Find Your Place: Simple Distributed Algorithms for Community Detection

Emanuele Natale[◇]

joint work with

Luca Becchetti[†], Andrea Clementi[⋆],

Francesco Pasquale[⋆] and Luca Trevisan[*]

[†] SAPIENZA UNIVERSITÀ DI ROMA

[⋆] UNIVERSITA' DEGLI STUDI DI ROMA Tor Vergata

[*] Berkeley UNIVERSITY OF CALIFORNIA

[◇] mpi max planck institut informatik

ACM-SIAM Symposium on Discrete Algorithms

16-19 January 2017 - Barcellona, Spain

Dynamics

Dynamics: For every graph, agent and round, states are updated according to fixed rule of current state and symmetric function of states of neighbors.

Dynamics

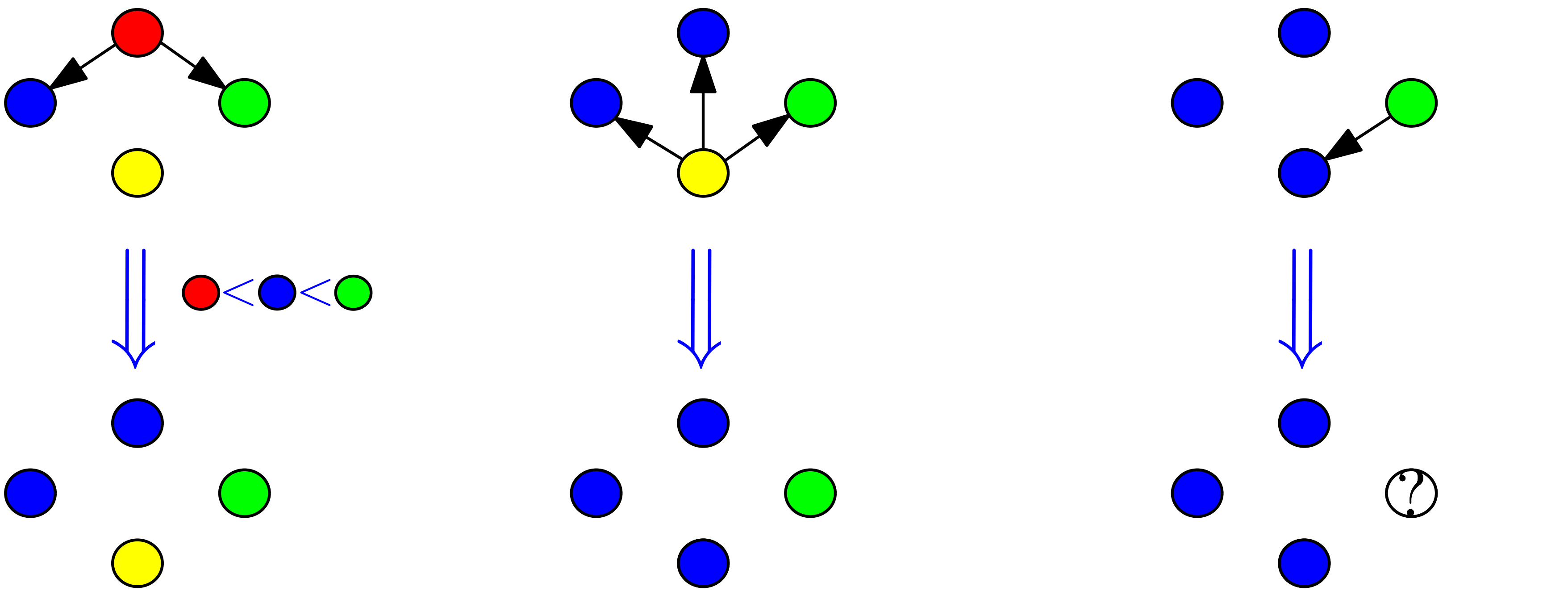
Dynamics: For every graph, agent and round, states are updated according to fixed rule of current state and symmetric function of states of neighbors.

Examples of Dynamics:

3-Median dyn. [Doerr et al. '11]

3-Majority dyn. [Becchetti et al. '14, '16]

Undecided-state dyn. [Becchetti et al. '15]

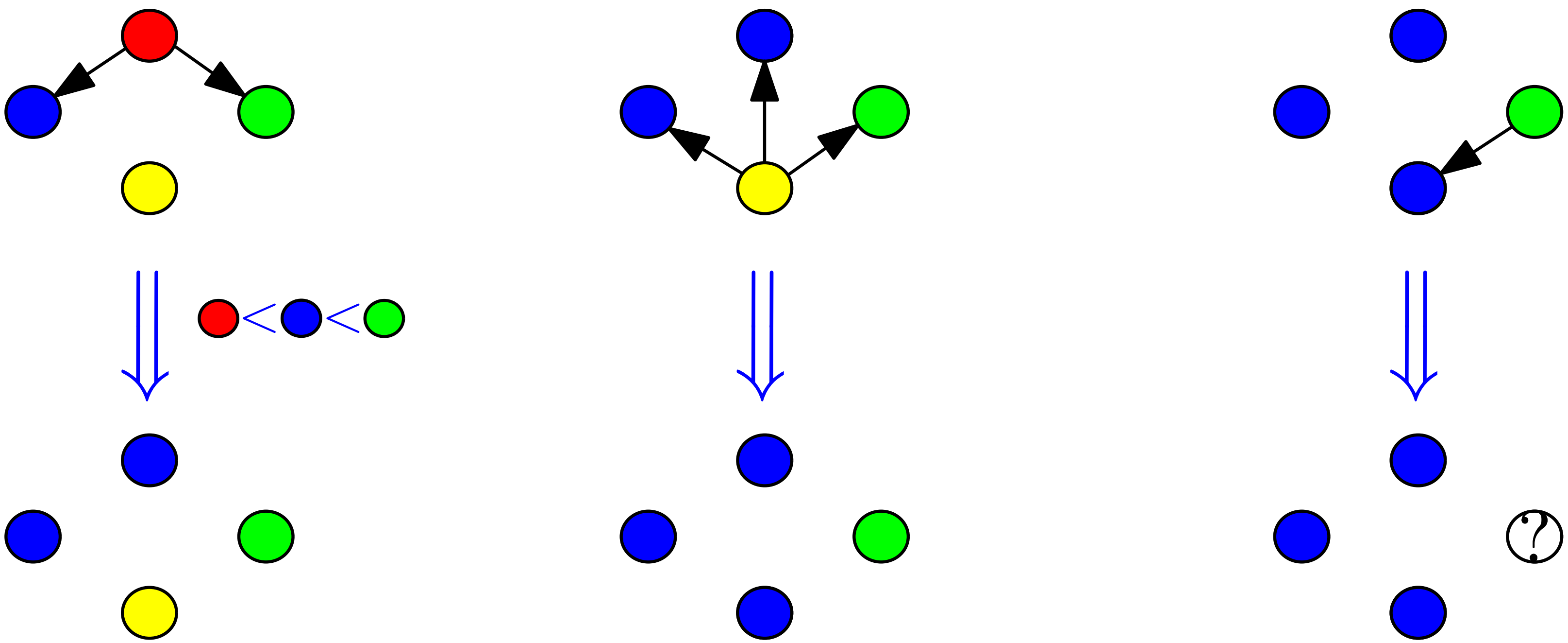


Dynamics

Dynamics: For every graph, agent and round, states are updated according to fixed rule of current state and symmetric function of states of neighbors.

Examples of Dynamics:

3-Median dyn. [Doerr et al. '11] | 3-Majority dyn. [Becchetti et al. '14, '16] | Undecided-state dyn. [Becchetti et al. '15]



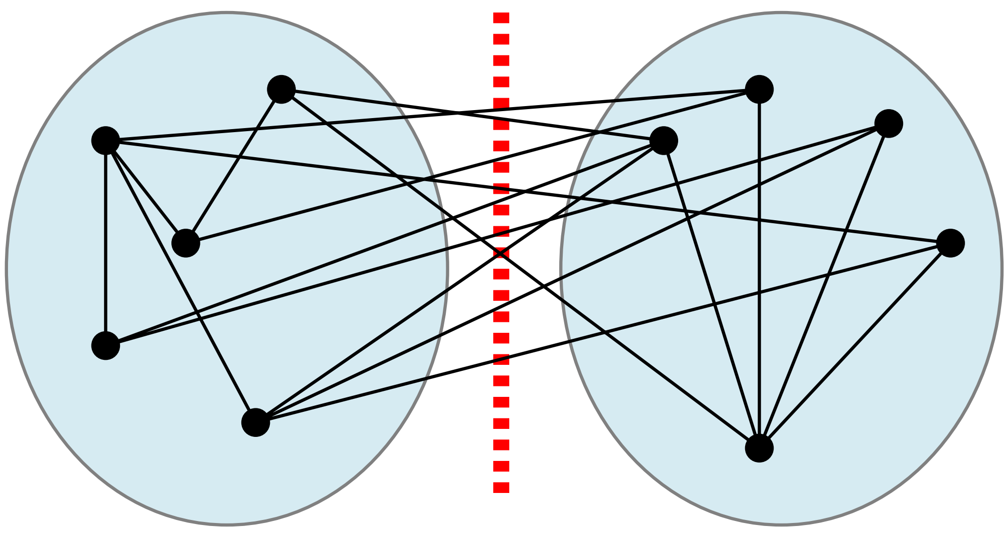
Can dynamics solve a problem non-trivial in centralized setting?

Community Detection as Minimum Bisection

Minimum Bisection Problem.

Input: a graph G with n nodes.

Output: $S = \arg\min_{\substack{S \subset V \\ |S| = n/2}} E(S, V - S)$.

