Community Detection on an Euclidean Random Graph

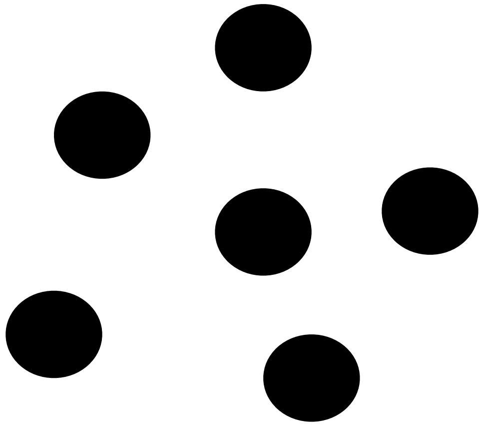
Abishek Sankararaman, Emmanuel Abbe and François Baccelli

Jan 2020

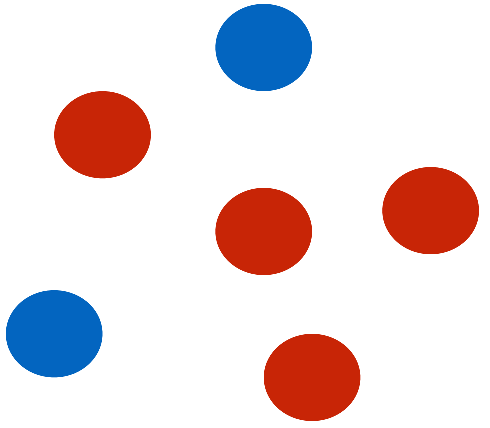
Community Detection - Abstract Definition

- Grouping objects given ***indirect information*** of memberships.

A population

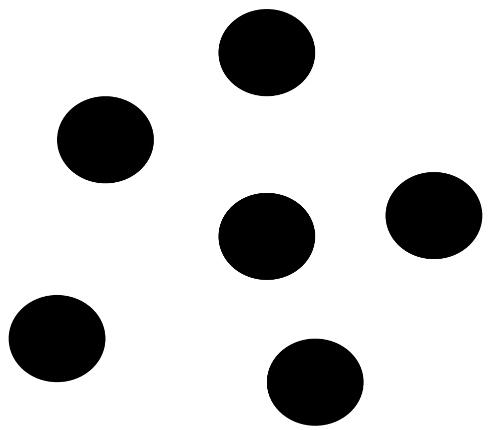
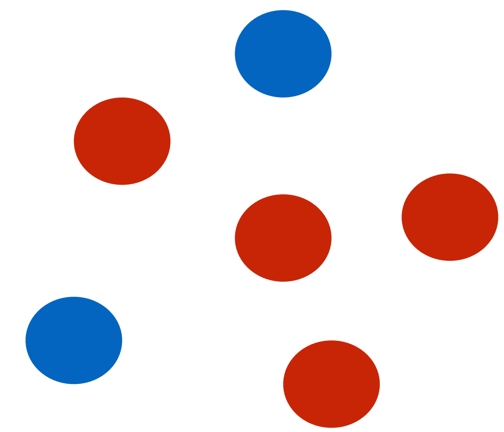


partitioned into groups



Community Detection - Examples

- Grouping objects given ***indirect information*** of memberships.

A population  partitioned into groups 

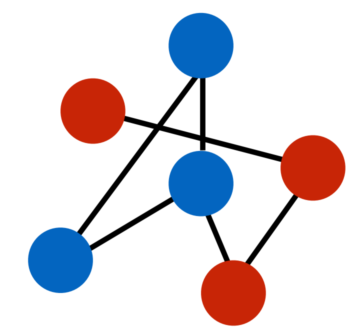
1. People on an Online Social Network. 

2. Proteins classified into groups based on their functional behavior.

3. Grouping Base-Stations based on similarities in traffic pattern.

Graph as Information

Important sub-class



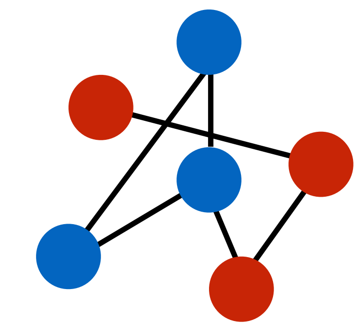
Population - Represented as nodes of a graph.

Membership Information - Encoded as labeled edges of the graph.

Graph Clustering Problem -

Given an unlabeled graph data, recover the partition of nodes.

Graph Clustering



<u>Graph Clustering</u> -

Given an <u>unlabeled graph data</u>, recover the partition of nodes.

What if there are additional contextual information on each node ?

Web-pages, the textual content in a page.

Social Networks - Personal information (age, location, income….)

Computational Biology - Metadata generated by measurements.

Planted Partition Random Connection Model

Planted Partition Random Connection Model

Vertex Set - $\{1, 2, \cdots, N_n\}$ N_n - # *nodes*

Each node $i \in [1, N_n]$ has two labels -

location label $X_i \in \mathbb{R}^d$ and a community label $Z_i \in \{-1, 1\}$

.

