually Q where $Q(\sigma) = P(\sigma | \sum_v \sigma_v = 0)$.

A Quiz

- **Q**: Given the $+/-$ status of the vertices what is the distribution of edges?
- **A:** $(+,+)$ with probability a/n and $(+,-)$ with probability b/n.
- **Q:** Given the graph G what is the distribution of the $+,-$ on the vertices?
- **A:** It's P

$$P(\sigma) = Z^{-1} a^{|\{(u,v)\in G : \sigma(u)=\sigma(v)\}|} b^{|\{(u,v)\in G : \sigma(u)\neq\sigma(v)\}|}$$

- Also known as the Ising model on G!
- Is this correct?
- Almost - it's actually Q where $Q(\sigma) = P(\sigma | \sum_v \sigma_v = 0)$.
- Well done!

The sparse case: a phase transition

Decelle-Krzakala-Moore-Zdeborová '11: "recovery" means getting a partition that is positively correlated with the truth.

Conjecture

If

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

then recovery is possible. If

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

then recovery is impossible.

The sparse case: a phase transition

Decelle-Krzakala-Moore-Zdeborová '11: "recovery" means getting a partition that is positively correlated with the truth.

Conjecture

If

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

then recovery is possible. If

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

then recovery is impossible.

"Physics" proof using belief propagation.

Non-Reconstruction results of M-Neeman-Sly

Theorem

If $a+b>2$ and $(a-b)^2 \leq 2(a+b)$ then, for any fixed vertices u and v,

$$\mathbb{P}_n(\sigma_u = + | G, \sigma_v = +) \to \frac{1}{2}$$

$\implies$ *impossible to recover a partition that is correlated with the true partition.*

Theorem

Let $\mathbb{P}'_n$ be the law of $G(n, \frac{a+b}{2n})$. If $(a-b)^2 < 2(a+b)$ then $\mathbb{P}_n$ and $\mathbb{P}'_n$ are mutually contiguous i.e., for a sequence of events A_n, $\mathbb{P}_n(A_n) \to 0$ if, and only if, $\mathbb{P}'_n(A_n) \to 0$.

Reconstruction Results of M-Neeman-Sly

Theorem

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

- *The parameters* a, b *are identifiable.*
- *Let* $\mathbb{P}'_n$ *be the law of* $G(n, \frac{a+b}{2n})$. *Then* $\mathbb{P}_n$ *and* $\mathbb{P}'_n$ *are asymptotically orthogonal. In other words, there exist events* A_n *such that* $\mathbb{P}_n(A_n) \to 1$ *and* $\mathbb{P}'_n(A_n) \to 0$.

Proof Sketch - non-reconstruction

- Tree neighborhood looks like a broadcast process on a tree.

Proof Sketch - non-reconstruction

- Tree neighborhood looks like a broadcast process on a tree.
- The reconstruction problem on the tree is not-solvable if $(a-b)^2 < 2(a+b)$ (EKPS 2000).

Proof Sketch - non-reconstruction

- Tree neighborhood looks like a broadcast process on a tree.
- The reconstruction problem on the tree is not-solvable if $(a-b)^2 < 2(a+b)$ (EKPS 2000).
- Given fixed u, v, condition on $\partial B(v,r)$ for some large r. If $u \notin B(v,r)$ then

Proof Sketch - non-reconstruction

- Tree neighborhood looks like a broadcast process on a tree.
- The reconstruction problem on the tree is not-solvable if $(a-b)^2 < 2(a+b)$ (EKPS 2000).
- Given fixed u, v, condition on $\partial B(v,r)$ for some large r. If $u \notin B(v,r)$ then
- $\sigma_u \perp \sigma_{B(v,r)}$ by non-reconstruction and

