On the optimality of local belief propagation under the degree-correlated stochastic block model

Jiaming Xu

Department of Statistics, The Wharton School
University of Pennsylvania

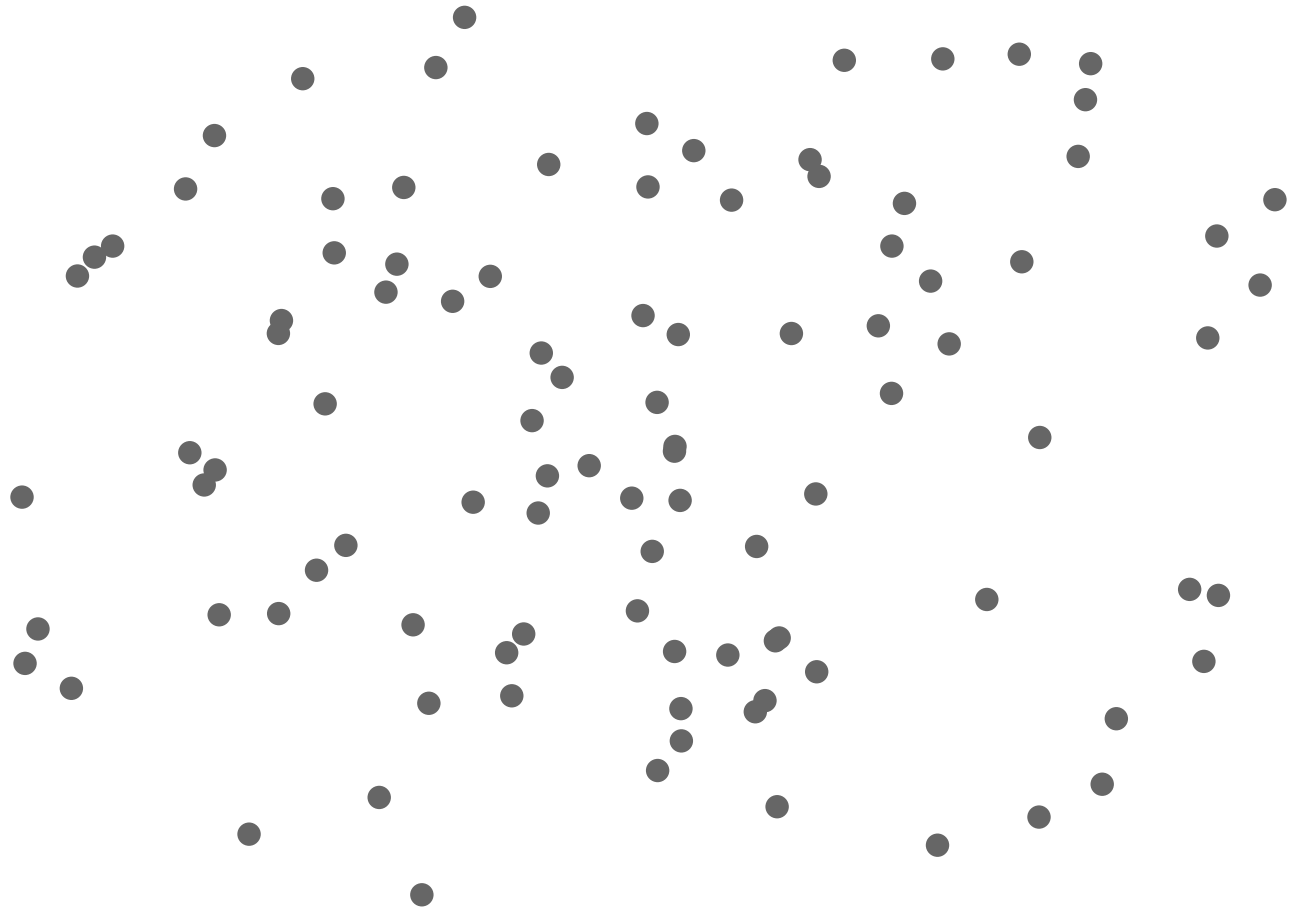
jiamingx@wharton.upenn.edu

Joint work with Elchanan Mossel (Penn and UC Berkeley)

October 14, 2015

Stochastic block model [Holland-Laskey-Leinhardt '83]

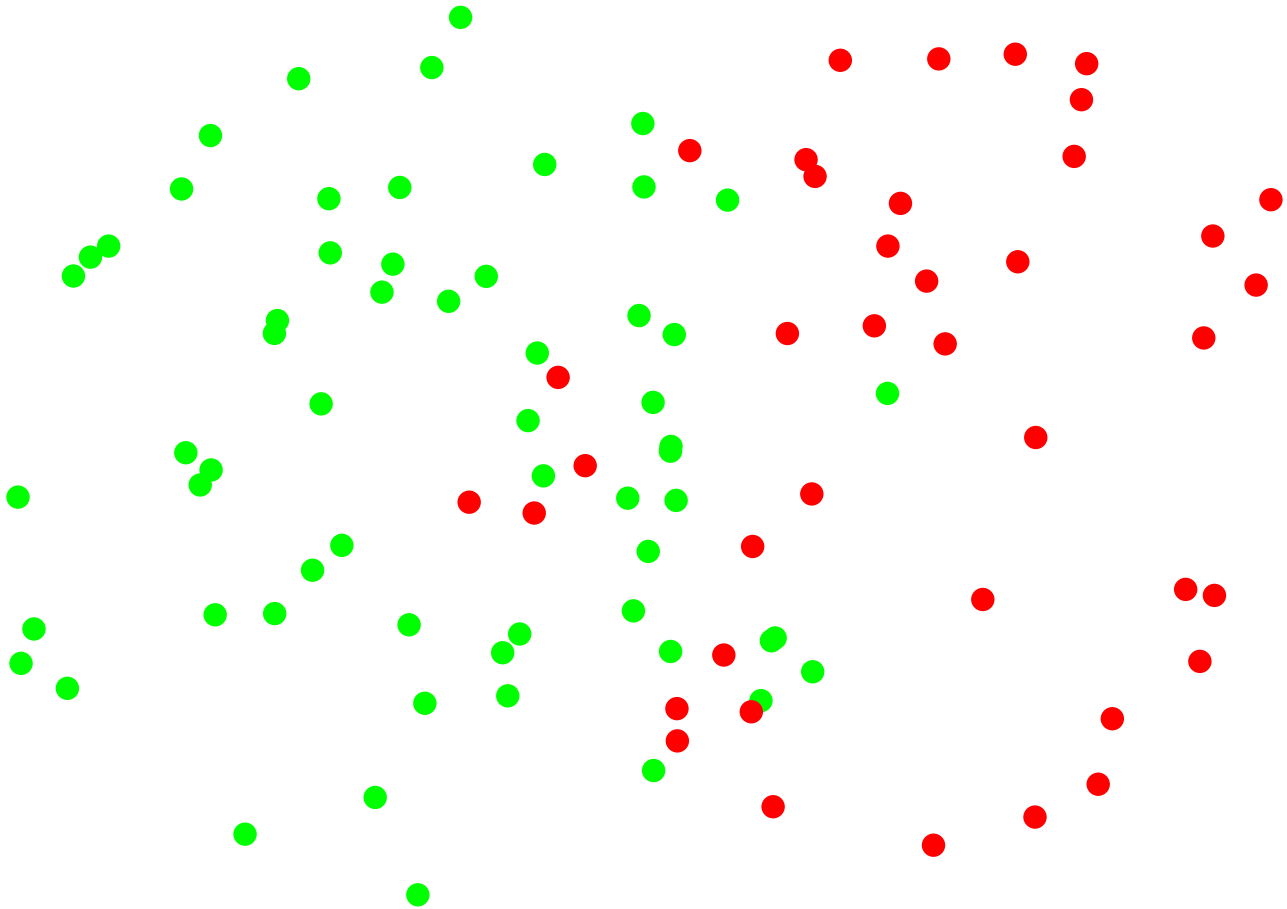
$$\mathcal{G}(n, \rho, a, b, c)$$



Stochastic block model [Holland-Laskey-Leinhardt '83]