Planted Partition Random Connection Model

Vertex Set - $\{1, 2, \cdots, N_n\}$ N_n - # *nodes*

Each node $i \in [1, N_n]$ has two labels -

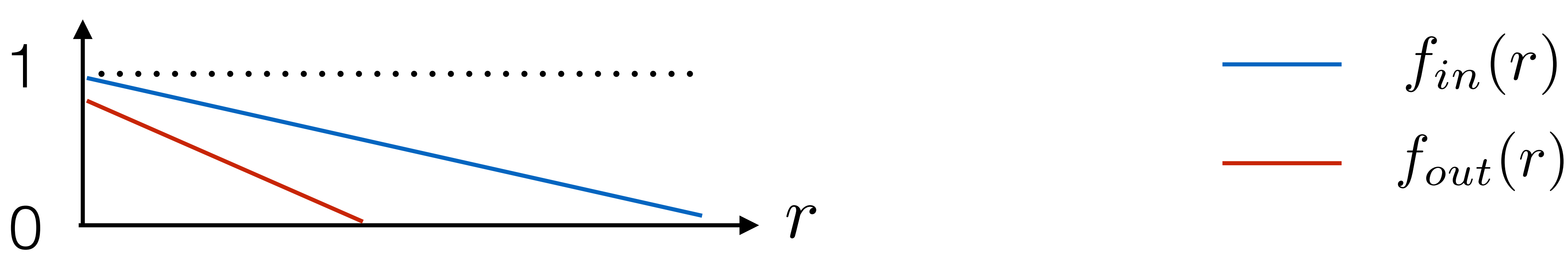
location label $X_i \in \mathbb{R}^d$ and a community label $Z_i \in \{-1, 1\}$

Random Graph Parameters

$\lambda > 0$ Intensity.

$d \geq 2$ Dimension of embedding.

$$f_{in}(\cdot), f_{out}(\cdot) : \mathbb{R}_+ \to [0, 1] \text{ s.t } \forall r \geq 0 \ , \ f_{in}(r) \geq f_{out}(r)$$



Planted Partition Random Connection Model

Planted Partition Random Connection Model

1) $N_n \sim \mathrm{Poisson}(\lambda n)$ number of nodes

On avg λ *points per unit area.*

Planted Partition Random Connection Model

1) $N_n \sim \mathrm{Poisson}(\lambda n)$ number of nodes

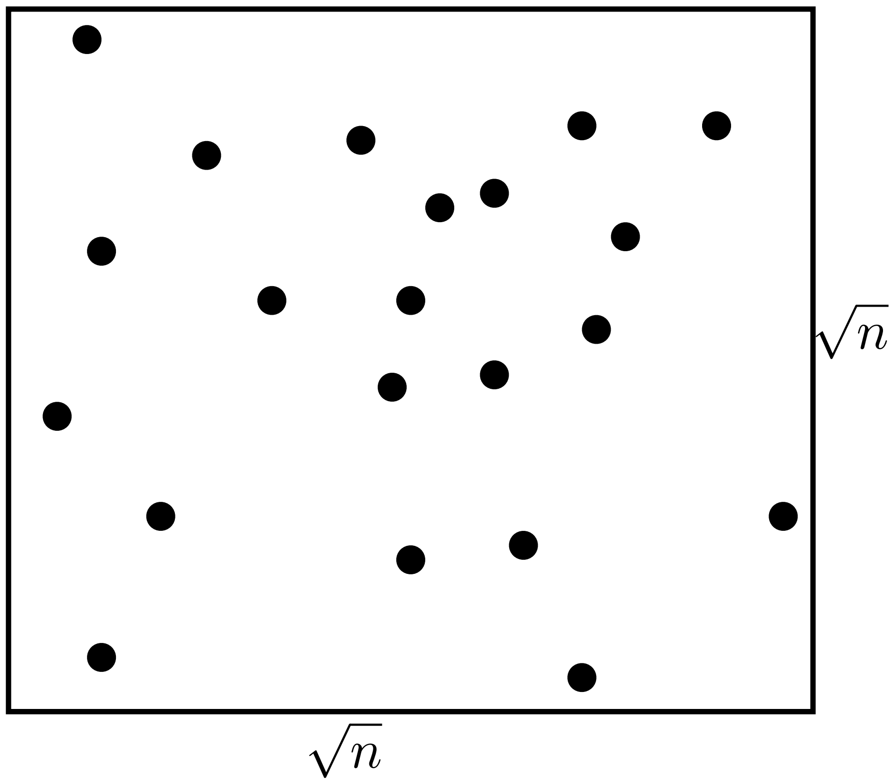
On avg λ points per unit area.