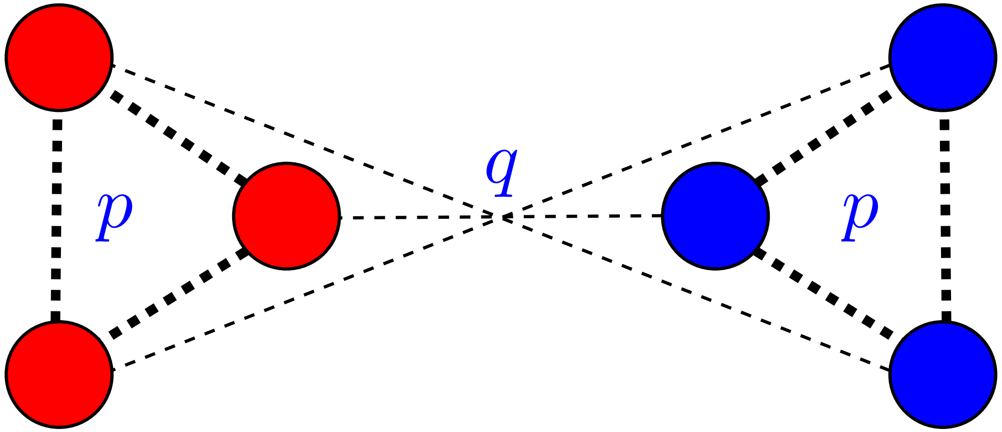


[Garey, Johnson, Stockmeyer '76]:

Min-Bisection is *NP-Complete*.

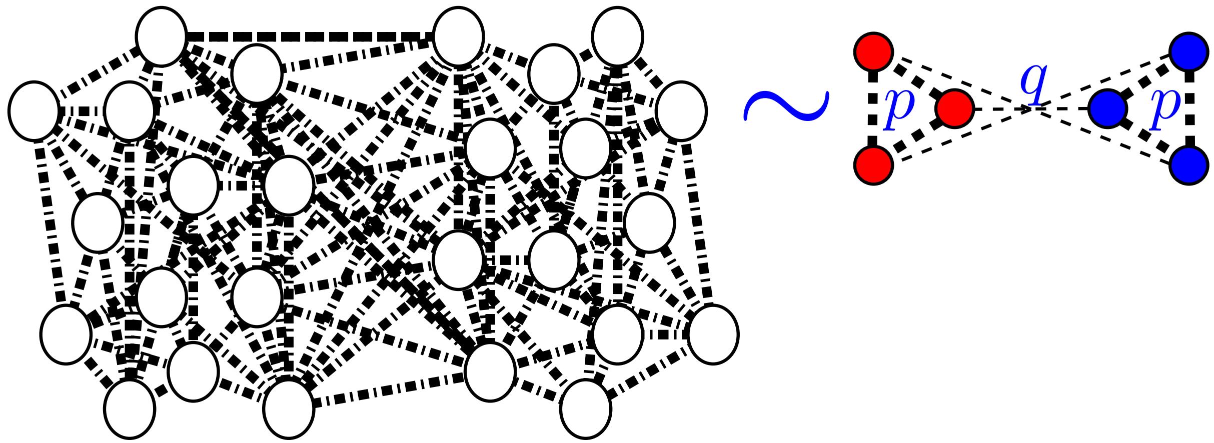
The Stochastic Block Model

Stochastic Block Model (SBM). Two "communities" of equal size V_1 and V_2, each edge inside a community included with probability p, each edge across communities included with probability $q < p$.

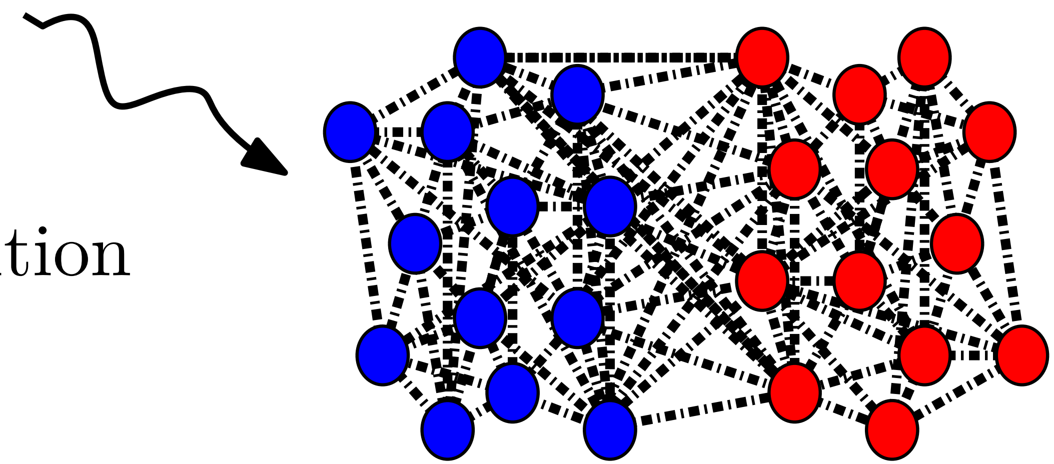


The Stochastic Block Model

Reconstruction problem:
Given graph generated by SBM,



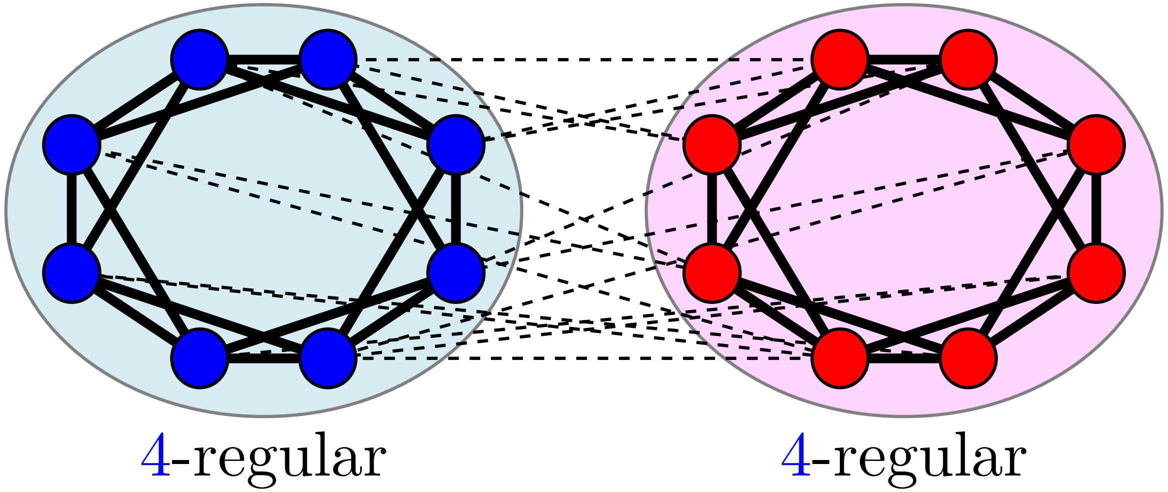
find original partition



Regular Stochastic Block Model

Regular SBM (RSBM) [Brito et al. SODA'16]. A graph $G = (V_1 \dot{\bigcup} V_2, E)$ s.t.

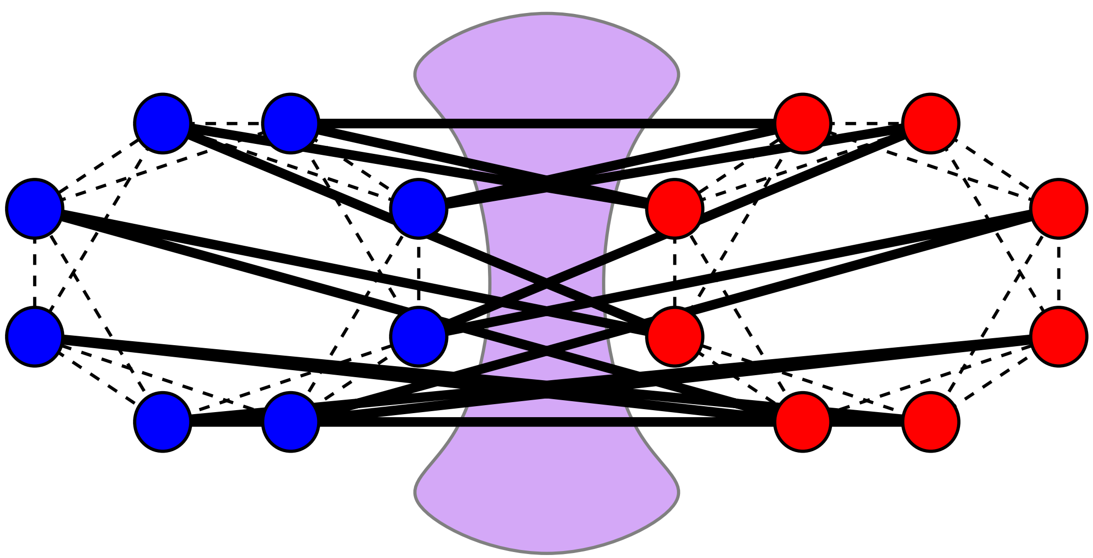
- $|V_1| = |V_2|$,
- $G\big|_{V_1}, G\big|_{V_2} \sim$ random a-regular graphs
- $G\big|_{E(V_1,V_2)} \sim$ random b-regular bipartite graph.



Regular Stochastic Block Model

Regular SBM (RSBM) [Brito et al. SODA'16]. A graph $G = (V_1 \dot{\bigcup} V_2, E)$ s.t.

- $|V_1| = |V_2|$,
- $G\big|_{V_1}, G\big|_{V_2}$ ~ random a-regular graphs
- $G\big|_{E(V_1,V_2)}$ ~ random b-regular bipartite graph.