$$\mathcal{G}(n, \rho, a, b, c)$$

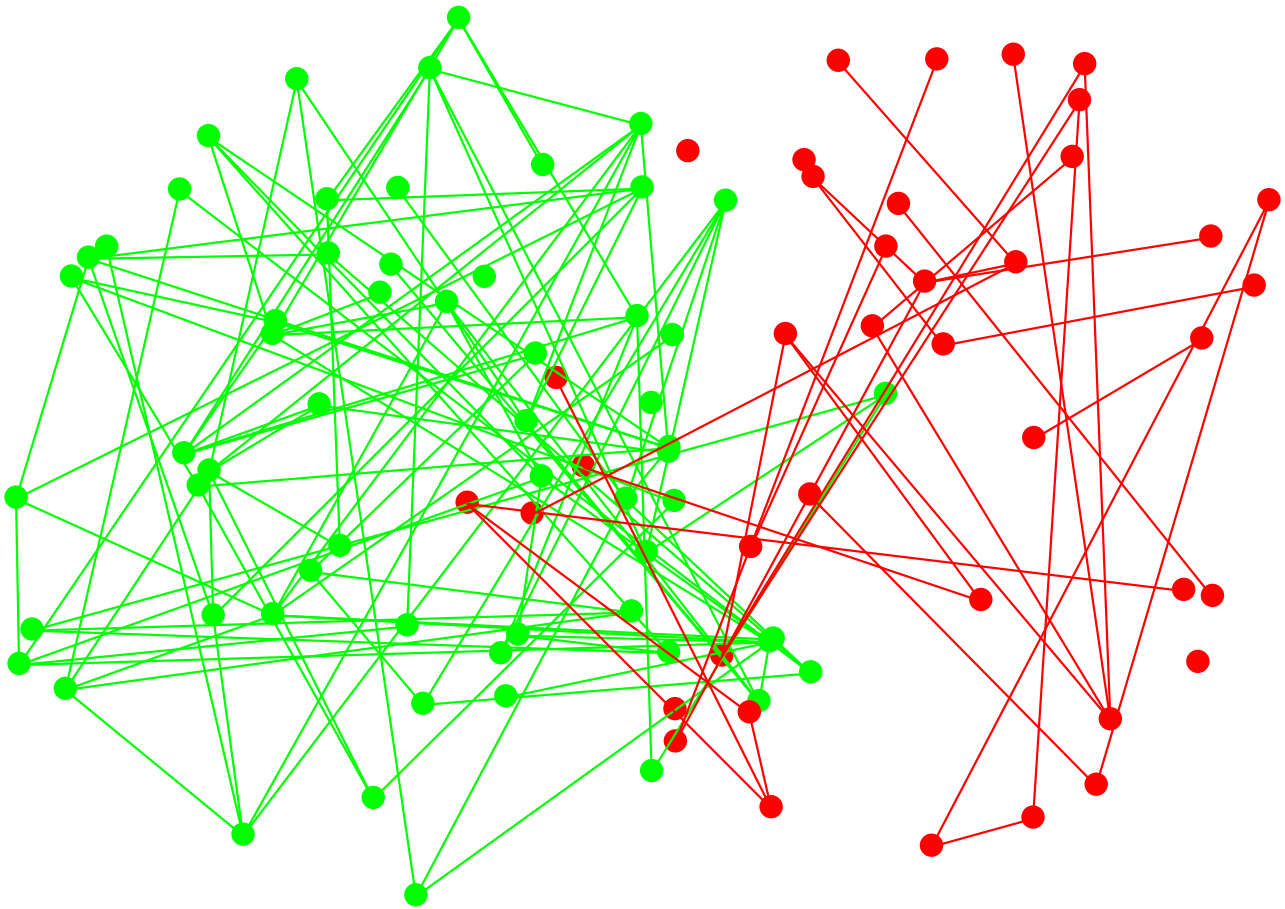
1. Color the vertices randomly: green ρ, red $\bar{\rho} \triangleq 1 - \rho$



Stochastic block model [Holland-Laskey-Leinhardt '83]

$$\mathcal{G}(n, \rho, a, b, c)$$

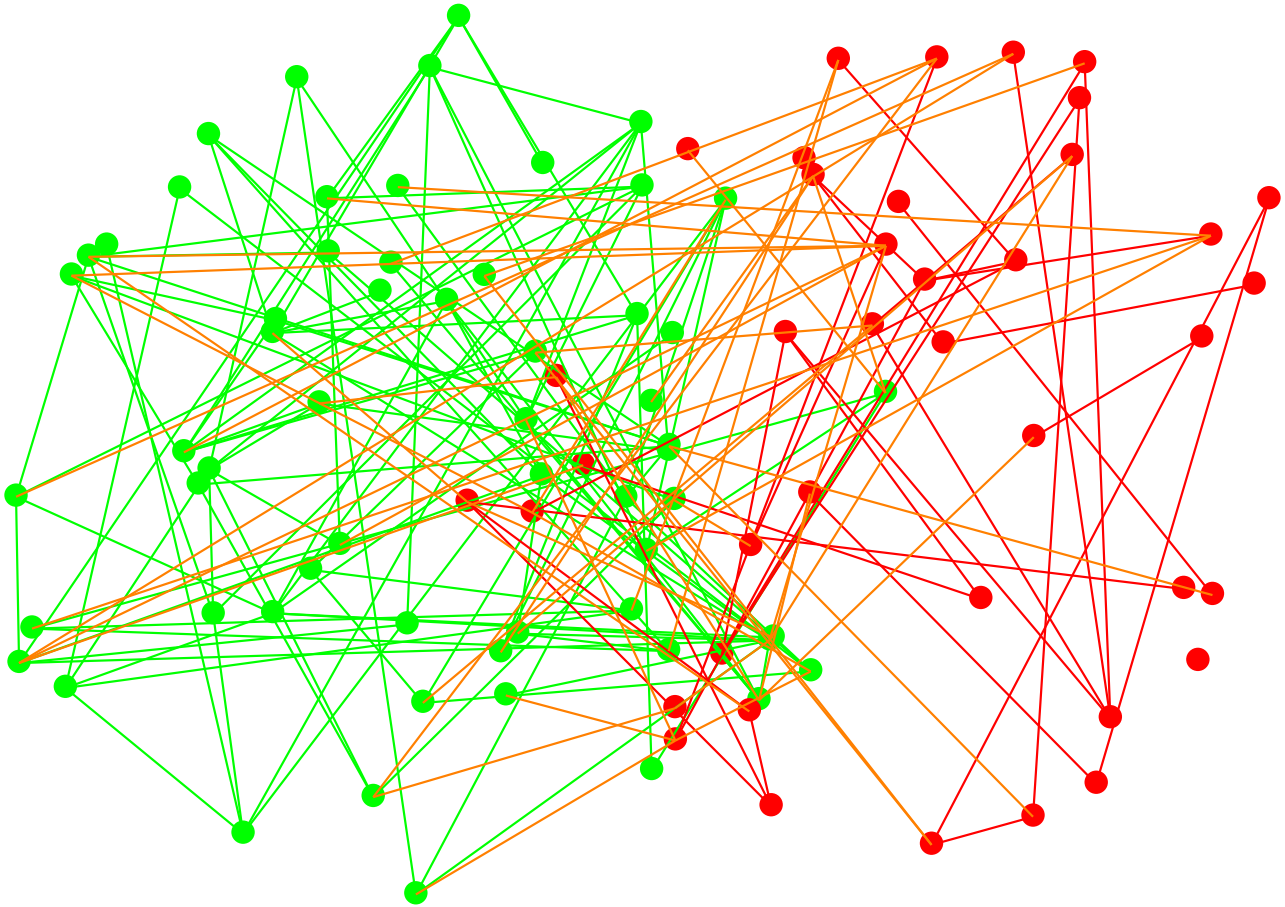
1. Color the vertices randomly: green ρ, red $\bar{\rho} \triangleq 1 - \rho$
2. For two nodes in green (red), add an edge w.p. $\frac{a}{n}$ ($\frac{c}{n}$)



Stochastic block model [Holland-Laskey-Leinhardt '83]