2) Each node $i \in [1, N_n]$, has a

- ***Location label*** $X_i \in \left[-\frac{n^{1/d}}{2}, \frac{n^{1/d}}{2}\right]$

sampled independently and uniformly

Planted Partition Random Connection Model



1) $N_n \sim \mathrm{Poisson}(\lambda n)$ number of nodes

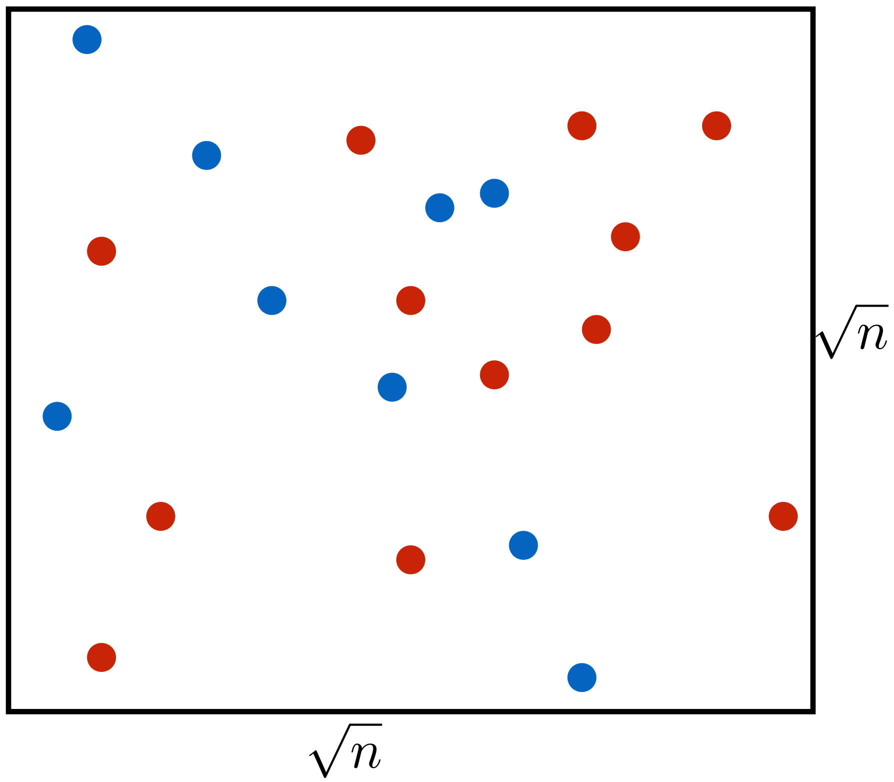
On avg λ points per unit area.

2) Each node $i \in [1, N_n]$, has a

- ***Location label*** $X_i \in \left[-\frac{n^{1/d}}{2}, \frac{n^{1/d}}{2}\right]$

sampled independently and uniformly

Planted Partition Random Connection Model



1) $N_n \sim \mathrm{Poisson}(\lambda n)$ number of nodes

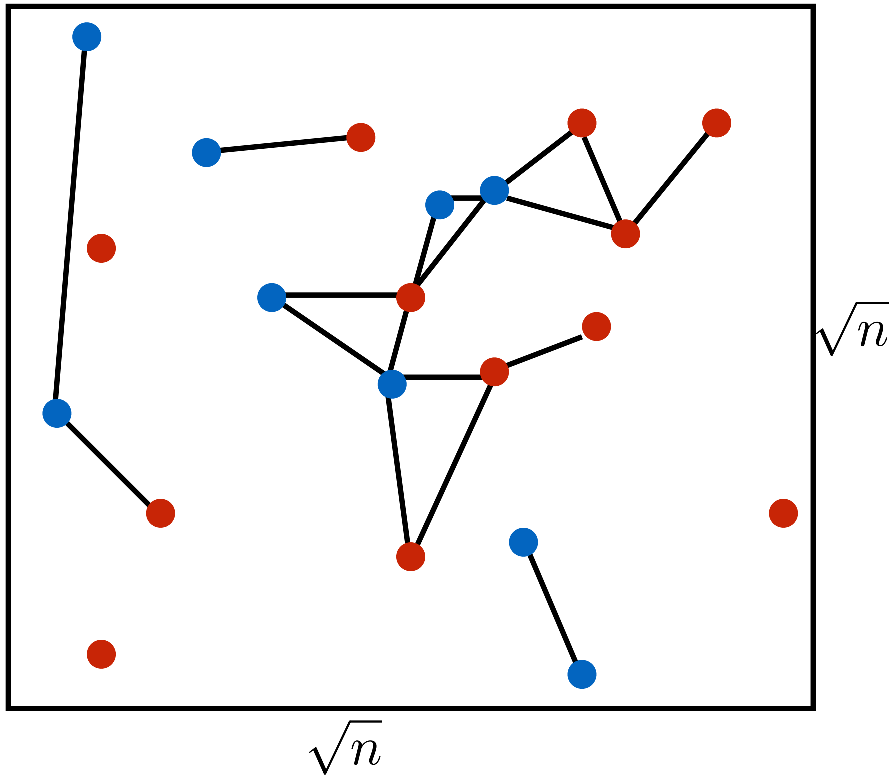
   On avg λ points per unit area.

2) Each node $i \in [1, N_n]$, has a

   - ***Location label*** $X_i \in \left[-\frac{n^{1/d}}{2}, \frac{n^{1/d}}{2}\right]$
   - ***Community label*** $Z_i \in \{-1, +1\}$

   sampled independently and uniformly

Planted Partition Random Connection Model



1) $N_n \sim \mathrm{Poisson}(\lambda n)$ number of nodes

On avg λ points per unit area.