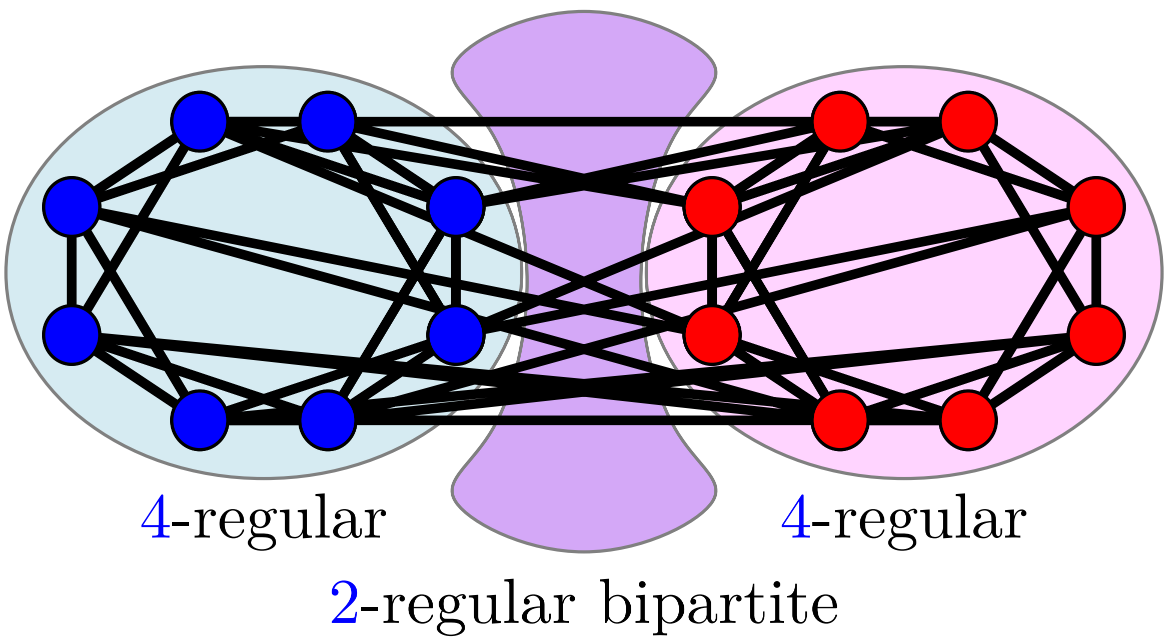


2-regular bipartite

Regular Stochastic Block Model

Regular SBM (RSBM) [Brito et al. SODA'16]. A graph $G = (V_1 \dot{\bigcup} V_2, E)$ s.t.

- $|V_1| = |V_2|$,
- $G\big|_{V_1}, G\big|_{V_2} \sim$ random a-regular graphs
- $G\big|_{E(V_1,V_2)} \sim$ random b-regular bipartite graph.



When is Reconstruction Possible?

[Decelle, Massoulie, Mossel, Brito, Abbe et al.]:
Reconstruction is possible iff

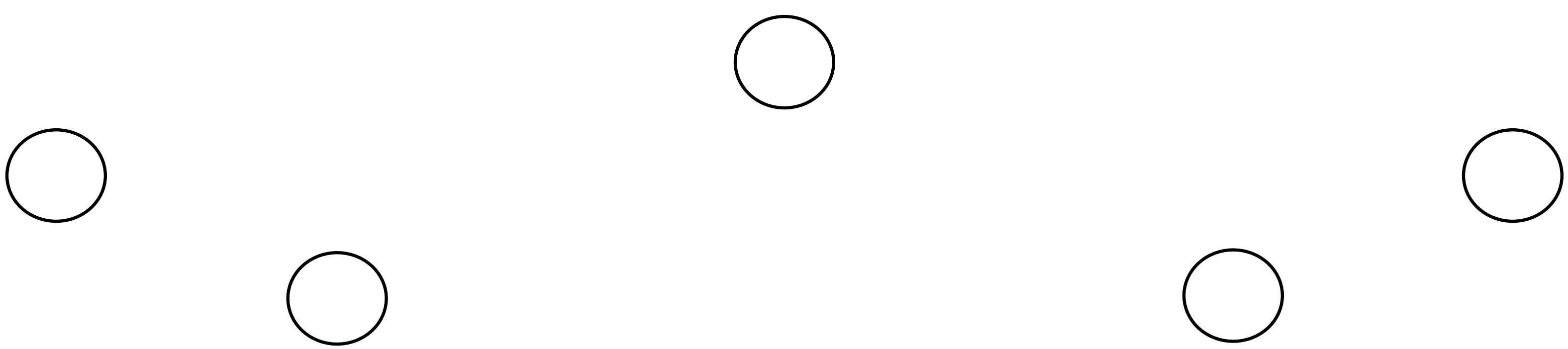
- $a - b > 2\sqrt{a+b}$ in SBM (weak)
- $a - b > 2(\sqrt{a} - \sqrt{b})\sqrt{b} + 2\log n$ in SBM (strong)
- $a - b > 2\sqrt{a+b-1}$ in Regular SBM (strong)

Upper bounds obtained by linearizations of *Belief Propagation*, advanced spectral methods (power and Lanczos method), SDP.

The Average Dynamics

Al nodes at the same time:

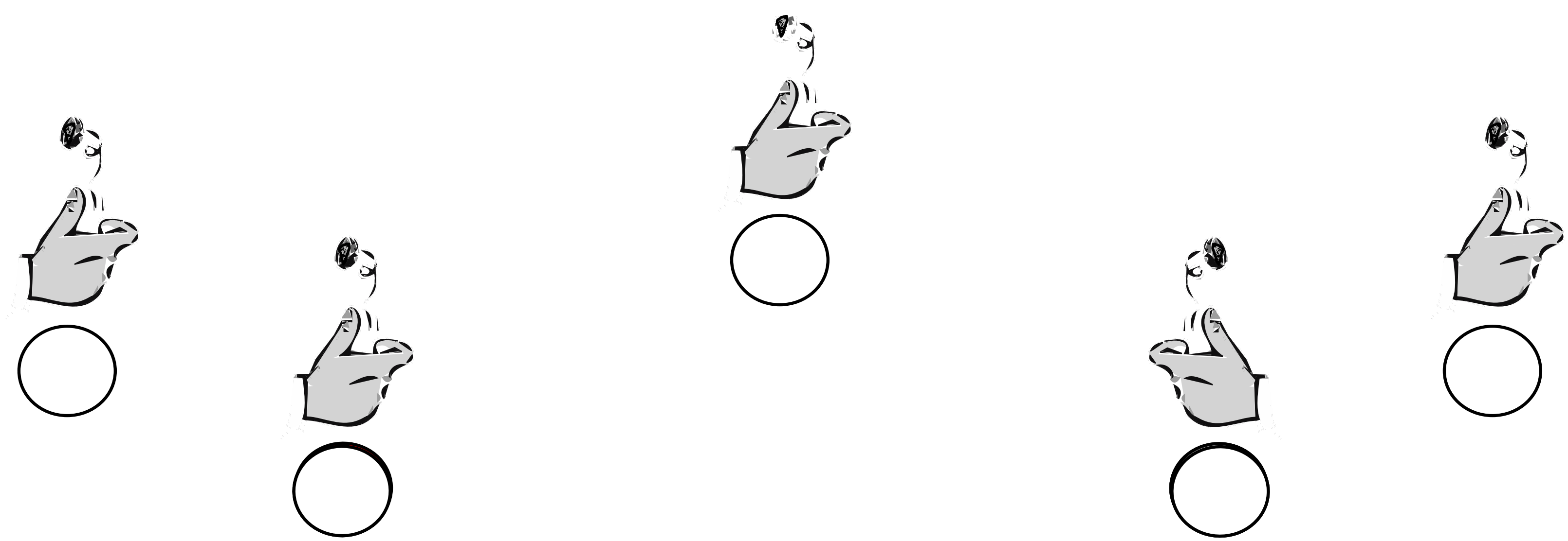
- At $t = 0$, randomly pick value $x^{(t)} \in \{\mathbf{+1}, \mathbf{-1}\}$.
- Then, at each round
  1. Set value $x^{(t)}$ to average of neighbors,
  2. Set label to **blue** if $x^{(t)} < x^{(t-1)}$, **red** otherwise.



The Average Dynamics

Al nodes at the same time:

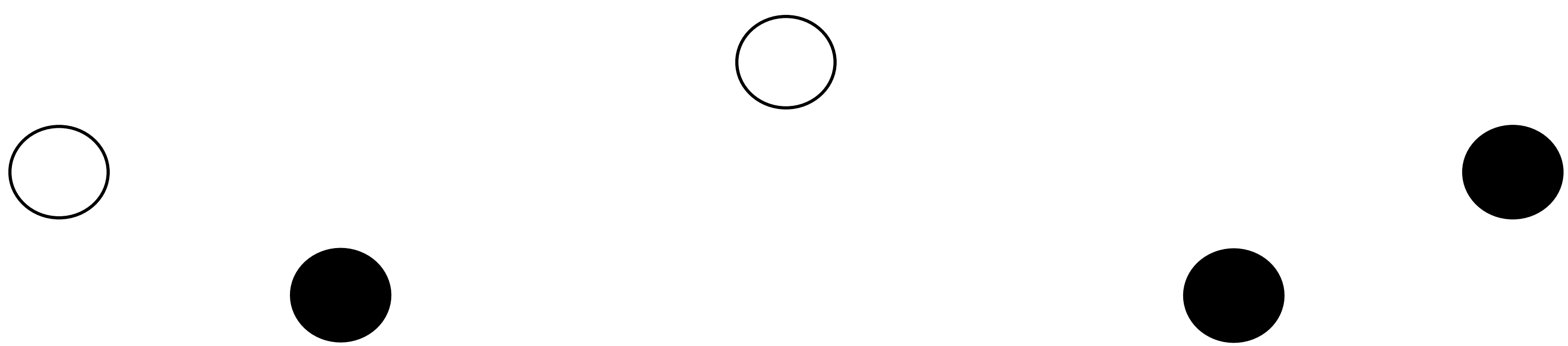
- At $t = 0$, randomly pick value $x^{(t)} \in \{+1, -1\}$.
- Then, at each round
  1. Set value $x^{(t)}$ to average of neighbors,
  2. Set label to **blue** if $x^{(t)} < x^{(t-1)}$, **red** otherwise.