$$\mathcal{G}(n, \rho, a, b, c)$$

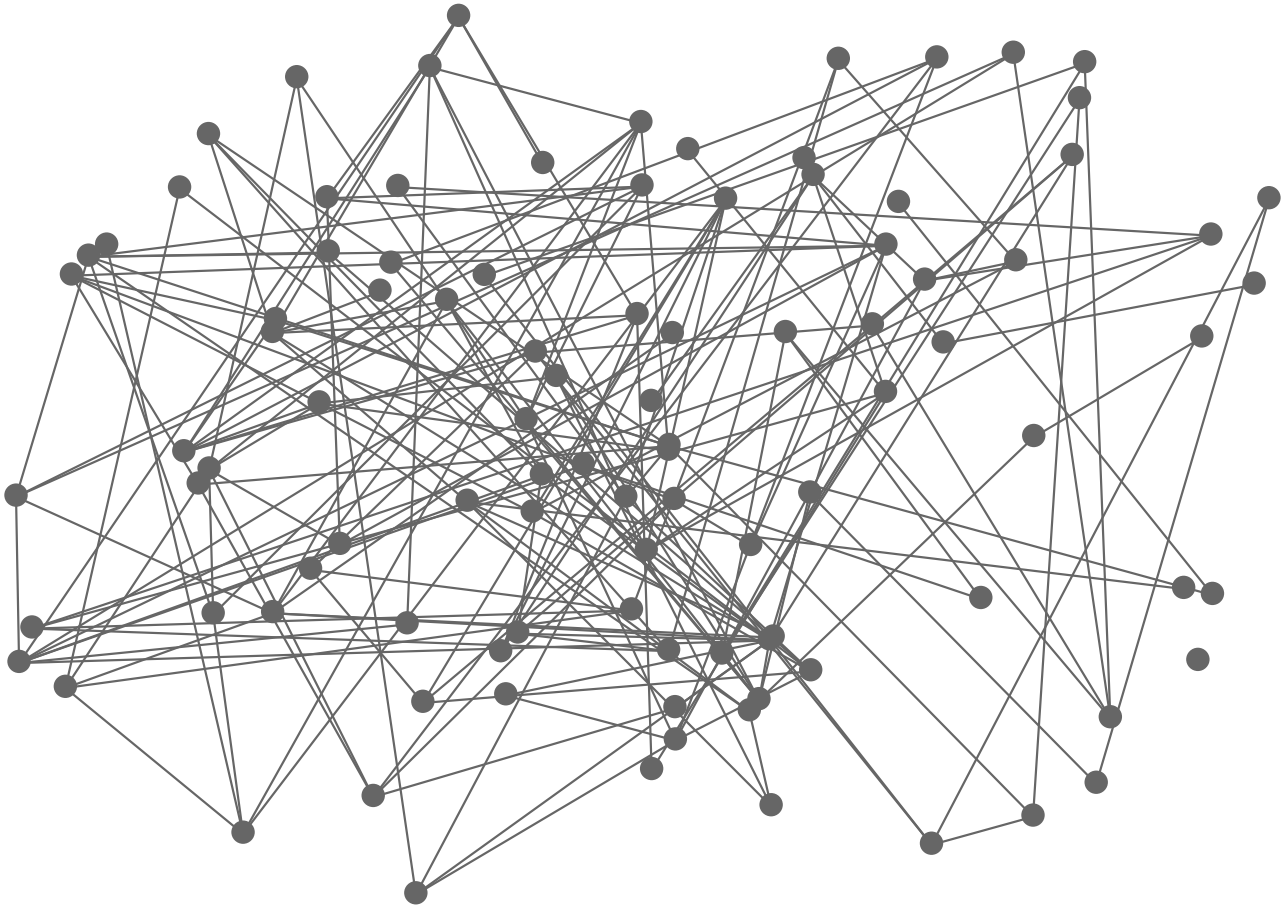
1. Color the vertices randomly: green ρ, red $\bar{\rho} \triangleq 1 - \rho$
2. For two nodes in green (red), add an edge w.p. $\frac{a}{n}$ ($\frac{c}{n}$)
3. For two nodes in different colors, add an edge w.p. $\frac{b}{n}$



Stochastic block model [Holland-Laskey-Leinhardt '83]

$$\mathcal{G}(n, \rho, a, b, c)$$

1. Color the vertices randomly: green ρ, red $\bar{\rho} \triangleq 1 - \rho$
2. For two nodes in green (red), add an edge w.p. $\frac{a}{n}$ ($\frac{c}{n}$)
3. For two nodes in different colors, add an edge w.p. $\frac{b}{n}$



Three recovery thresholds: $\rho = 1/2$ and $a = c$

ϵ: fraction of misclassified vertices

- Correlated recovery ($\epsilon < \frac{1}{2}$): [Mossel-Neeman-Sly '13] [Massoulié '13]

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

Three recovery thresholds: $\rho = 1/2$ and $a = c$

ϵ: fraction of misclassified vertices

- Correlated recovery ($\epsilon < \frac{1}{2}$): [Mossel-Neeman-Sly '13] [Massoulié '13]

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

- Weak recovery ($\epsilon = o(1)$): [Mossel-Neeman-Sly '14]

$$\frac{(a-b)^2}{(a+b)} \to \infty$$

Three recovery thresholds: $\rho = 1/2$ and $a = c$

ϵ: fraction of misclassified vertices

- Correlated recovery ($\epsilon < \frac{1}{2}$): [Mossel-Neeman-Sly '13] [Massoulié '13]

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

- Weak recovery ($\epsilon = o(1)$): [Mossel-Neeman-Sly '14]

$$\frac{(a-b)^2}{(a+b)} \to \infty$$

- Exact recovery ($\epsilon = 0$): [Abbe-Bandeira-Hall '14] [Mossel-Neeman-Sly '14]

$$\frac{(\sqrt{a}-\sqrt{b})^2}{\log n} > 2$$

Three recovery thresholds: $\rho = 1/2$ and $a = c$

ϵ: fraction of misclassified vertices

- Correlated recovery ($\epsilon < \frac{1}{2}$): [Mossel-Neeman-Sly '13] [Massoulié '13]

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

- Weak recovery ($\epsilon = o(1)$): [Mossel-Neeman-Sly '14]

$$\frac{(a-b)^2}{(a+b)} \to \infty$$

- Exact recovery ($\epsilon = 0$): [Abbe-Bandeira-Hall '14] [Mossel-Neeman-Sly '14]

$$\frac{(\sqrt{a}-\sqrt{b})^2}{\log n} > 2$$

Q: What is the threshold for a fixed $\epsilon \in (0, 1/2)$

Three recovery thresholds: $\rho = 1/2$ and $a = c$

ϵ: fraction of misclassified vertices

- Correlated recovery ($\epsilon < \frac{1}{2}$): [Mossel-Neeman-Sly '13] [Massoulié '13]

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

- Weak recovery ($\epsilon = o(1)$): [Mossel-Neeman-Sly '14]

$$\frac{(a-b)^2}{(a+b)} \to \infty$$

- Exact recovery ($\epsilon = 0$): [Abbe-Bandeira-Hall '14] [Mossel-Neeman-Sly '14]

$$\frac{(\sqrt{a}-\sqrt{b})^2}{\log n} > 2$$

Q: What is the threshold for a fixed $\epsilon \in (0, 1/2)$

Need to look at regime $\frac{(a-b)^2}{a+b} = \Theta(1)$ and get $\epsilon_{\min}(a, b)$

Outline of the talk

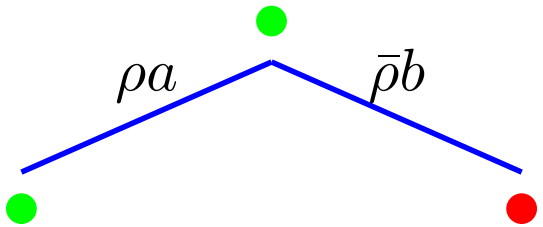
1. Tree reconstruction problems
2. Analysis of BP on tree via Gaussian approximations
3. Main results
4. Conjectures and open problems