2) Each node $i \in [1, N_n]$, has a

- ***Location label*** $X_i \in \left[-\frac{n^{1/d}}{2}, \frac{n^{1/d}}{2}\right]$
- ***Community label*** $Z_i \in \{-1, +1\}$

sampled independently and uniformly

3) Edge between $i, j \in [1, N_n]$ with probability either

$f_{in}(||X_i - X_j||)$ - If $Z_i = Z_j$ (***same colors***)

$f_{out}(||X_i - X_j||)$ - If $Z_i \neq Z_j$ (***different colors***)

$\forall r \geq 0, 1 \geq f_{in}(r) \geq f_{out}(r) \geq 0$

More edges within communities than across.

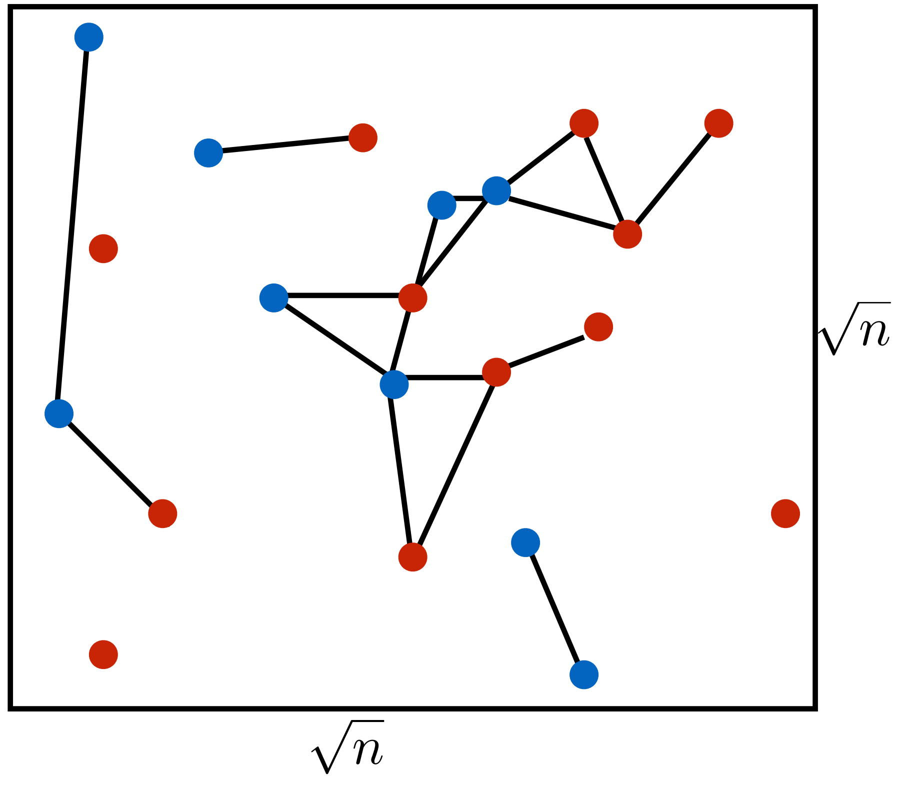
Conditional on node labels, edges are independent

Planted Partition Random Connection Model

1) $\{X_i\}_{i \in \mathbb{N}}$ - a ***Poisson Point Process*** on $\mathbb{R}^d$ with intensity λ

2) Independently ***mark*** it $\{Z_i\}_{i \in \mathbb{N}}$ each of which is uniform over $\{-1, 1\}$

3) Connect any two nodes $i \neq j \in \mathbb{N}$ with probability
$f_{in}(||X_i - X_j||)\mathbf{1}_{Z_i = Z_j} + f_{out}(||X_i - X_j||)\mathbf{1}_{Z_i \neq Z_j}$ independently for all pairs



$G_n \stackrel{d}{=} G$ restricted to $\left[-\frac{n^{1/d}}{2}, \frac{n^{1/d}}{2}\right]^d$

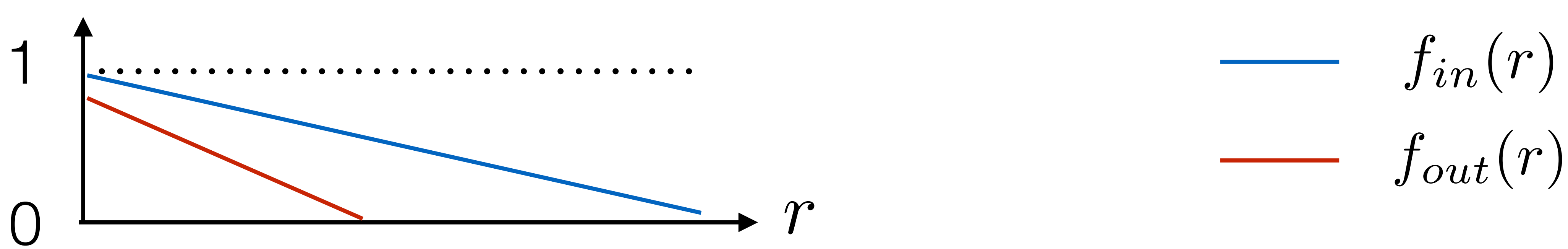
Planted Partition Random Connection Model