The Average Dynamics

Al nodes at the same time:

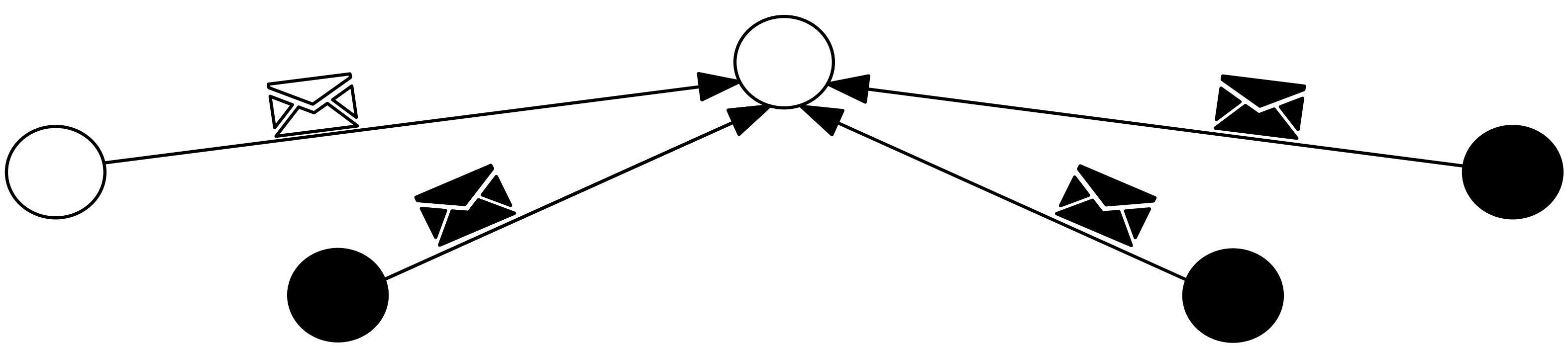
- At $t = 0$, randomly pick value $x^{(t)} \in \{\mathbf{+1}, \mathbf{-1}\}$.
- Then, at each round
  1. Set value $x^{(t)}$ to average of neighbors,
  2. Set label to **blue** if $x^{(t)} < x^{(t-1)}$, **red** otherwise.



The Average Dynamics

Al nodes at the same time:

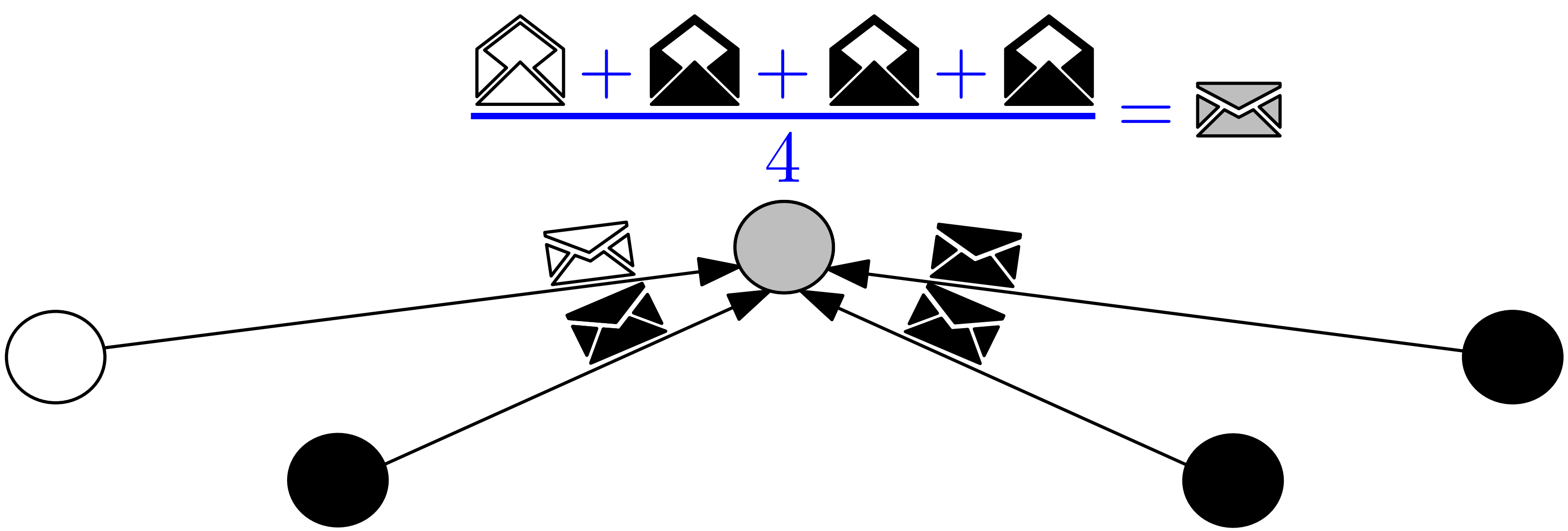
- At $t = 0$, randomly pick value $x^{(t)} \in \{$**+1**, **-1**$\}$.
- Then, at each round
  1. Set value $x^{(t)}$ to average of neighbors,
  2. Set label to **blue** if $x^{(t)} < x^{(t-1)}$, **red** otherwise.



The Average Dynamics

Al nodes at the same time:

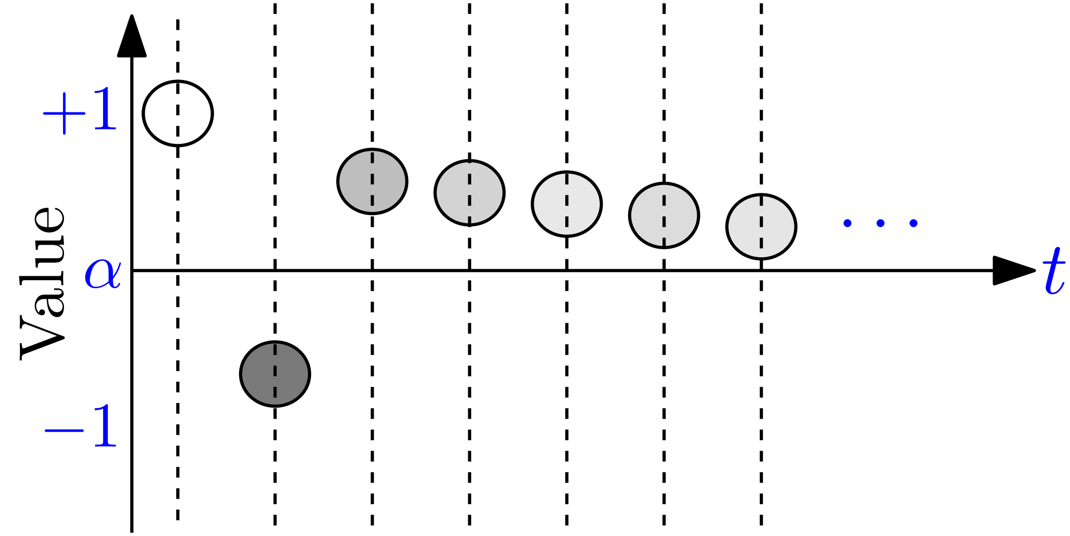
- At $t = 0$, randomly pick value $x^{(t)} \in \{\mathbf{+1}, \mathbf{-1}\}$.
- Then, at each round
  1. Set value $x^{(t)}$ to average of neighbors,
  2. Set label to **blue** if $x^{(t)} < x^{(t-1)}$, **red** otherwise.

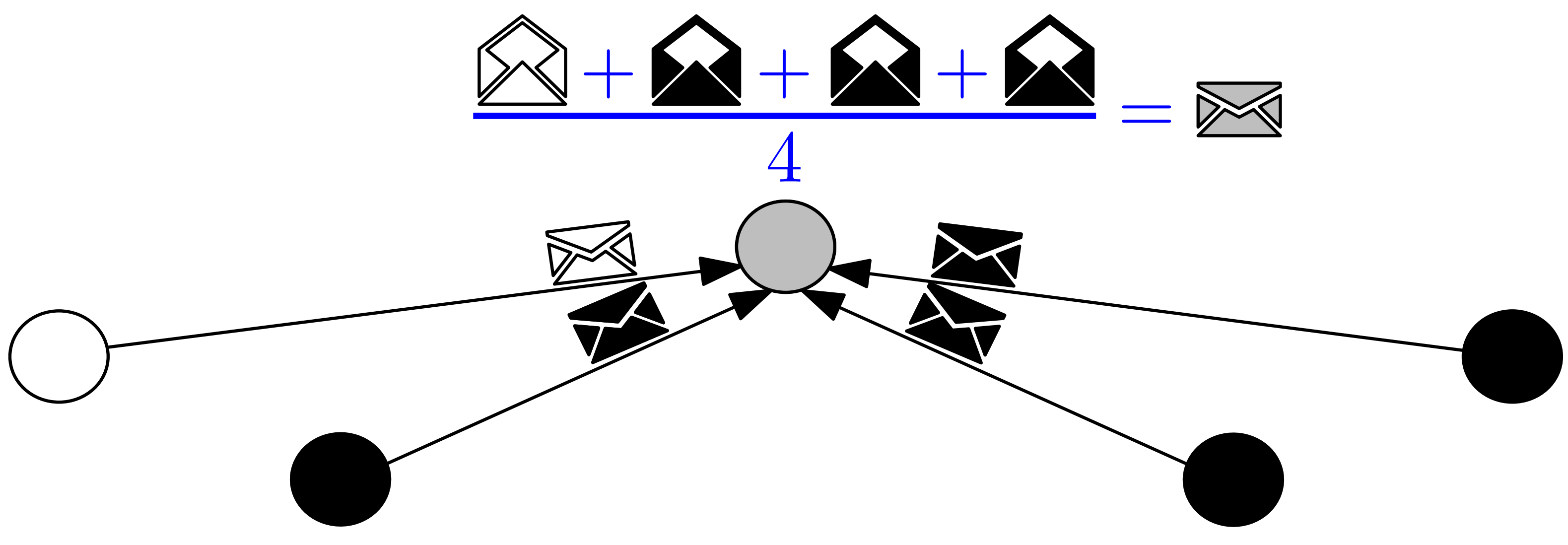


The Average Dynamics

Al nodes at the same time:

- At $t = 0$, randomly pick value $x^{(t)} \in \{+1, -1\}$.
- Then, at each round
  1. Set value $x^{(t)}$ to average of neighbors,
  2. Set label to **blue** if $x^{(t)} < x^{(t-1)}$, **red** otherwise.