Proof Sketch - non-reconstruction

- Tree neighborhood looks like a broadcast process on a tree.
- The reconstruction problem on the tree is not-solvable if $(a-b)^2 < 2(a+b)$ (EKPS 2000).
- Given fixed u, v, condition on $\partial B(v, r)$ for some large r. If $u \notin B(v, r)$ then
- $\sigma_u \perp \sigma_{B(v,r)}$ by non-reconstruction and
- $\sigma_u \perp \sigma_v | \sigma_{B(v,r)}$ (Markovian property) so

Proof Sketch - non-reconstruction

- Tree neighborhood looks like a broadcast process on a tree.
- The reconstruction problem on the tree is not-solvable if $(a-b)^2 < 2(a+b)$ (EKPS 2000).
- Given fixed u, v, condition on $\partial B(v,r)$ for some large r. If $u \notin B(v,r)$ then
- $\sigma_u \perp \sigma_{B(v,r)}$ by non-reconstruction and
- $\sigma_u \perp \sigma_v | \sigma_{B(v,r)}$ (Markovian property) so
- $\sigma_u \perp \sigma_v$.

Proof Sketch - non-reconstruction

- Tree neighborhood looks like a broadcast process on a tree.
- The reconstruction problem on the tree is not-solvable if $(a-b)^2 < 2(a+b)$ (EKPS 2000).
- Given fixed u, v, condition on $\partial B(v, r)$ for some large r. If $u \notin B(v, r)$ then
- $\sigma_u \perp \sigma_{B(v,r)}$ by non-reconstruction and
- $\sigma_u \perp \sigma_v | \sigma_{B(v,r)}$ (Markovian property) so
- $\sigma_u \perp \sigma_v$.
- This argument is not correct since we have global information on the number of $+$s.

Proof Sketch - non-reconstruction

- Tree neighborhood looks like a broadcast process on a tree.
- The reconstruction problem on the tree is not-solvable if $(a-b)^2 < 2(a+b)$ (EKPS 2000).
- Given fixed u, v, condition on $\partial B(v,r)$ for some large r. If $u \notin B(v,r)$ then
- $\sigma_u \perp \sigma_{B(v,r)}$ by non-reconstruction and
- $\sigma_u \perp \sigma_v | \sigma_{B(v,r)}$ (Markovian property) so
- $\sigma_u \perp \sigma_v$.
- This argument is not correct since we have global information on the number of $+$s.
- Still approximately correct for small enough neighborhoods.

Identifiability when $(a-b)^2 > 2(a+b)$

- $a+b$ is identifiable by looking at the total number of edges.

Identifiability when $(a-b)^2 > 2(a+b)$

- $a+b$ is identifiable by looking at the total number of edges.
- Claim: Let $X_{k,n}$ denote the number of k-cycles where $k = O(\log^{1/4}(n))$. Then

$$X_{k,n} \to Pois\left(\frac{1}{k2^{k+1}}((a+b)^k + (a-b)^k)\right).$$

$$\lim \mathbb{E}\left[X_{k,n}\right] = \frac{1}{2k}\left(\left(\frac{a+b}{2}\right)^k + \left(\frac{a-b}{2}\right)^k\right)$$

$$\lim Var[X_{k,n}] = \frac{1}{2k}(1+o_k(1))\left(\frac{a+b}{2}\right)^k.$$

Identifiability when $(a-b)^2 > 2(a+b)$

- $a+b$ is identifiable by looking at the total number of edges.
- Claim: Let $X_{k,n}$ denote the number of k-cycles where $k = O(\log^{1/4}(n))$. Then

$$X_{k,n} \to Pois\left(\frac{1}{k2^{k+1}}((a+b)^k + (a-b)^k)\right).$$

$$\lim \mathbb{E}\left[X_{k,n}\right] = \frac{1}{2k}\left(\left(\frac{a+b}{2}\right)^k + \left(\frac{a-b}{2}\right)^k\right)$$

$$\lim Var[X_{k,n}] = \frac{1}{2k}(1+o_k(1))\left(\frac{a+b}{2}\right)^k.$$

- For large k if

$$\left(\frac{a-b}{2}\right)^2 > \frac{a+b}{2}$$

can detect the difference in mean.