Model Parameters

$\lambda > 0$ Intensity

$d \geq 2$ Dimension of embedding

$$f_{in}(\cdot), f_{out}(\cdot) : \mathbb{R}_+ \to [0,1] \text{ s.t } \forall r \geq 0 \ , \ f_{in}(r) \geq f_{out}(r)$$

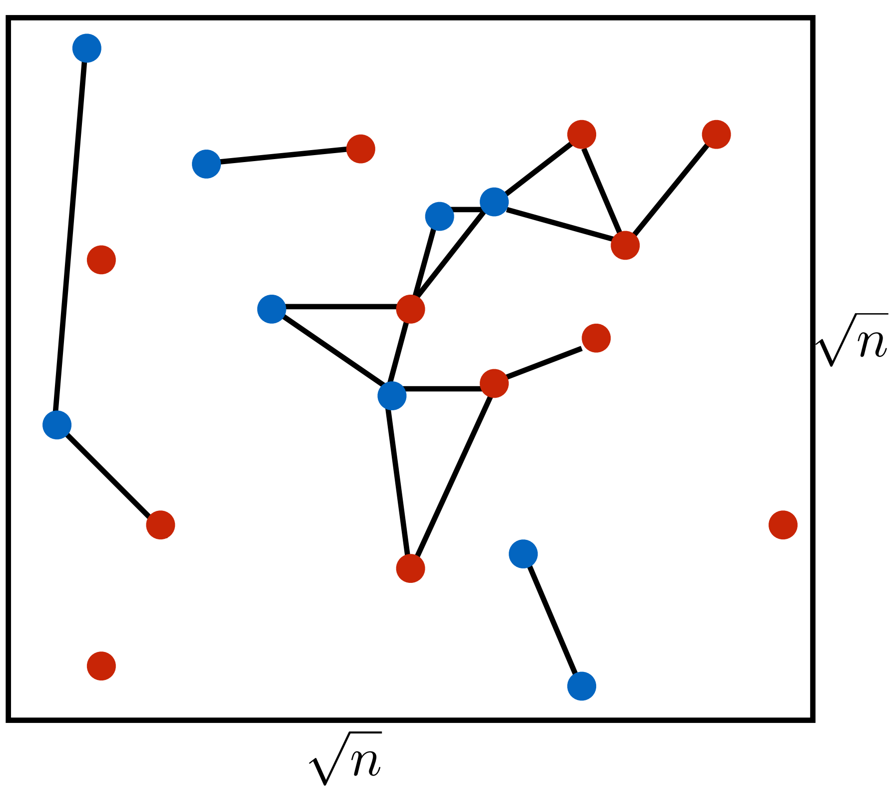


Planted Partition Random Connection Model

Assume $\displaystyle\int_{x\in\mathbb{R}^d} f_{out}(||x||)dx \leq \int_{x\in\mathbb{R}^d} f_{in}(||x||)dx < \infty$

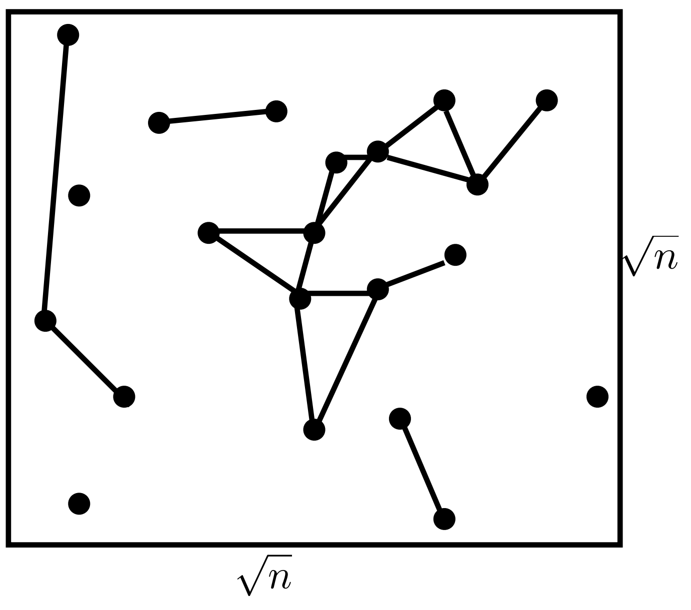
Avg # of neighbors in

- same community is - $\displaystyle(\lambda/2)\int_{x\in\mathbb{R}^d} f_{in}(||x||)dx - o(1)$
- opposite community is - $\displaystyle(\lambda/2)\int_{x\in\mathbb{R}^d} f_{out}(||x||)dx - o(1)$