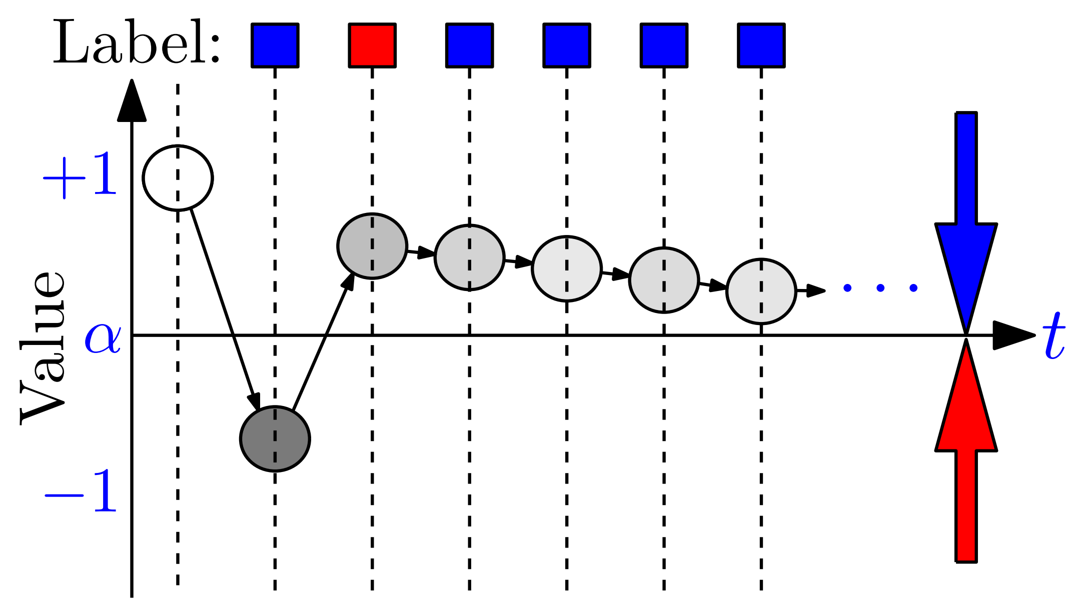


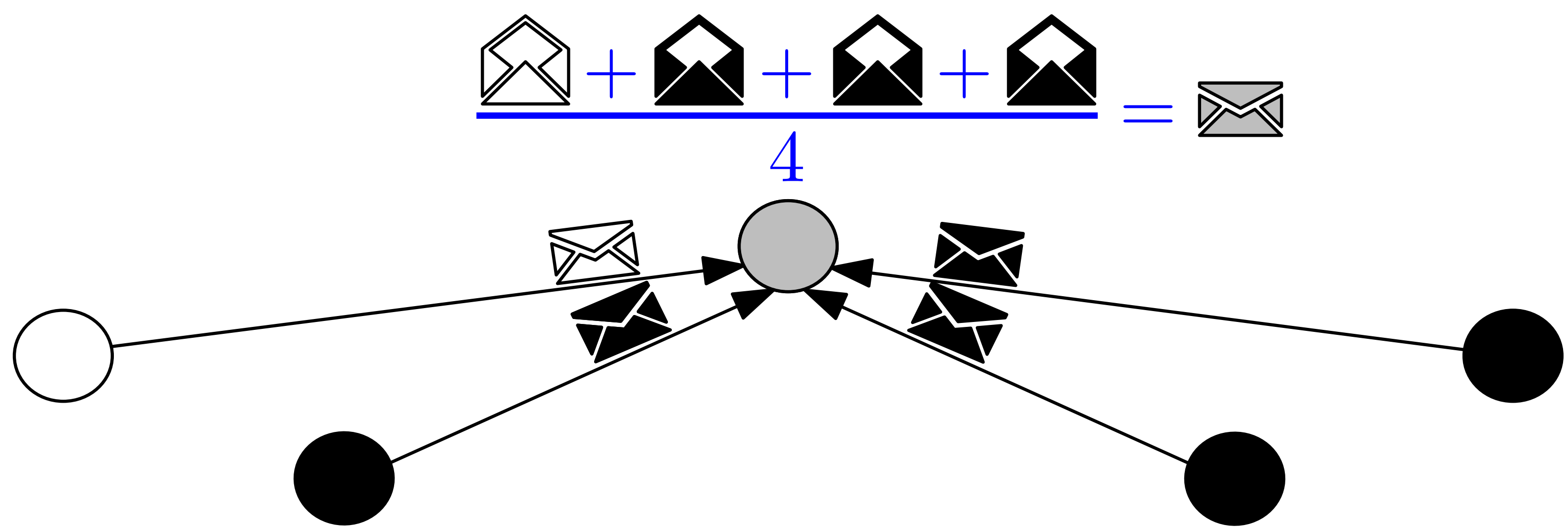


The Average Dynamics

Al nodes at the same time:

- At $t = 0$, randomly pick value $x^{(t)} \in \{\mathbf{+1}, \mathbf{-1}\}$.
- Then, at each round
  1. Set value $x^{(t)}$ to average of neighbors,
  2. Set label to **blue** if $x^{(t)} < x^{(t-1)}$, **red** otherwise.





Properties of the Averaging Dynamics

Al nodes at the same time:

- At $t = 0$, randomly pick value $x^{(t)} \in \{\textbf{blue}, \textbf{red}\}$.
- Then, at each round
  1. Set color $x^{(t)}$ to average of neighbors,
  2. Set label to **blue** if $x^{(t)} < x^{(t-1)}$, **red** otherwise.

P transition matrix of simple random walk on the graph

Averaging is a **linear** dynamics

$$\mathbf{x}^{(t)} = \begin{pmatrix} \circ \\ \bullet \\ \circ \\ \bullet \\ \bullet \end{pmatrix}$$

$$\boxed{\mathbf{x}^{(t)} = P \cdot \mathbf{x}^{(t-1)} = P^{t} \cdot \mathbf{x}^{(0)}}$$

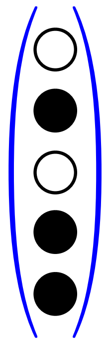
Properties of the Averaging Dynamics

Al nodes at the same time:

- At $t = 0$, randomly pick value $x^{(t)} \in \{\textbf{blue}, \textbf{red}\}$.
- Then, at each round
  1. Set color $x^{(t)}$ to average of neighbors,
  2. Set label to **blue** if $x^{(t)} < x^{(t-1)}$, **red** otherwise.

P transition matrix of simple random walk on the graph

Averaging is a **linear** dynamics

$\mathbf{x}^{(t)} =$ 

$$\mathbf{x}^{(t)} = P \cdot \mathbf{x}^{(t-1)} = P^t \cdot \mathbf{x}^{(0)}$$

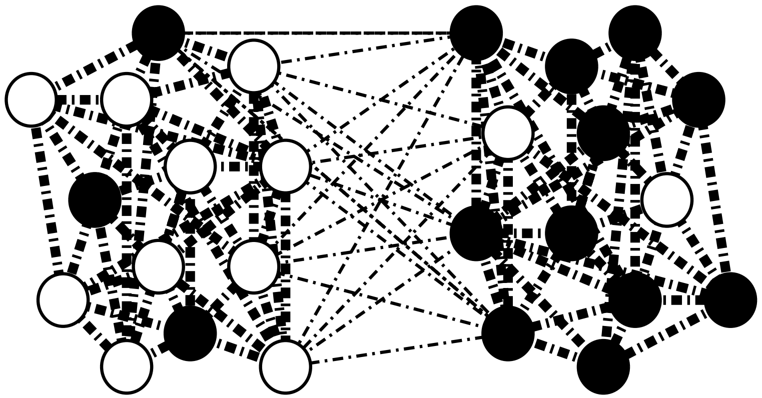
Bottleneck of mixing time for spectral methods:

Distributed computation of second eigenvector

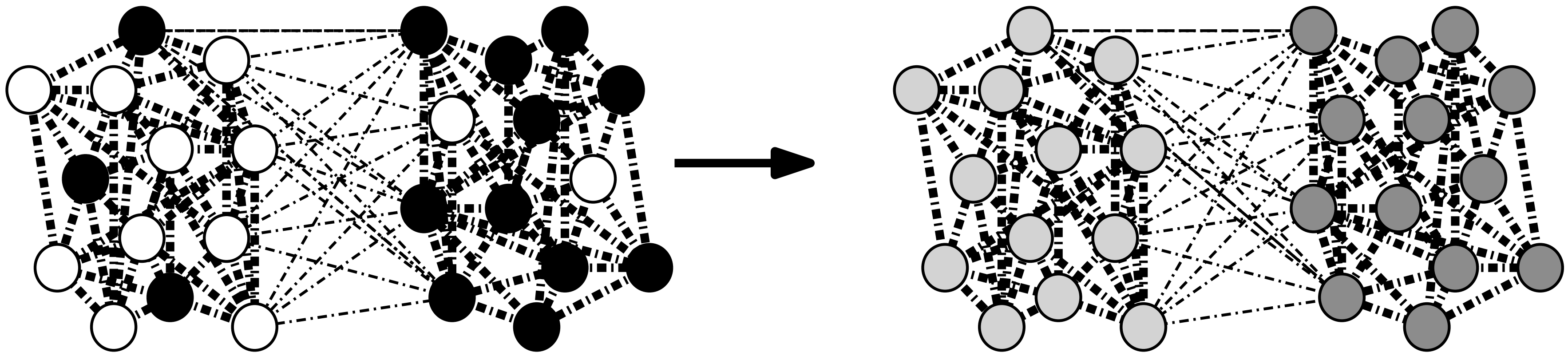
[Kempe & McSherry '08]: $\mathcal{O}(\tau_{mix} \log^2 n)$.

$\lambda_2(P) \approx \frac{a-b}{a+b} \implies$ mixing time of a random walk on $\mathcal{G}_{n,p,q}$ is $\geq \frac{1}{1-\lambda_2} \approx \frac{a+b}{2b}$.