Two-type branching process

1. Start with any given vertex and color it randomly: green ρ, Red $\bar{\rho}$

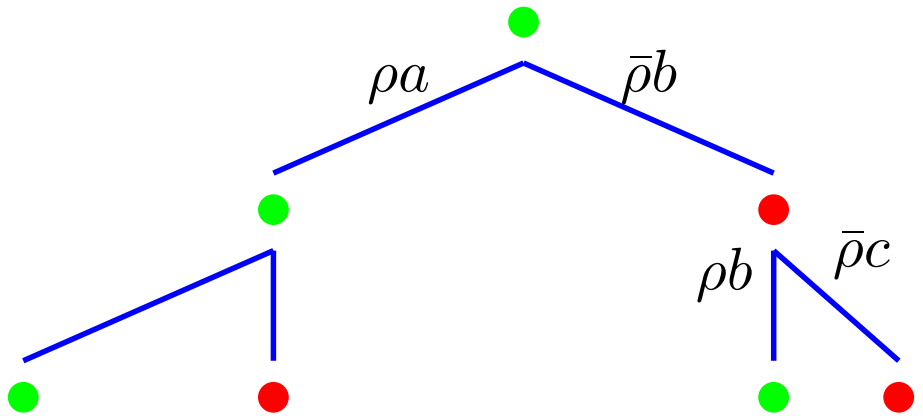
Two-type branching process

1. Start with any given vertex and color it randomly: green ρ, Red $\bar{\rho}$
2. Generate Poisson children



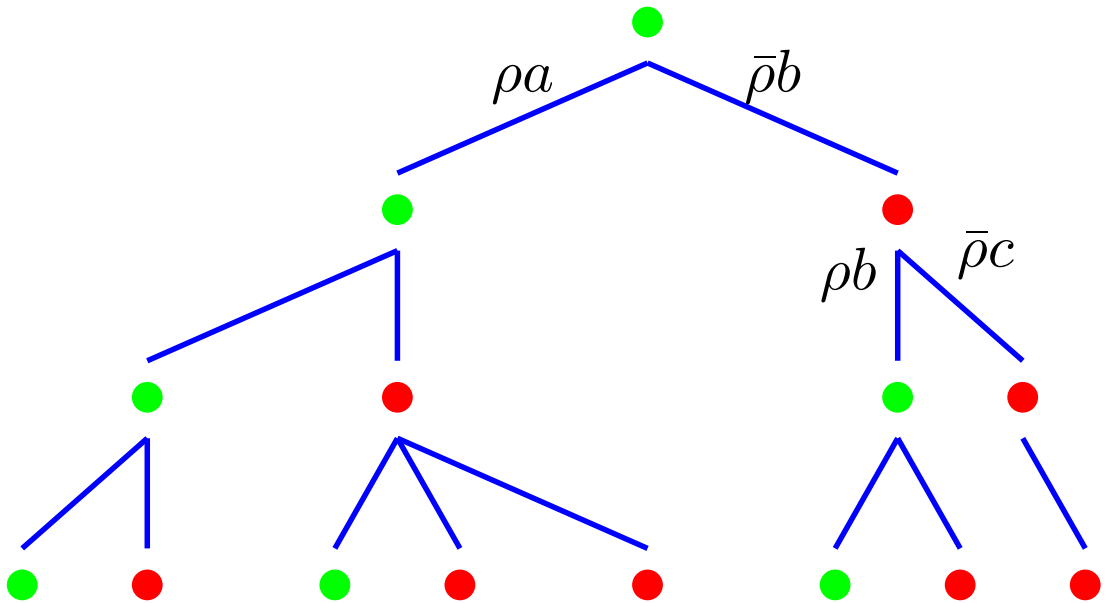
Two-type branching process

1. Start with any given vertex and color it randomly: green ρ, Red $\bar{\rho}$
2. Generate Poisson children
3. Repeat



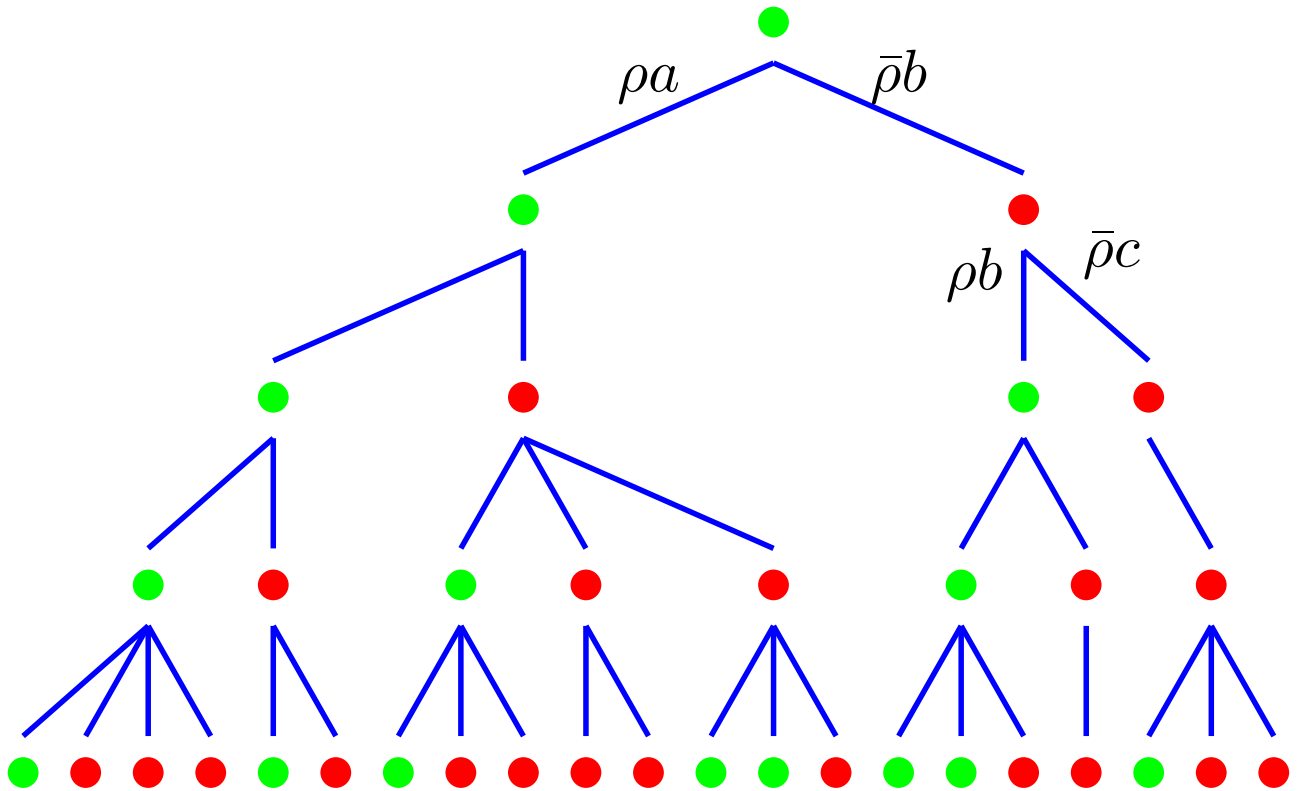
Two-type branching process

1. Start with any given vertex and color it randomly: green ρ, Red $\bar{\rho}$
2. Generate Poisson children
3. Repeat

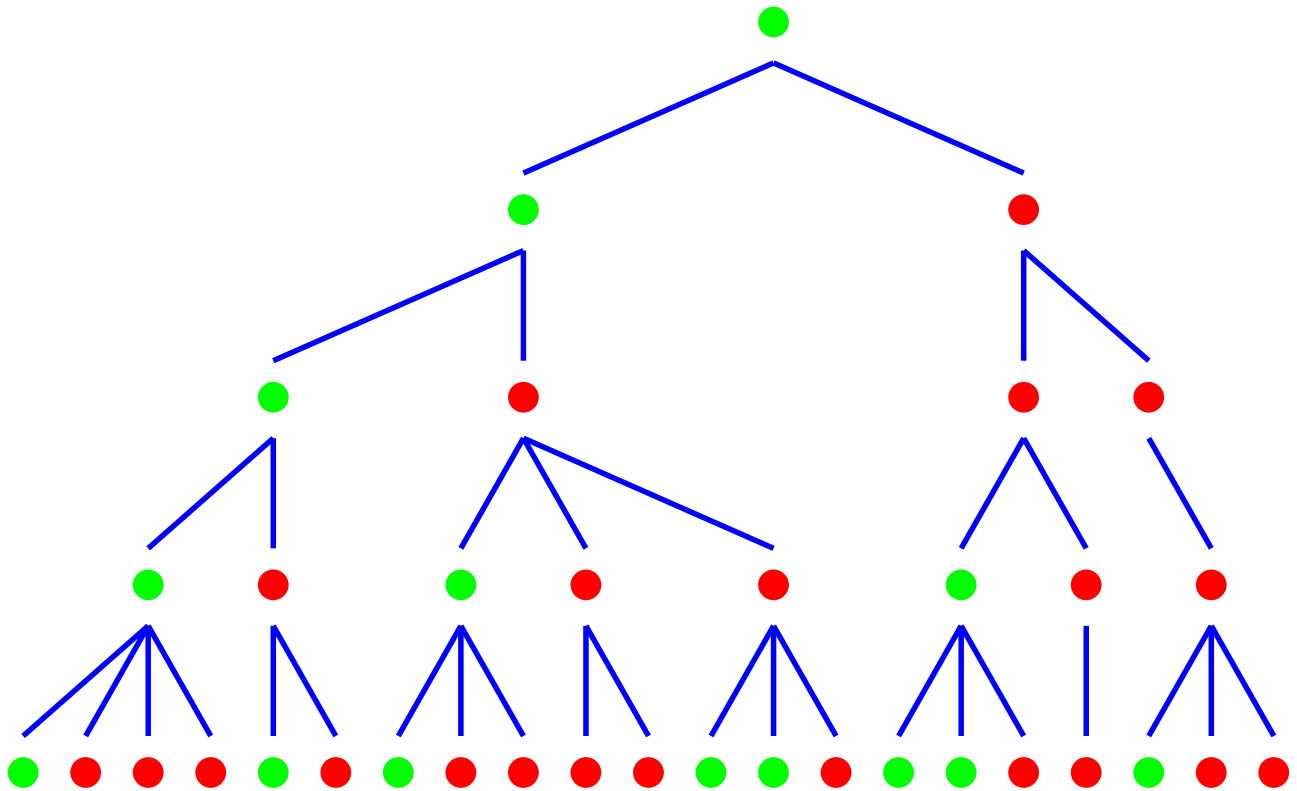


Two-type branching process

1. Start with any given vertex and color it randomly: green ρ, Red $\bar{\rho}$
2. Generate Poisson children
3. Repeat

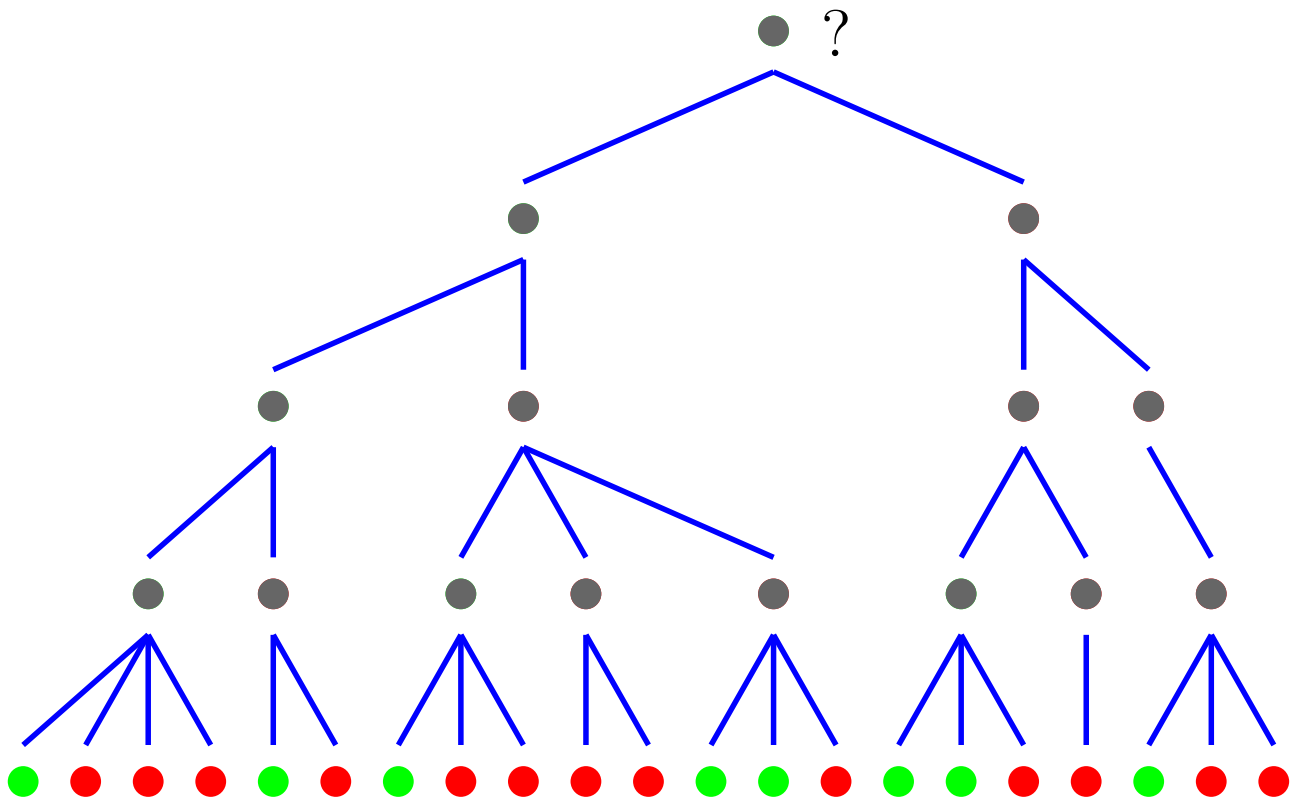