Constant avg degree

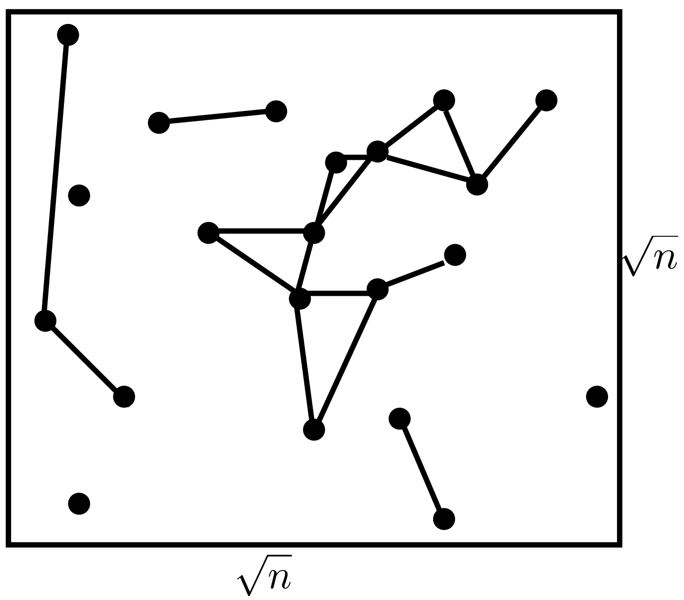
Community Detection Problem



Given G_n and $\{X_i\}_{i \in [0, N_n]}$, estimate $\{Z_i\}_{i \in [1, N_n]}$

$\{\tau_i\}_{i \in [0, N_n]}$- Community estimates

Community Detection Problem



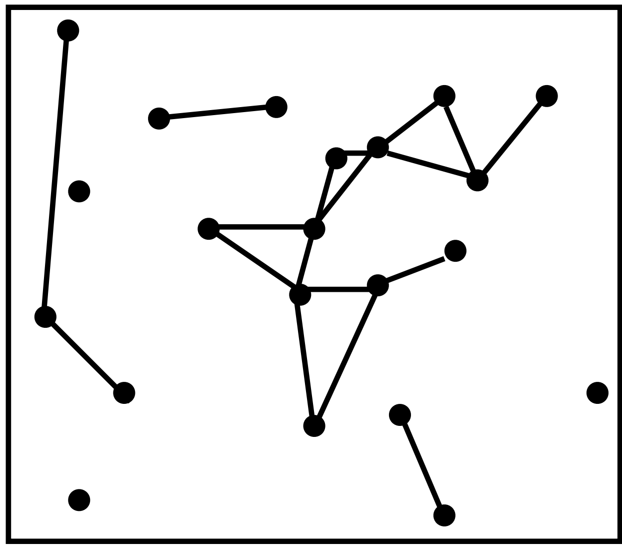
Given G_n and $\{X_i\}_{i \in [0, N_n]}$, estimate $\{Z_i\}_{i \in [1, N_n]}$

$\{\tau_i\}_{i \in [0, N_n]}$- Community estimates

$$\mathcal{O}_\tau := \quad \frac{1}{N_n}\left|\sum_{i=1}^{N_n} Z_i \tau_i\right|$$ ***overlap*** of the estimator

$\mathcal{O}_\tau :=$| Fraction of correctly classified nodes - Fraction of incorrectly classified nodes |

Community Detection Problem



Given G_n and $\{X_i\}_{i\in[0,N_n]}$, estimate $\{Z_i\}_{i\in[1,N_n]}$

$\{\tau_i\}_{i\in[0,N_n]}$- Community estimates

$$\mathcal{O}_\tau := \frac{1}{N_n}\left|\sum_{i=1}^{N_n} Z_i\tau_i\right|$$ ***overlap*** of the estimator

$\mathcal{O}_\tau :=$| Fraction of correctly classified nodes - Fraction of incorrectly classified nodes |

Community Detection is ***solvable*** if there exists an estimator $\{\tau_i\}_{i\in[0,N_n]}$ for every n, and some $\gamma > 0$ s.t. $\lim_{n\to\infty}\mathbb{P}[\mathcal{O}_\tau > \gamma] = 1$

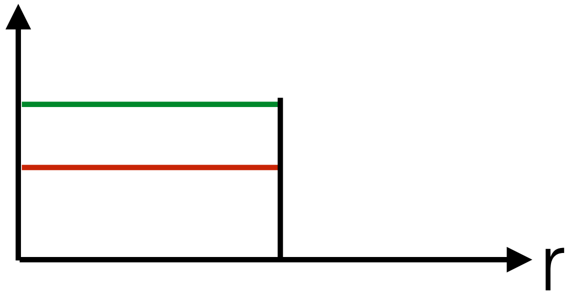
SLLN gives $\sum_{I=1}^{N_n}\frac{\tau_i Z_i}{N_n} \to 0$ *for blind guessing*

Solvability $\approx$ *asymptotically beating a random guess*

Community Detection Problem

Consider the example $f_{in}(r) = a\mathbf{1}_{r \leq R}$ $\qquad f_{out}(r) = b\mathbf{1}_{r \leq R}$