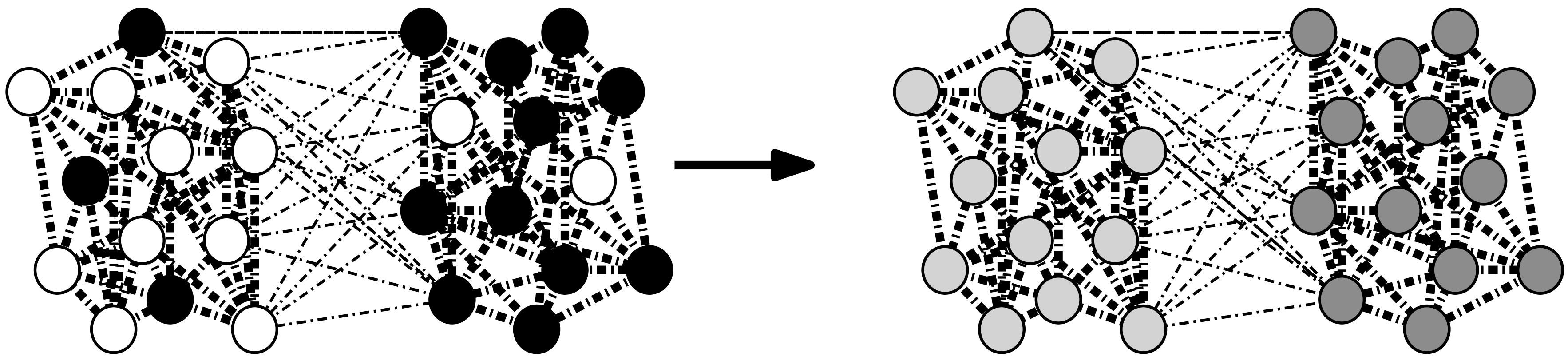
Our Results



Our Results



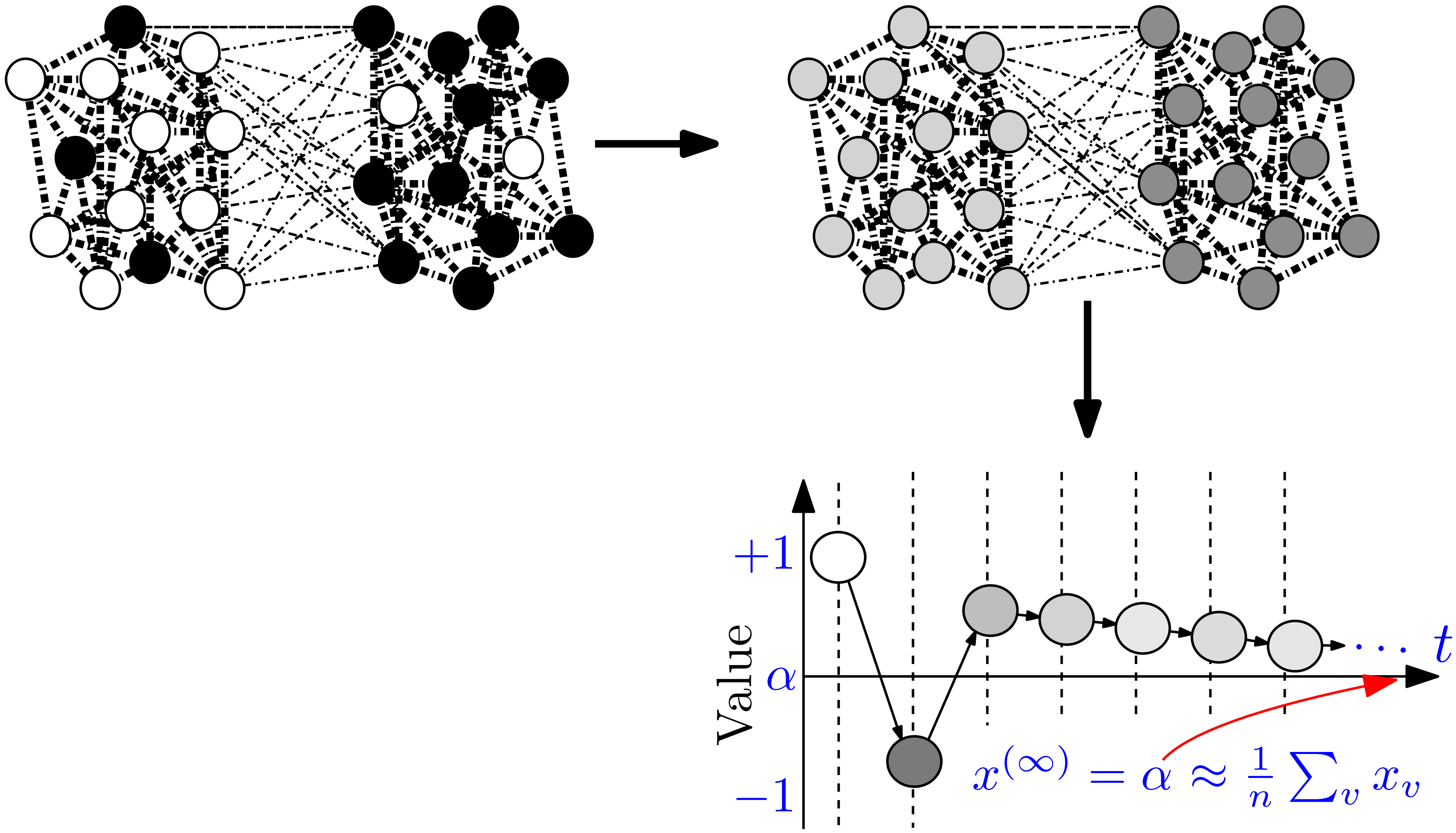
Our Results



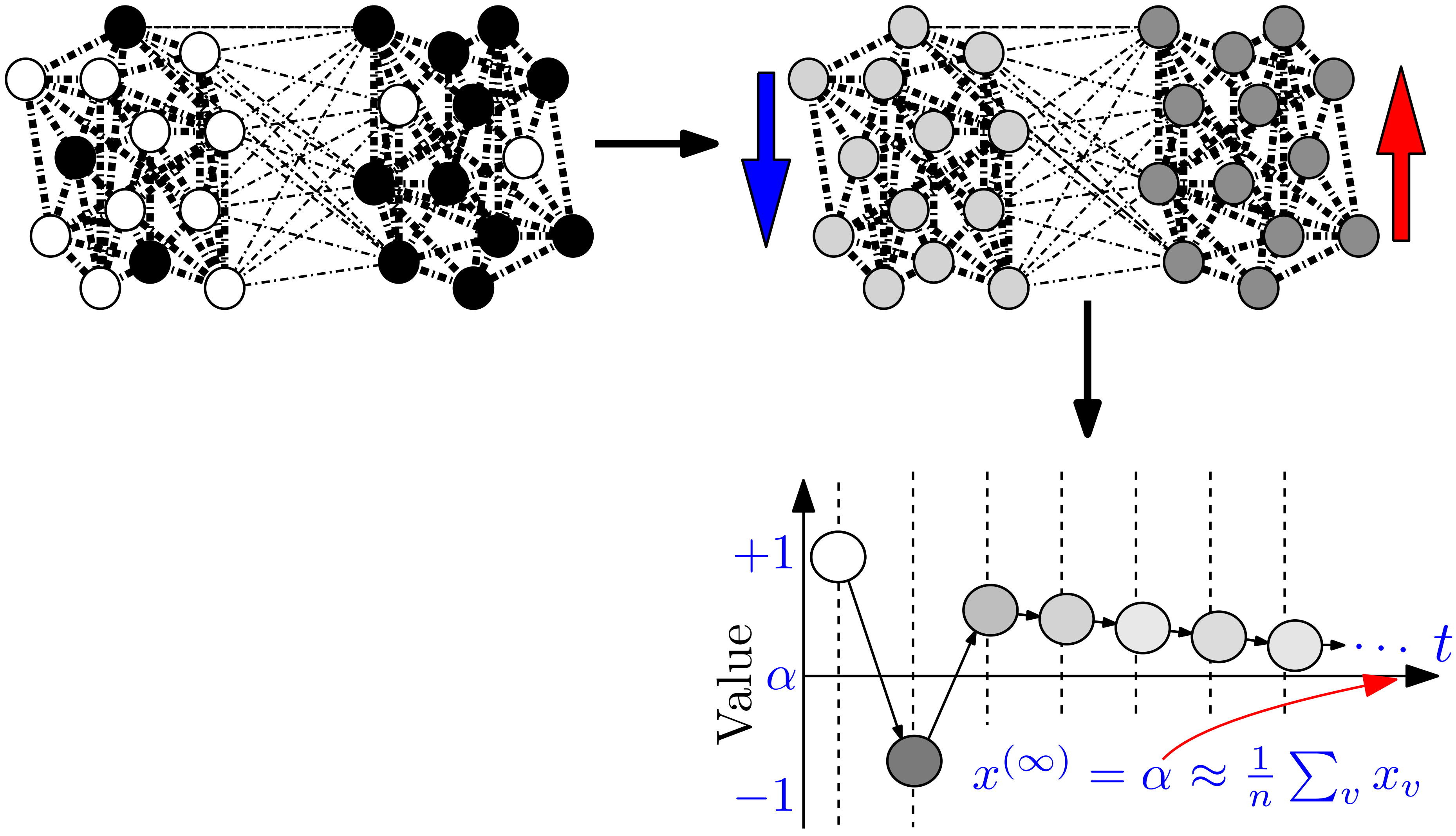
Let's say nodes are in the same community if their distance is at least ϵ...

- How to set ϵ?
- Not a global clustering.