Reconstruction problems on trees



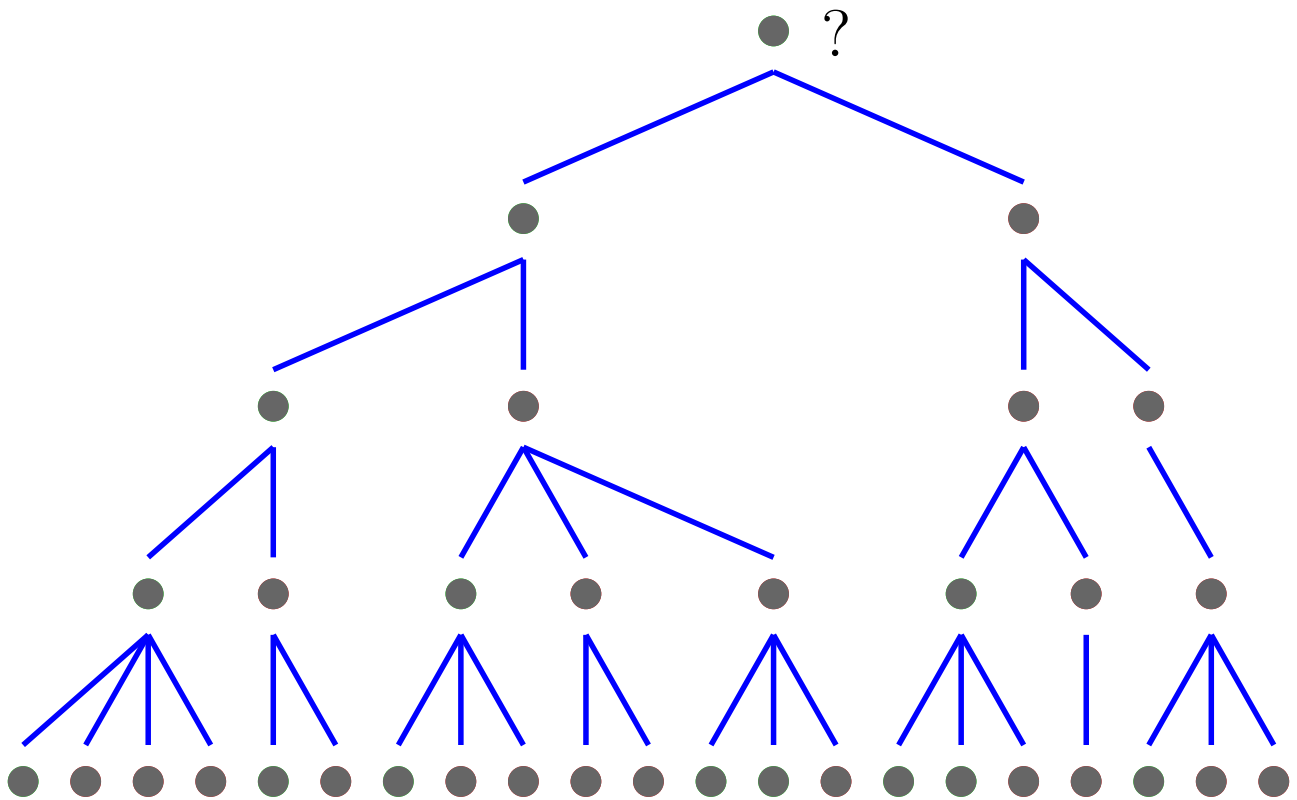
Reconstruction problems on trees

1. Exact colors at boundary [Evans-Kenyon-Peres-Schulman '00]



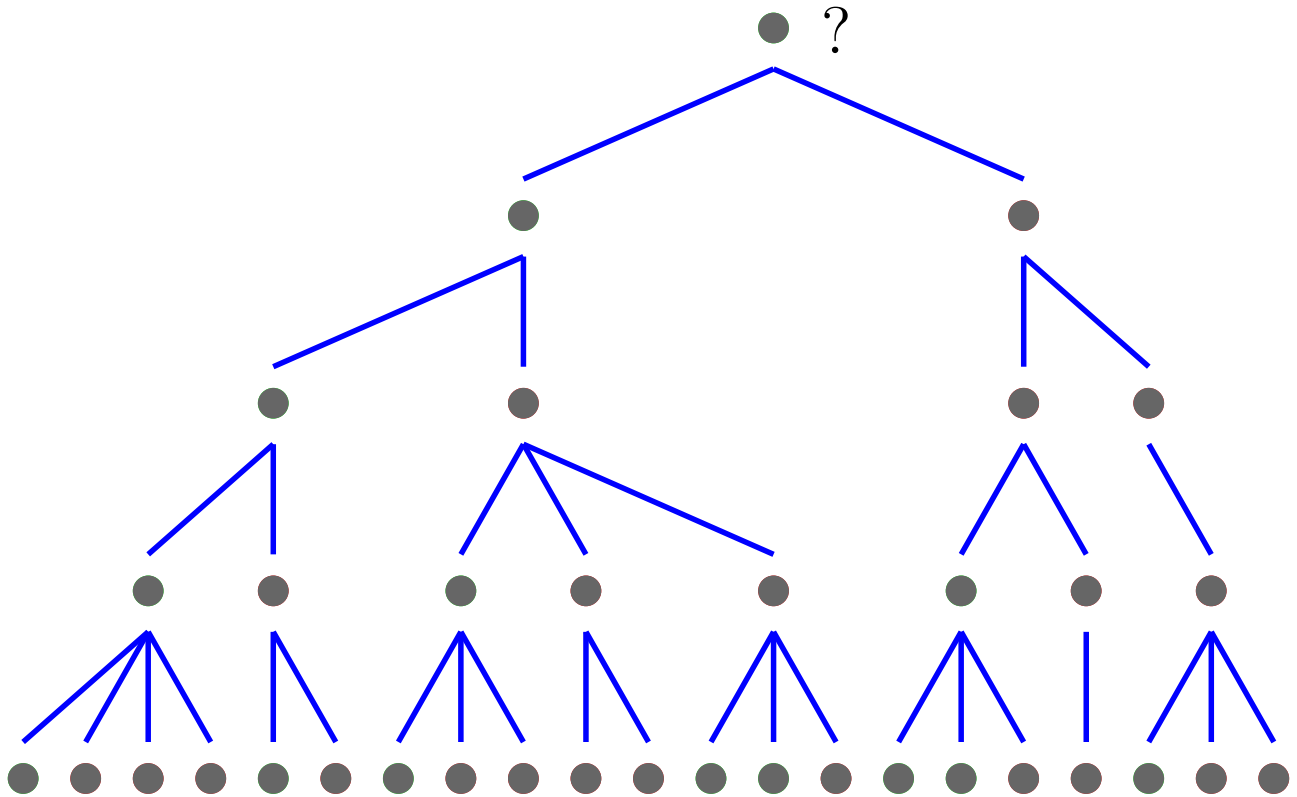
Reconstruction problems on trees

1. Exact colors at boundary [Evans-Kenyon-Peres-Schulman '00]
2. No colors at boundary



Reconstruction problems on trees

1. Exact colors at boundary [Evans-Kenyon-Peres-Schulman '00]
2. No colors at boundary



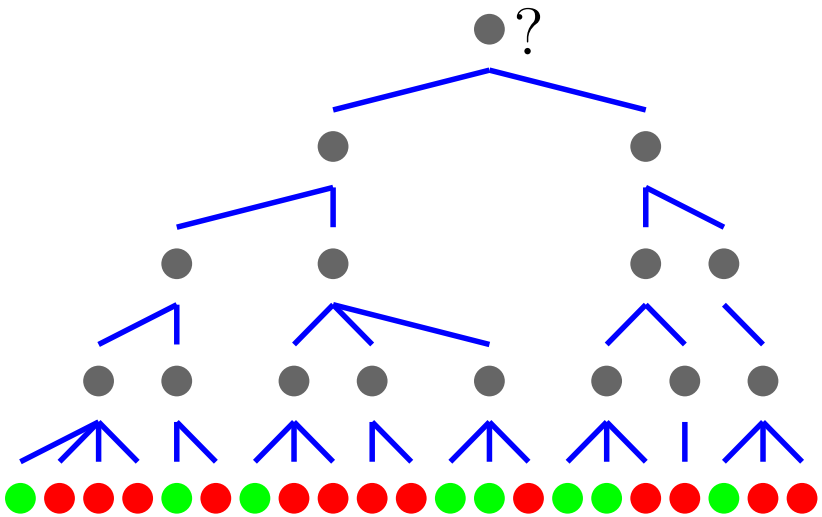
For both problems, MAP can be computed via belief propagation

Coupling between tree model and stochastic block model

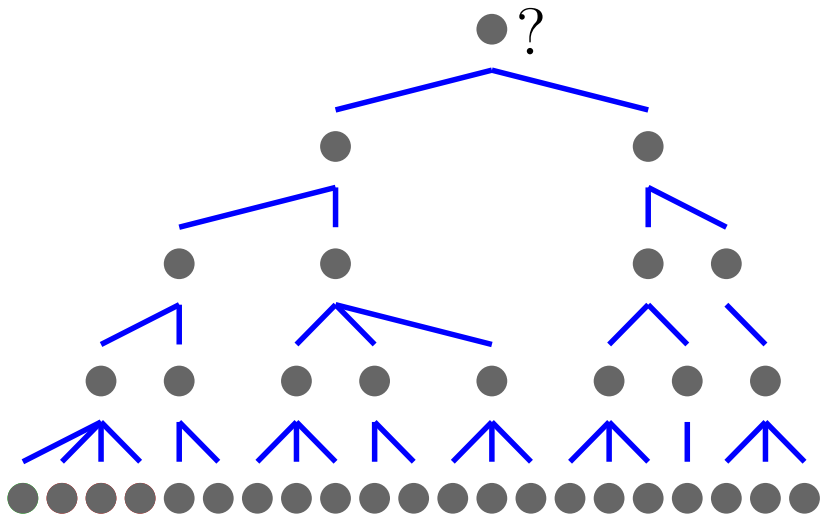
The t-local neighborhood of any given vertex can be coupled with the tree model with high probability, if $b^t = n^{o(1)}$

Coupling between tree model and stochastic block model

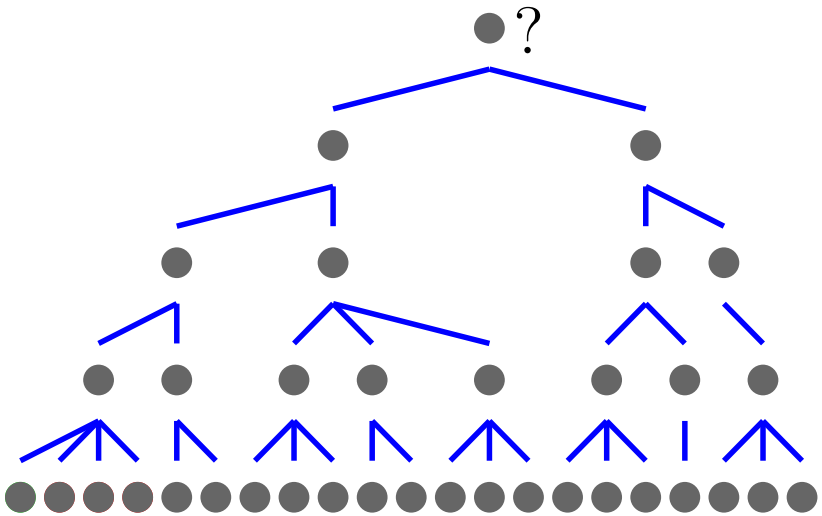