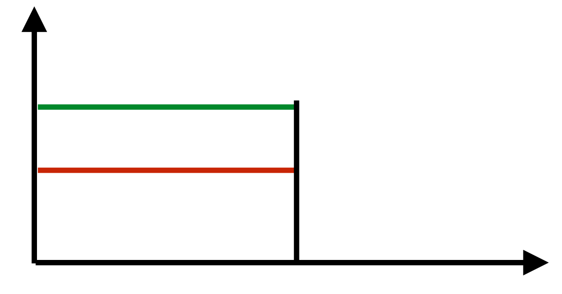
$$0 \leq b < a \leq 1$$

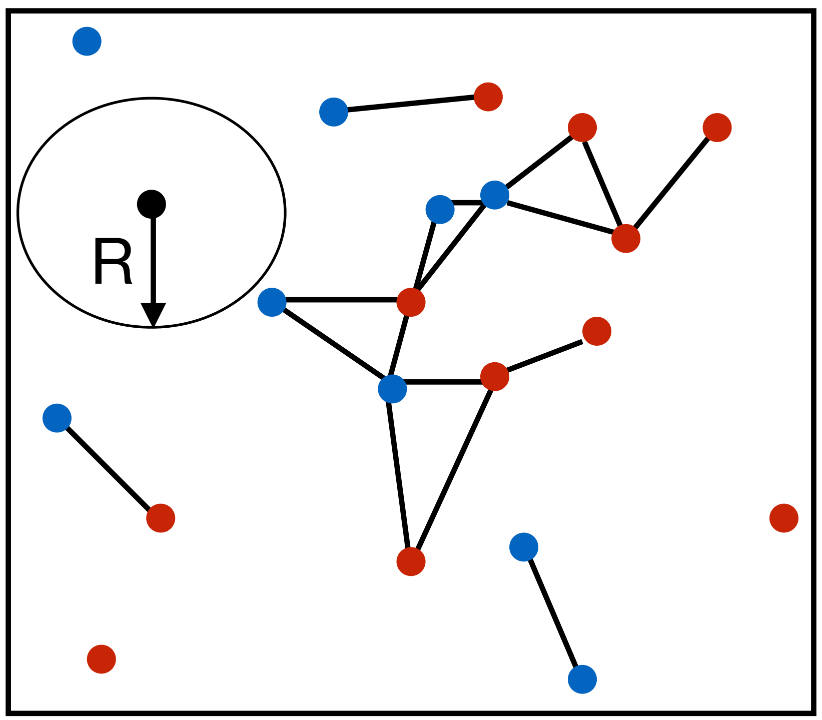


Community Detection Problem

Consider the example $f_{in}(r) = a\mathbf{1}_{r \leq R}$ $\quad f_{out}(r) = b\mathbf{1}_{r \leq R}$

$$0 \leq b < a \leq 1$$





Isolated Nodes = No interaction with other points

Clearly $\mathcal{O}_\tau \leq 1 - e^{-\lambda \nu_d(1) R^d} < 1$

$\nu_d(1)$ Vol of unit ball in d dimensions

Solvability Phase Transition

An overlap of γ is ***achievable*** if there exists an estimator $\{\tau_i\}_{i=1}^{N_n}$ such that $\lim_{n\to\infty} \mathbb{P}[\mathcal{O}_\tau > \gamma] = 1$

Solvability Phase Transition

An overlap of γ is ***achievable*** if there exists an estimator $\{\tau_i\}_{i=1}^{N_n}$ such that $\lim_{n\to\infty} \mathbb{P}[\mathcal{O}_\tau > \gamma] = 1$

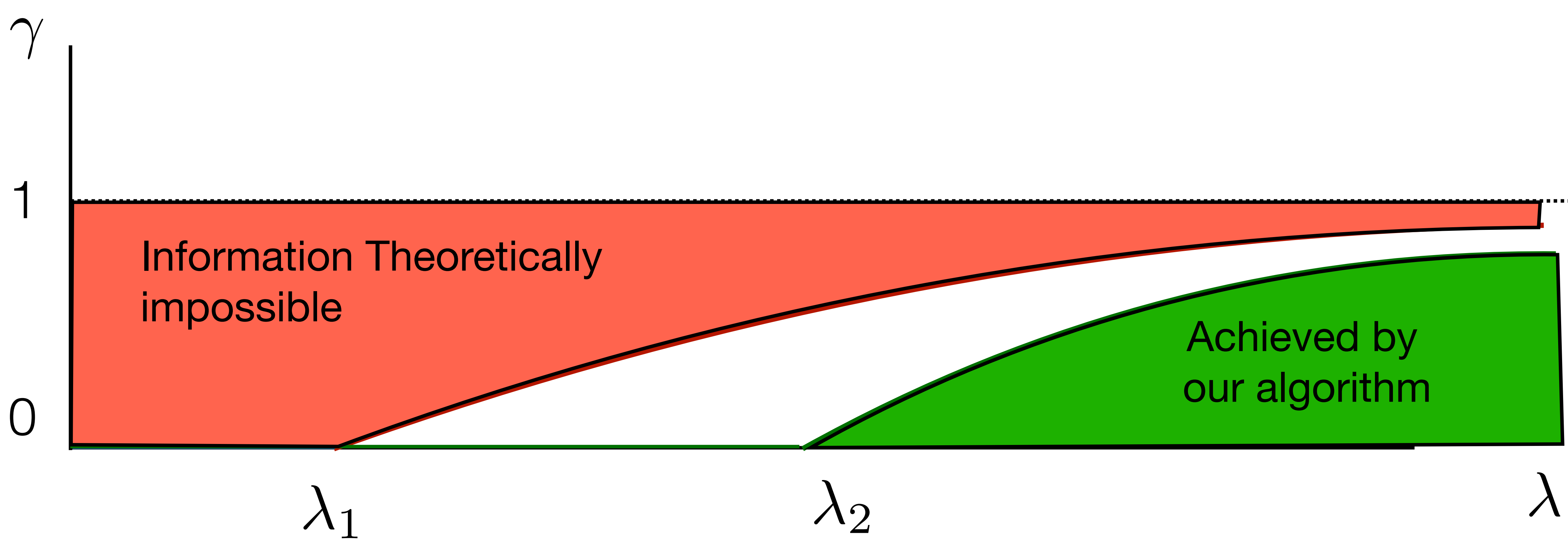
Solvability iff any $\gamma > 0$ is achievable