Orthogonality to $G(n, \frac{a+b}{2}n)$ when $(a-b)^2 > 2(a+b)$

- Claim: Let $X_{k,n}$ denote the number of k-cycles where $k = O(\log^{1/4}(n))$. Then

$$X_{k,n} \to Pois\left(\frac{1}{k2^{k+1}}((a+b)^k + (a-b)^k)\right).$$

Orthogonality to $G(n, \frac{a+b}{2}n)$ when $(a-b)^2 > 2(a+b)$

- Claim: Let $X_{k,n}$ denote the number of k-cycles where $k = O(\log^{1/4}(n))$. Then

$$X_{k,n} \to Pois\left(\frac{1}{k2^{k+1}}((a+b)^k + (a-b)^k)\right).$$

- If $Y_{n,k}$ is the corresponding variable for $G(n, \frac{a+b}{2n})$ then

$$\lim \mathbb{E}\left[X_{k,n} - Y_{k,n}\right] = \frac{1}{2k}\left(\frac{a-b}{2}\right)^k$$

$$\lim Var[X_{k,n}], \lim Var[Y_{k,n}] = \frac{1}{2k}(1+o_k(1))\left(\frac{a+b}{2}\right)^k.$$

Orthogonality to $G(n, \frac{a+b}{2}n)$ when $(a-b)^2 > 2(a+b)$

- Claim: Let $X_{k,n}$ denote the number of k-cycles where $k = O(\log^{1/4}(n))$. Then

$$X_{k,n} \to Pois\left(\frac{1}{k2^{k+1}}((a+b)^k + (a-b)^k)\right).$$

- If $Y_{n,k}$ is the corresponding variable for $G(n, \frac{a+b}{2n})$ then

$$\lim \mathbb{E}\left[X_{k,n} - Y_{k,n}\right] = \frac{1}{2k}\left(\frac{a-b}{2}\right)^k$$

$$\lim Var[X_{k,n}], \lim Var[Y_{k,n}] = \frac{1}{2k}(1+o_k(1))\left(\frac{a+b}{2}\right)^k.$$

- For large k if

$$\left(\frac{a-b}{2}\right)^2 > \frac{a+b}{2}$$

then $X_{n,k}$ and $Y_{n,k}$ are almost orthogonal.

Contiguous with $G(n, \frac{a+b}{2}n)$ when $(a-b)^2 < 2(a+b)$

- Let P_n, Q_n be the distributions corresponding to $G(n, \frac{a}{n}, \frac{b}{n}), G(n, \frac{a+b}{2n})$.

Contiguous with $G(n, \frac{a+b}{2}n)$ when $(a-b)^2 < 2(a+b)$

- Let P_n, Q_n be the distributions corresponding to $G(n, \frac{a}{n}, \frac{b}{n}), G(n, \frac{a+b}{2n})$.
- We extend Q_n by letting $Q_n(\sigma|G) = \frac{P_n(G|\sigma)}{Z_n(G)}$.

Contiguous with $G(n, \frac{a+b}{2}n)$ when $(a-b)^2 < 2(a+b)$

- Let P_n, Q_n be the distributions corresponding to $G(n, \frac{a}{n}, \frac{b}{n}), G(n, \frac{a+b}{2n})$.
- We extend Q_n by letting $Q_n(\sigma|G) = \frac{P_n(G|\sigma)}{Z_n(G)}$.
- Let

$$Y_n := \frac{P_n(G,\sigma)}{Q_n(G,\sigma)} = \frac{P_n(\sigma)P_n(G|\sigma)Z_n(G)}{P_n(G|\sigma)} = 2^{-n}\frac{Z_n(G)}{Q_n(G)}.$$