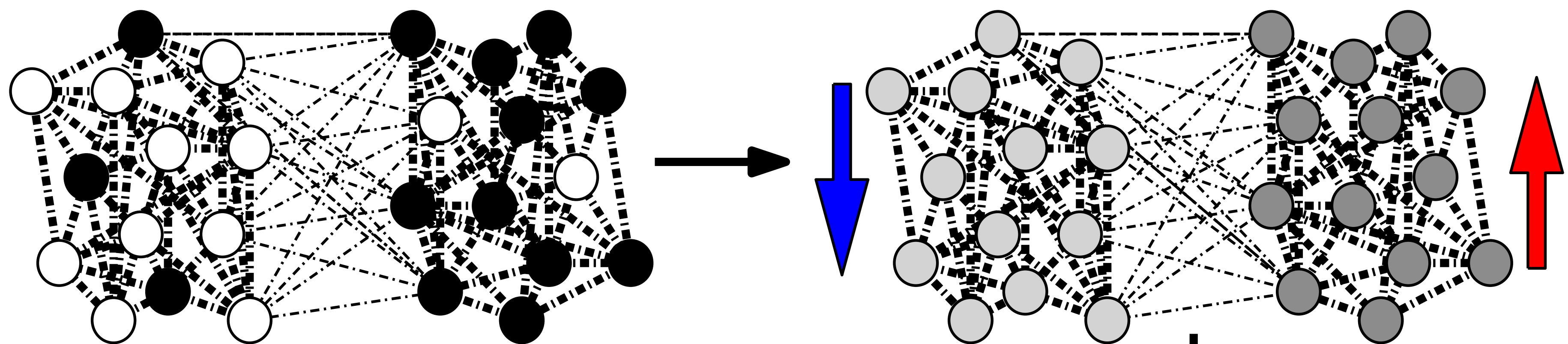
Our Results



Our Results



Our Results

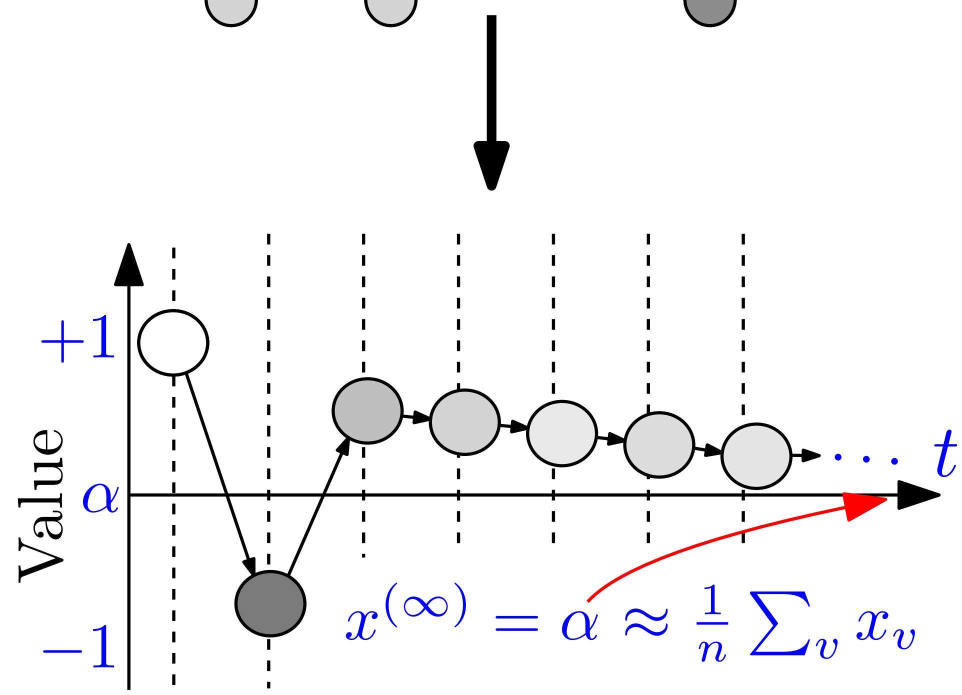


$v_1, \ldots, v_n$ eigenvectors of random walk matrix P:

$$v_1 = \mathbb{1} = (1, \ldots, 1)$$

$$v_2 \approx \chi = (1, \ldots, 1, -1, \ldots, -1)$$

"nice" graph



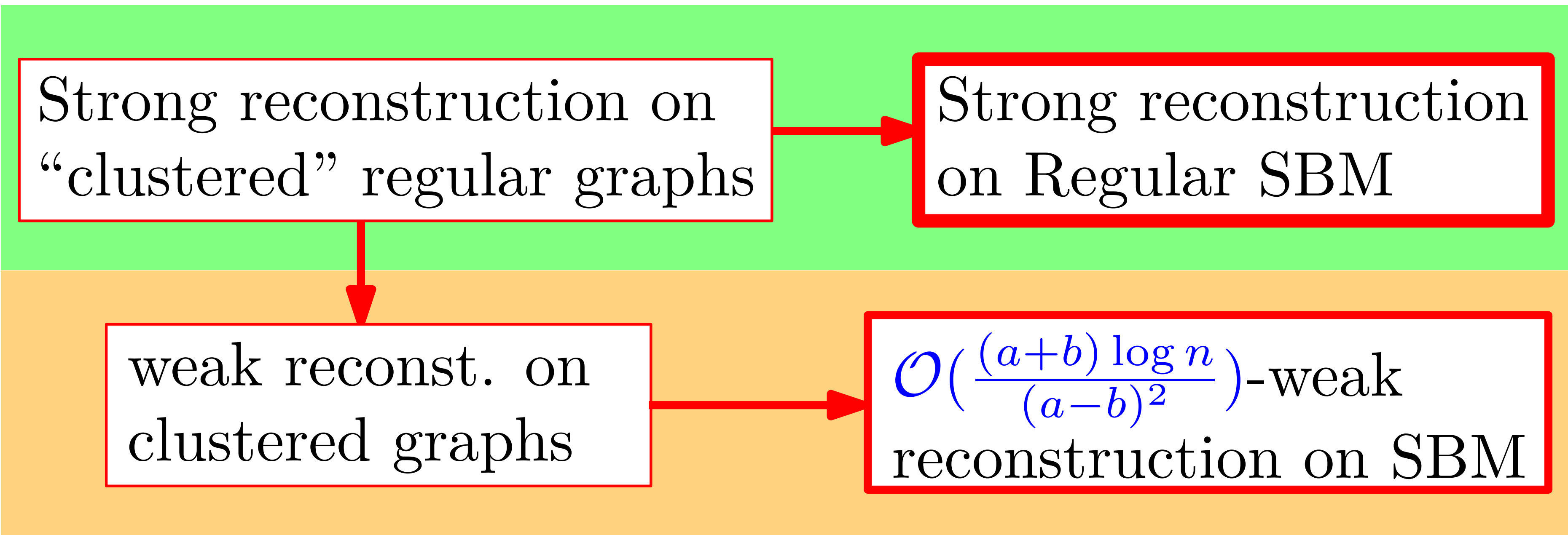
Our Results

(Informal) Theorem. $G = (V_1 \dot{\bigcup} V_2, E)$ s.t.
i) $\chi = \mathbf{1}_{V_1} - \mathbf{1}_{V_2}$ close to right-eigenvector of eigenvalue λ_2 of transition matrix of G, and
ii) gap between λ_2 and $\lambda = \max\{\lambda_3, |\lambda_n|\}$ sufficiently large, then
Averaging (approximately) identifies (V_1, V_2).

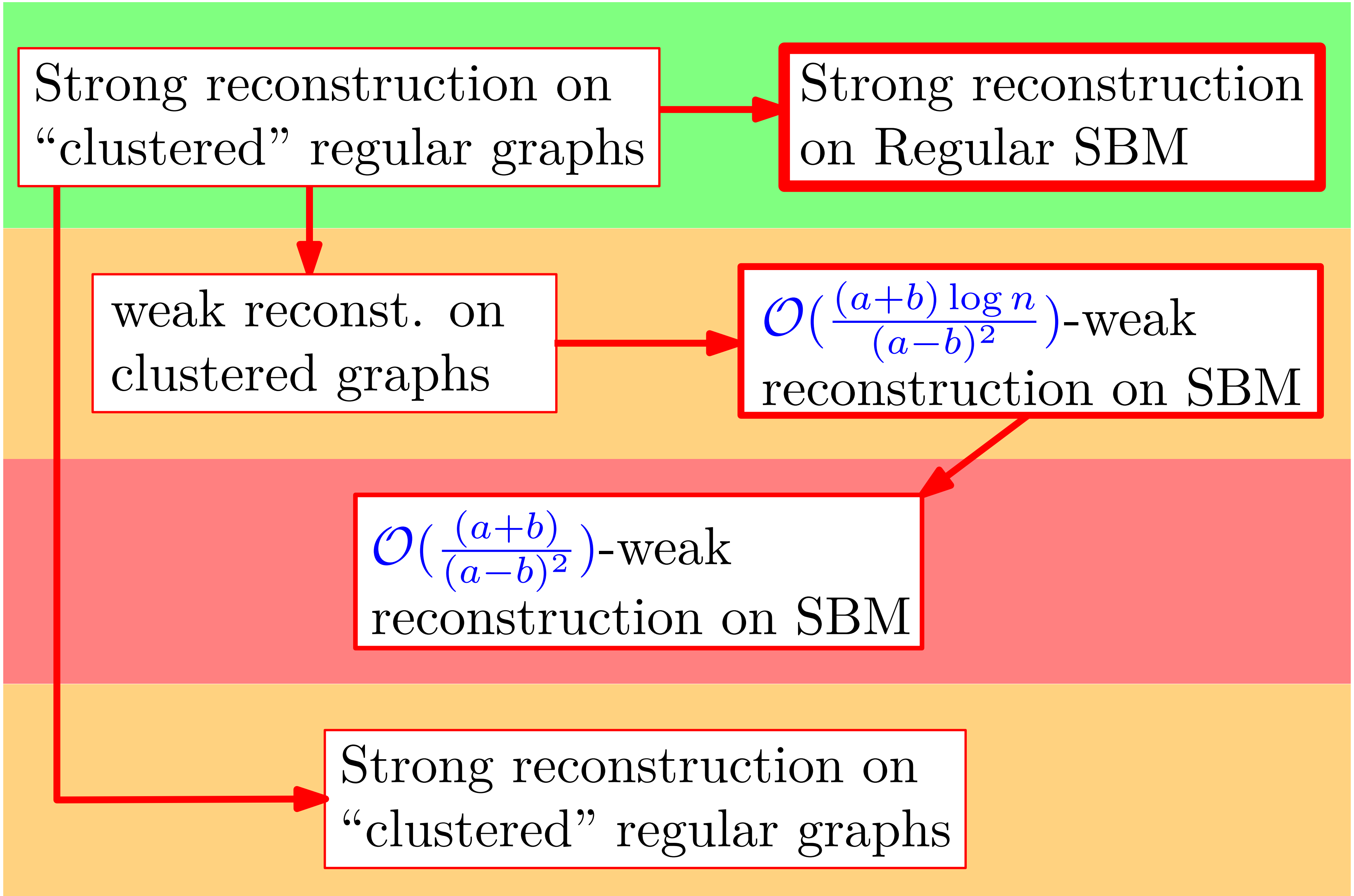
Above conditions are met w.h.p. if

- in Regular SBM, $a - b > 2\sqrt{a + b - 1}$ (Strong reconstruction)
- in SBM, if $a - b > \sqrt{(a + b)\log n}$ and $b > \frac{\log n}{n^2}$ ($\mathcal{O}\left(\frac{(a+b)\log n}{(a-b)^2}\right)$-weak reconstruction.)