The t-local neighborhood of any given vertex can be coupled with the tree model with high probability, if $b^t = n^{o(1)}$



Provides lower bound on misclassification fraction under SBM



Provides upper bound on misclassification fraction under SBM



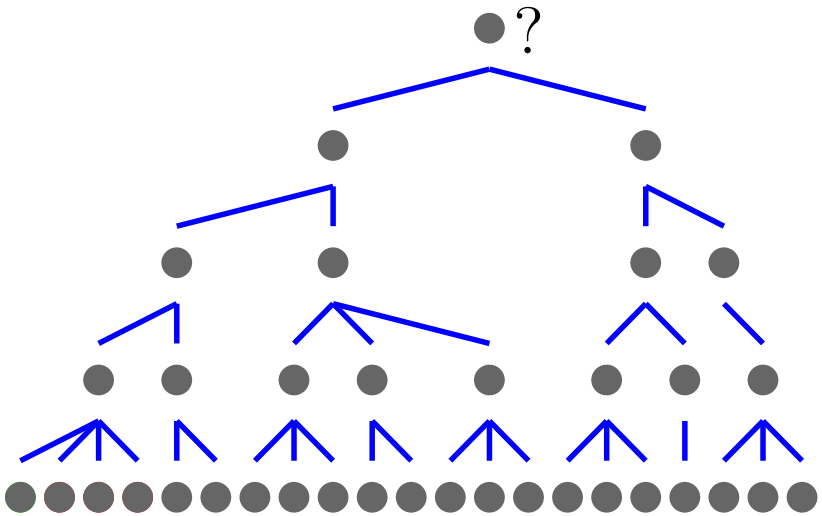
$$\hat{\tau}_{\mathrm{MAP}} = \mathbf{1}_{\{\Lambda_u^t \geq -\varphi\}}$$

$$\Lambda_u^t \triangleq \frac{1}{2} \log \frac{\mathbb{P}\left\{T_u^t | \tau_u = +\right\}}{\mathbb{P}\left\{T_u^t | \tau_u = -\right\}}$$

$$\varphi \triangleq \frac{1}{2} \log(\rho/\bar{\rho})$$

BP iterations: recursive formula of log likelihood ratios

$$\Lambda_{i \to \pi(i)}^{t+1} = \frac{-d_+ + d_-}{2} + \sum_{\ell \in \partial i} f(\Lambda_{\ell \to i}^t)$$



$$\hat{\tau}_{\mathrm{MAP}} = \mathbf{1}_{\{\Lambda_u^t \geq -\varphi\}}$$

$$\Lambda_u^t \triangleq \frac{1}{2} \log \frac{\mathbb{P}\left\{T_u^t | \tau_u = +\right\}}{\mathbb{P}\left\{T_u^t | \tau_u = -\right\}}$$