Contiguous with $G(n, \frac{a+b}{2}n)$ when $(a-b)^2 < 2(a+b)$

- Let P_n, Q_n be the distributions corresponding to $G(n, \frac{a}{n}, \frac{b}{n}), G(n, \frac{a+b}{2n})$.
- We extend Q_n by letting $Q_n(\sigma|G) = \frac{P_n(G|\sigma)}{Z_n(G)}$.
- Let

$$Y_n := \frac{P_n(G,\sigma)}{Q_n(G,\sigma)} = \frac{P_n(\sigma)P_n(G|\sigma)Z_n(G)}{P_n(G|\sigma)} = 2^{-n}\frac{Z_n(G)}{Q_n(G)}.$$

- Working with the measure Q_n we see that $\mathbb{E}[Y_n] = 1$. Moreover, we show

$$\mathbb{E}\left[Y_n^2\right] = (1+o(1))\frac{e^{-t/2-t^2/4}}{\sqrt{1-t}}, \quad t = \frac{(a-b)^2}{2(a+b)}.$$

Contiguous with $G(n, \frac{a+b}{2}n)$ when $(a-b)^2 < 2(a+b)$

- Let P_n, Q_n be the distributions corresponding to $G(n, \frac{a}{n}, \frac{b}{n}), G(n, \frac{a+b}{2n})$.
- We extend Q_n by letting $Q_n(\sigma|G) = \frac{P_n(G|\sigma)}{Z_n(G)}$.
- Let

$$Y_n := \frac{P_n(G,\sigma)}{Q_n(G,\sigma)} = \frac{P_n(\sigma)P_n(G|\sigma)Z_n(G)}{P_n(G|\sigma)} = 2^{-n}\frac{Z_n(G)}{Q_n(G)}.$$

- Working with the measure Q_n we see that $\mathbb{E}\left[Y_n\right] = 1$. Moreover, we show

$$\mathbb{E}\left[Y_n^2\right] = (1+o(1))\frac{e^{-t/2-t^2/4}}{\sqrt{1-t}}, \quad t = \frac{(a-b)^2}{2(a+b)}.$$

- This already shows that $\lim_{\epsilon\to 0}\lim_{n\to\infty}\mathbb{P}(Y_n > \epsilon^{-1}) = 0$ goes to zero.

Contiguous with $G(n, \frac{a+b}{2}n)$ when $(a-b)^2 < 2(a+b)$

- Let P_n, Q_n be the distributions corresponding to $G(n, \frac{a}{n}, \frac{b}{n}), G(n, \frac{a+b}{2n})$.
- We extend Q_n by letting $Q_n(\sigma|G) = \frac{P_n(G|\sigma)}{Z_n(G)}$.
- Let

$$Y_n := \frac{P_n(G,\sigma)}{Q_n(G,\sigma)} = \frac{P_n(\sigma)P_n(G|\sigma)Z_n(G)}{P_n(G|\sigma)} = 2^{-n}\frac{Z_n(G)}{Q_n(G)}.$$

- Working with the measure Q_n we see that $\mathbb{E}\left[Y_n\right] = 1$. Moreover, we show

$$\mathbb{E}\left[Y_n^2\right] = (1+o(1))\frac{e^{-t/2-t^2/4}}{\sqrt{1-t}}, \quad t = \frac{(a-b)^2}{2(a+b)}.$$

- This already shows that $\lim_{\epsilon\to 0}\lim_{n\to\infty}\mathbb{P}(Y_n > \epsilon^{-1}) = 0$ goes to zero.
- Most of the work is devoted to apply the "small graph conditioning method" to show that $\lim_{\epsilon\to 0}\lim_{n\to\infty}\mathbb{P}(Y_n < \epsilon) = 0$.

Morals and open problems

- Cluster information is detectible even in cases where it is impossible to identify with certainty the cluster identity of any individual node.
- Algorithms such as Belief Propagation are likely to be effective in detecting community structures.
- Challenge: efficient algorithm for community detection.
- Challenge: Extension to other models.