Analysis: Roadmap



Analysis: Roadmap



Analysis on Regular SBM

P $\longrightarrow$ symmetric $\Longrightarrow$ orthonormal eigenvectors $\mathbf{v}_1, ..., \mathbf{v}_n$ and real eigenvalues $\lambda_1, ..., \lambda_n$.

Analysis on Regular SBM

P → symmetric $\implies$ orthonormal eigenvectors $\mathbf{v}_1, ..., \mathbf{v}_n$ and real eigenvalues $\lambda_1, ..., \lambda_n$.

$$\mathbf{x}^{(t)} = P^t \cdot \mathbf{x}^{(0)} = \sum_i \lambda_i^t (\mathbf{v}_i^\intercal \mathbf{x}^{(0)}) \mathbf{v}_i$$

Analysis on Regular SBM

P ⟶ symmetric $\Longrightarrow$ orthonormal eigenvectors $\mathbf{v}_1, ..., \mathbf{v}_n$ and real eigenvalues $\lambda_1, ..., \lambda_n$.

$$\mathbf{x}^{(t)} = P^t \cdot \mathbf{x}^{(0)} = \sum_i \lambda_i^t (\mathbf{v}_i^\intercal \mathbf{x}^{(0)}) \mathbf{v}_i$$

$$\mathbf{v}_1 = \frac{1}{\sqrt{n}} \mathbf{1}$$

Regular SBM $\Longrightarrow$ $P\chi = \left(\frac{a-b}{a+b}\right) \cdot \chi$

Analysis on Regular SBM

P ⟶ symmetric $\implies$ orthonormal eigenvectors $\mathbf{v}_1, \ldots, \mathbf{v}_n$ and real eigenvalues $\lambda_1, \ldots, \lambda_n$.

$$\mathbf{x}^{(t)} = P^t \cdot \mathbf{x}^{(0)} = \sum_i \lambda_i^t (\mathbf{v}_i^\intercal \mathbf{x}^{(0)}) \mathbf{v}_i$$

$$\mathbf{v}_1 = \frac{1}{\sqrt{n}} \mathbf{1}$$

Regular SBM $\implies$ $P\chi = (\frac{a-b}{a+b}) \cdot \chi$

$$\frac{1}{a+b} \left(\begin{array}{c|c} \cdots\cdots\cdots & \cdots\cdots\cdots \\ \cdots a \text{ “1”s} \cdots & \cdots b \text{ “1”s} \cdots \\ \cdots\cdots\cdots & \cdots\cdots\cdots \\ \hline \cdots\cdots\cdots & \cdots\cdots\cdots \\ \cdots b \text{ “1”s} \cdots & \cdots a \text{ “1”s} \cdots \\ \cdots\cdots\cdots & \cdots\cdots\cdots \end{array} \right) \cdot \begin{pmatrix} 1 \\ \vdots \\ 1 \\ -1 \\ \vdots \\ -1 \end{pmatrix} = \frac{a-b}{a+b} \begin{pmatrix} 1 \\ \vdots \\ 1 \\ -1 \\ \vdots \\ -1 \end{pmatrix}$$

Analysis on Regular SBM

P → symmetric $\implies$ orthonormal eigenvectors $\mathbf{v}_1, ..., \mathbf{v}_n$ and real eigenvalues $\lambda_1, ..., \lambda_n$.

$$\mathbf{x}^{(t)} = P^t \cdot \mathbf{x}^{(0)} = \sum_i \lambda_i^t (\mathbf{v}_i^\intercal \mathbf{x}^{(0)}) \mathbf{v}_i$$

$$\mathbf{v}_1 = \tfrac{1}{\sqrt{n}} \mathbf{1}$$

Regular SBM $\implies$ $P\chi = (\frac{a-b}{a+b}) \cdot \chi$

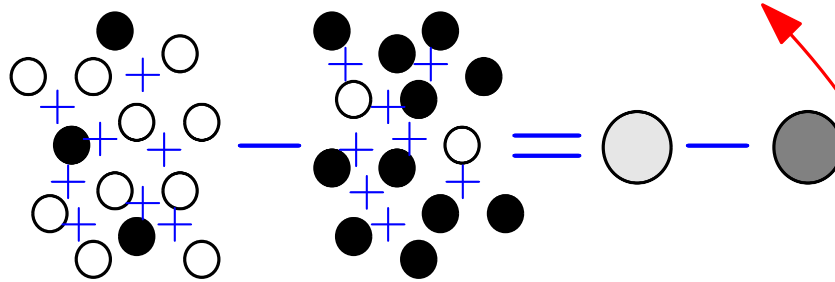
W.h.p. $\max\{\lambda_3, |\lambda_n|\}(1+\delta) < \frac{a-b}{a+b} = \lambda_2$, then

$$\mathbf{x}^{(t+1)} = \frac{1}{n}(\mathbf{1}^\intercal \mathbf{x}^{(0)})\mathbf{1} + \lambda_2^t \frac{1}{n}(\chi^\intercal \mathbf{x}^{(0)})\chi + \mathbf{e}^{(t)}$$

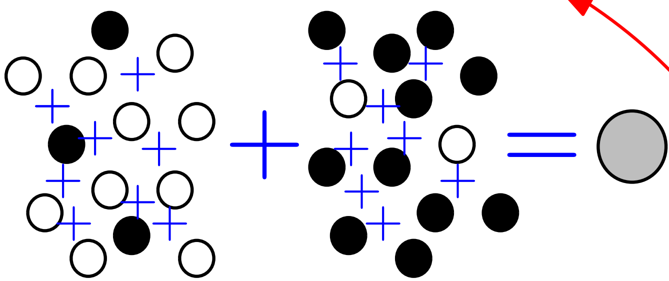
with $\|\mathbf{e}^{(t)}\| \leq (\max\{\lambda_3, |\lambda_n|\})^t \sqrt{n}$

Analysis on Regular SBM

$$\left(\frac{1}{n}\sum_{u\in V_1}\mathbf{x}^{(0)}(u) - \frac{1}{n}\sum_{u\in V_2}\mathbf{x}^{(0)}(u)\right)$$



$$\frac{1}{n}\sum_{u\in V}\mathbf{x}^{(0)}(u)$$



W.h.p. $\max\{\lambda_3, |\lambda_n|\}(1+\delta) < \frac{a-b}{a+b} = \lambda_2$, then

$$\mathbf{x}^{(t+1)} = \frac{1}{n}(\mathbf{1}^\intercal \mathbf{x}^{(0)})\mathbf{1} + \lambda_2^t \frac{1}{n}(\chi^\intercal \mathbf{x}^{(0)})\chi + \mathbf{e}^{(t)}$$

with $\|\mathbf{e}^{(t)}\| \leq (\max\{\lambda_3, |\lambda_n|\})^t \sqrt{n}$

Analysis on Regular SBM

$$\mathbf{x}^{(t)} = \frac{1}{n}(\mathbf{1}^\intercal \mathbf{x}^{(0)})\mathbf{1} + \lambda_2^t \frac{1}{n}(\chi^\intercal \mathbf{x}^{(0)})\chi + \mathbf{e}^{(t)}$$

Analysis on Regular SBM

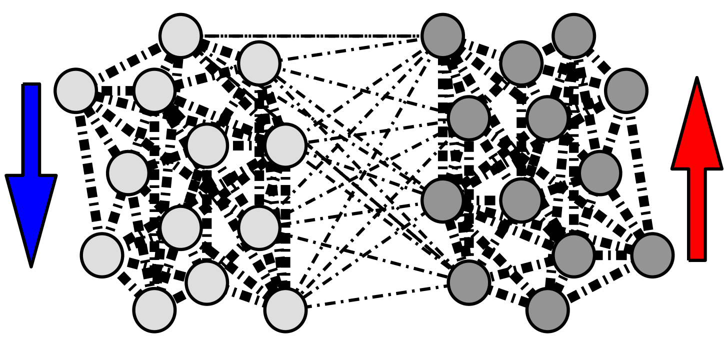
$$\mathbf{x}^{(t)} = \frac{1}{n}(\mathbf{1}^\intercal \mathbf{x}^{(0)})\mathbf{1} + \lambda_2^t \frac{1}{n}(\chi^\intercal \mathbf{x}^{(0)})\chi + \mathbf{e}^{(t)}$$

$$\mathbf{x}^{(t)} - \mathbf{x}^{(t-1)} = (\chi^\intercal \mathbf{x}^{(0)})\lambda_2^{t-1}(\lambda_2 - 1)\chi + \underbrace{\mathbf{e}^{(t)} - \mathbf{e}^{(t-1)}}_{\ll \lambda_2^{t-1} \text{ if } t = \Omega(\log n)}$$

Analysis on Regular SBM

$$\mathbf{x}^{(t)} = \frac{1}{n}(\mathbf{1}^\intercal \mathbf{x}^{(0)})\mathbf{1} + \lambda_2^t \frac{1}{n}(\chi^\intercal \mathbf{x}^{(0)})\chi + \mathbf{e}^{(t)}$$

$$\mathbf{x}^{(t)} - \mathbf{x}^{(t-1)} = (\chi^\intercal \mathbf{x}^{(0)})\lambda_2^{t-1}(\lambda_2 - 1)\chi + \underbrace{\mathbf{e}^{(t)} - \mathbf{e}^{(t-1)}}_{\ll \lambda_2^{t-1} \text{ if } t=\Omega(\log n)}$$



$$\mathrm{sign}(\mathbf{x}^{(t)}(u) - \mathbf{x}^{(t-1)}(u)) \propto \mathrm{sign}(\chi(u))$$

Future Work: Sparsification

At each round, pick an edge u.a.r.
(*population protocols*):
those two nodes averages their values.

Simulations. Does not seem to work for
$a - b \ll \log n$.

Analysis. A version with $\log n$ parallel instances (say two nodes are in same community only iff at least a certain fraction of instances agree), works for
$a - b \gg \log^{\Theta(1)} n$.

Thank

You!