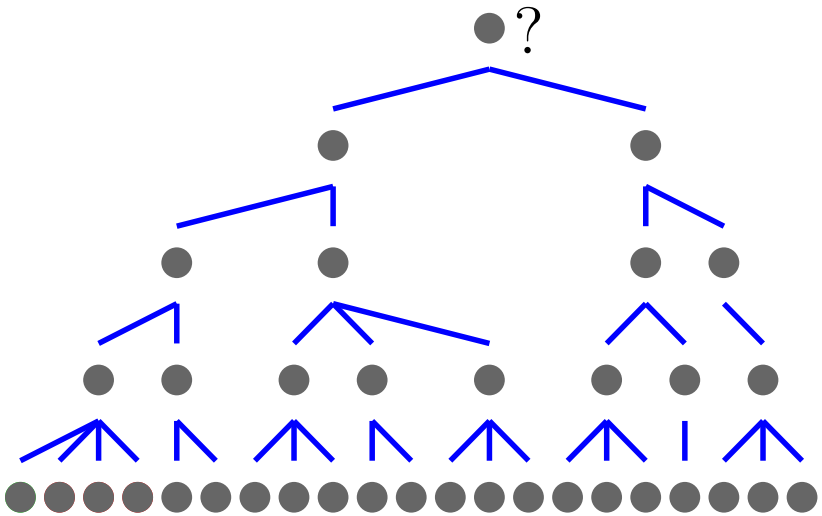
$$\varphi \triangleq \frac{1}{2} \log(\rho/\bar{\rho})$$

BP iterations: recursive formula of log likelihood ratios

$$\Lambda_{i \to \pi(i)}^{t+1} = \frac{-d_+ + d_-}{2} + \sum_{\ell \in \partial i} f(\Lambda_{\ell \to i}^t)$$

Remarks

- $f(x) = \frac{1}{2} \log \left(\frac{\mathrm{e}^{2x} \rho a + \bar{\rho} b}{\mathrm{e}^{2x} \rho b + \bar{\rho} c} \right)$
- No colors at boundary: $\Lambda_{\ell \to i}^0 \equiv 0$; Exact colors: $\Lambda_{\ell \to i}^0 = \pm \infty$



$$\hat{\tau}_{\mathrm{MAP}} = \mathbf{1}_{\{\Lambda_u^t \geq -\varphi\}}$$

$$\Lambda_u^t \triangleq \frac{1}{2}\log\frac{\mathbb{P}\left\{T_u^t|\tau_u=+\right\}}{\mathbb{P}\left\{T_u^t|\tau_u=-\right\}}$$

$$\varphi \triangleq \frac{1}{2}\log(\rho/\bar{\rho})$$

BP iterations: recursive formula of log likelihood ratios

$$\Lambda_{i\to\pi(i)}^{t+1} = \frac{-d_+ + d_-}{2} + \sum_{\ell\in\partial i} f(\Lambda_{\ell\to i}^t)$$

Remarks

- $f(x) = \frac{1}{2}\log\left(\frac{\mathrm{e}^{2x}\rho a + \bar{\rho}b}{\mathrm{e}^{2x}\rho b + \bar{\rho}c}\right)$
- No colors at boundary: $\Lambda_{\ell\to i}^0 \equiv 0$; Exact colors: $\Lambda_{\ell\to i}^0 = \pm\infty$
- Recursion for the distribution of Λ's: Let $Z_\pm^t \stackrel{d}{=} (\Lambda_u^t|\tau_u = \pm)$

$$\Phi : \mathcal{L}(Z_+^t) \to \mathcal{L}(Z_+^{t+1})$$

Analysis of BP via Gaussian approximations

Assume ρ fixed with $\varphi \triangleq \frac{1}{2}\log(\rho/\bar{\rho})$ and

$$b \to \infty, \qquad b = n^{o(1)}, \qquad \mu = \frac{a-b}{\sqrt{b}}, \qquad \nu = \frac{c-b}{\sqrt{b}}$$

BP messages are approximately Gaussian

$\mathcal{L}(Z_{\pm}^t) \Rightarrow \mathcal{N}(\pm v_t, v_t)$, where $Z \sim \mathcal{N}(0,1)$,

$$v_{t+1} = \frac{\rho(\mu-\nu)^2}{8} + \frac{(1-2\rho)\nu^2}{4} + \frac{\rho(\mu+\nu)^2}{8}\mathbb{E}\left[\tanh(v_t + \sqrt{v_t}Z + \varphi)\right].$$

Analysis of BP via Gaussian approximations

Assume ρ fixed with $\varphi \triangleq \frac{1}{2}\log(\rho/\bar{\rho})$ and

$$b \to \infty, \qquad b = n^{o(1)}, \qquad \mu = \frac{a-b}{\sqrt{b}}, \qquad \nu = \frac{c-b}{\sqrt{b}}$$

BP messages are approximately Gaussian

$\mathcal{L}(Z_{\pm}^t) \Rightarrow \mathcal{N}(\pm v_t, v_t)$, where $Z \sim \mathcal{N}(0,1)$,

$$v_{t+1} = \frac{\rho(\mu-\nu)^2}{8} + \frac{(1-2\rho)\nu^2}{4} + \frac{\rho(\mu+\nu)^2}{8}\mathbb{E}\left[\tanh(v_t + \sqrt{v_t}Z + \varphi)\right].$$

Remarks

- If $\nu = 0$ (single community), it is derived in [Montanari '15]
- No colors at boundary: $v_0 = 0$; Exact colors at boundary: $v_0 = \infty$
- Upper and lower bounds match if v_t converges to the same fixed point

Minimum fraction of misclassified vertices: $\rho = 1/2$

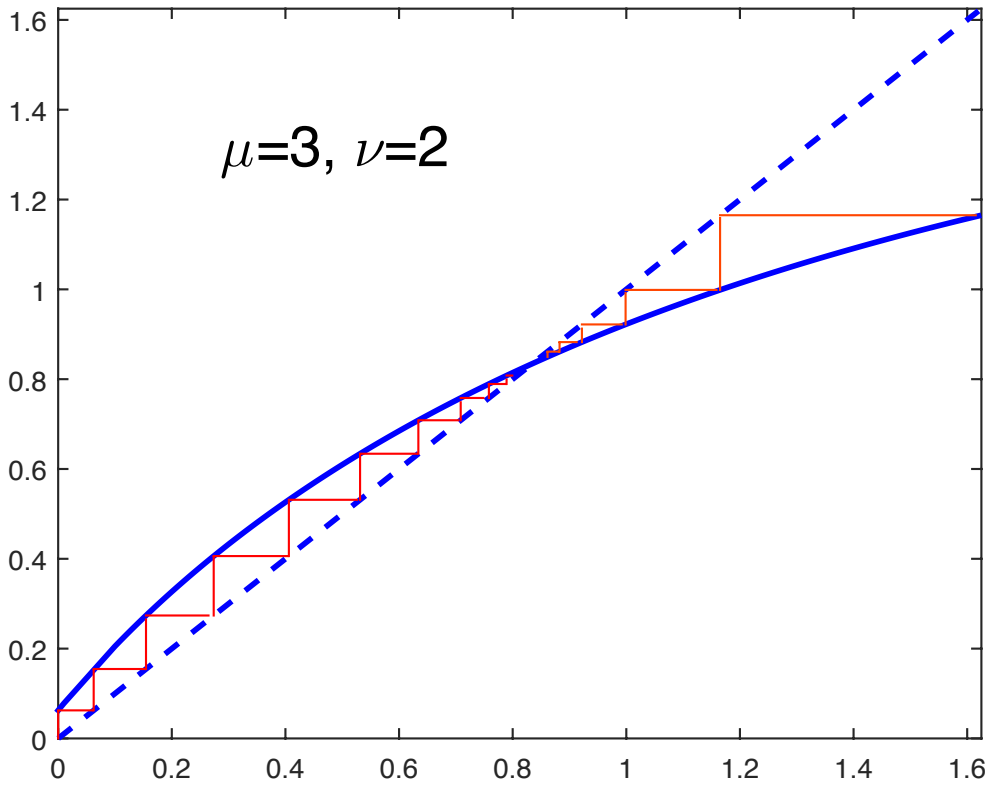
Theorem

The minimum fraction of misclassified vertices on average $\epsilon_{\min} = Q(\sqrt{v^*})$, *where* v^* *is the unique stable fixed point of*

$$v = \frac{(\mu-\nu)^2}{16} + \frac{(\mu+\nu)^2}{16}\mathbb{E}\left[\tanh(v+\sqrt{v}Z)\right]$$

Degree-correlated: $\mu \neq \nu$

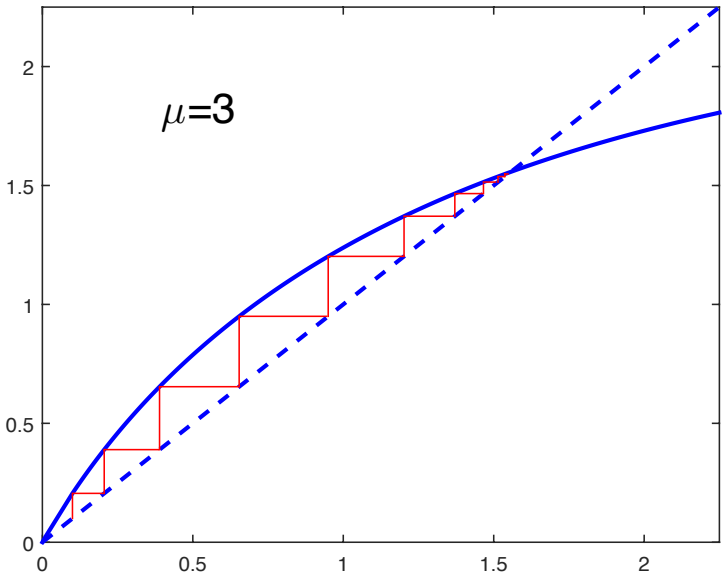
$$v = \frac{(\mu-\nu)^2}{16} + \frac{(\mu+\nu)^2}{16}\mathbb{E}\left[\tanh(v+\sqrt{v}Z)\right] \triangleq h(v)$$

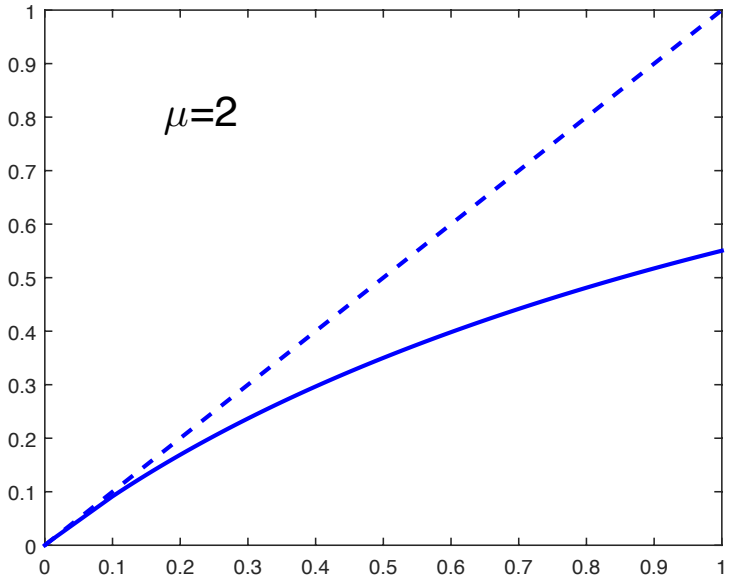


- Upper bound: v_t with $v_0 = 0$ is attained by BP with zero init.
- Lower bound: v_t with $v_0 = \infty$ is attained by BP with exact init.

Degree-uncorrelated: $\mu = \nu$

$$v = \frac{\mu^2}{4}\mathbb{E}\left[\tanh(v + \sqrt{v}Z)\right] \triangleq h(v)$$





- Phase transition at $\mu = \frac{(a-b)^2}{a+b} = 2$ (correlated recovery threshold)
- Local algorithms get stuck at the trivial zero fixed point
- Local BP + correlated init. is optimal [Mossel-Neeman-Sly '14]

Key of the proofs in case $\rho = 1/2$

To show $v \to \mathbb{E}\left[\tanh(v + \sqrt{v}Z\right]$ is concave, also proved by [Deshpande-Abbe-Montanari '15]

1. Compute first derivative using integration by parts
2. Use properties of symmetric random variables [Montanari '05] + first-order stochastic dominance

Key of the proofs in case $\rho = 1/2$

To show $v \to \mathbb{E}\left[\tanh(v + \sqrt{v}Z\right]$ is concave, also proved by [Deshpande-Abbe-Montanari '15]

1. Compute first derivative using integration by parts
2. Use properties of symmetric random variables [Montanari '05] + first-order stochastic dominance

Question

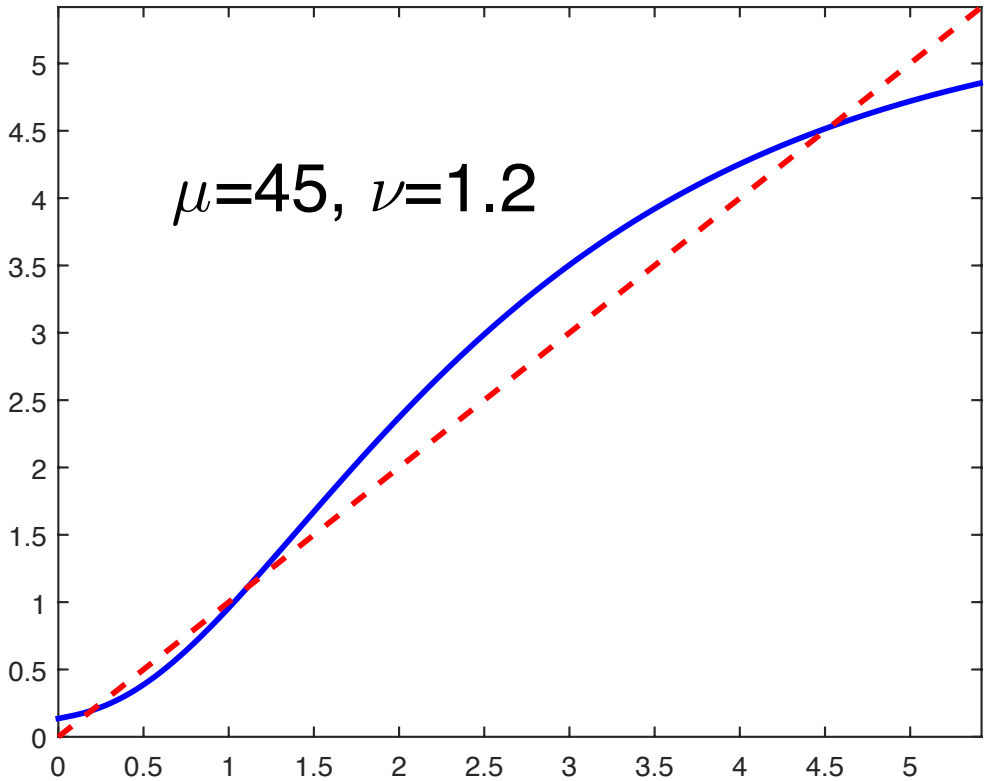
Is $\rho = 1/2$ special?

Open problems

- Numerical experiments show $v \to \mathbb{E}\left[\tanh(v + \sqrt{v}Z)\right]$ is concave for $\rho \geq 0.2 \Rightarrow$ local BP is optimal

Open problems

- Numerical experiments show $v \to \mathbb{E}\left[\tanh(v + \sqrt{v}Z)\right]$ is concave for $\rho \geq 0.2 \Rightarrow$ local BP is optimal
- If $\rho = 0.01$, $v = \frac{(\mu-\nu)^2}{16} + \frac{(\mu+\nu)^2}{16}\mathbb{E}\left[\tanh(v + \sqrt{v}Z)\right]$ has multiple fixed points $\Rightarrow$ local BP might be strictly suboptimal



Conclusions

- For two equal-sized communities: $\epsilon_{\min} = Q(\sqrt{v^*})$, where v^* is the unique stable fixed point of

$$v = \frac{(\mu-\nu)^2}{16} + \frac{(\mu+\nu)^2}{16}\mathbb{E}\left[\tanh(v+\sqrt{v}Z)\right]$$

Conclusions

- For two equal-sized communities: $\epsilon_{\min} = Q(\sqrt{v^*})$, where v^* is the unique stable fixed point of

$$v = \frac{(\mu-\nu)^2}{16} + \frac{(\mu+\nu)^2}{16}\mathbb{E}\left[\tanh(v+\sqrt{v}Z)\right]$$

- Conjecture: Local BP is still optimal for $\rho \geq 0.2$, but it is strictly suboptimal for $\rho \leq 0.01$

Conclusions

- For two equal-sized communities: $\epsilon_{\min} = Q(\sqrt{v^*})$, where v^* is the unique stable fixed point of

$$v = \frac{(\mu-\nu)^2}{16} + \frac{(\mu+\nu)^2}{16}\mathbb{E}\left[\tanh(v+\sqrt{v}Z)\right]$$

- Conjecture: Local BP is still optimal for $\rho \geq 0.2$, but it is strictly suboptimal for $\rho \leq 0.01$

Reference:

E. Mossel and J. Xu, "Density evolution in the degree-correlated stochastic block model," *arXiv:1509.03281*, Sept. 2015.