Solvability Phase Transition

An overlap of γ is ***achievable*** if there exists an estimator $\{\tau_i\}_{i=1}^{N_n}$

such that $\lim_{n\to\infty} \mathbb{P}[\mathcal{O}_\tau > \gamma] = 1$

Solvability iff any $\gamma > 0$ is achievable



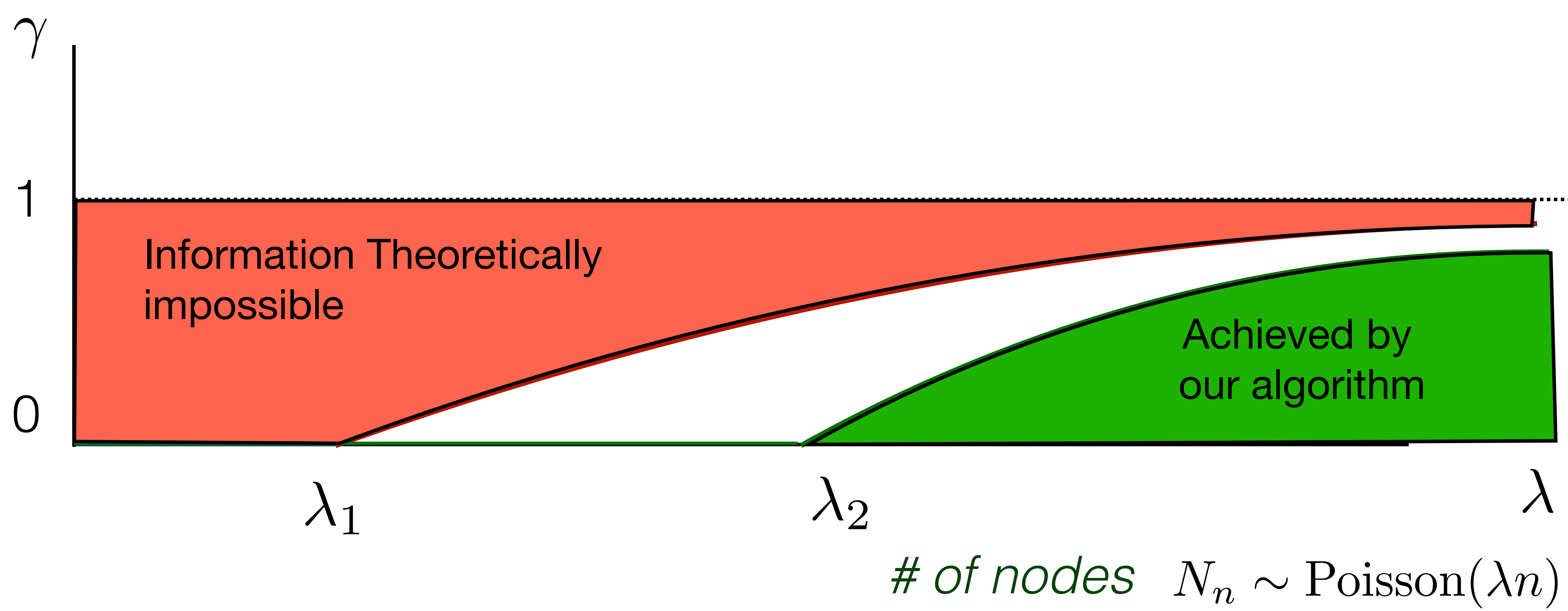
# of nodes $N_n \sim \mathrm{Poisson}(\lambda n)$

Solvability Phase Transition

<u>Theorem</u> - $\forall f_{in}(\cdot), f_{out}(\cdot), d \geq 2,\ \exists\ 0 < \lambda_1 \leq \lambda_2 < \infty$ such that -

$\lambda < \lambda_1 \implies$ Community Detection is not solvable

$\lambda > \lambda_2 \implies$ Our algorithm solves Community Detection efficiently



Our algorithm is *asymptotically optimal*.

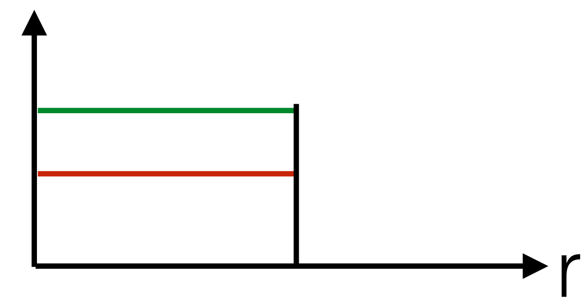
Algorithm Idea

Algorithm Idea

Spatial graph - *Locally dense* but *globally sparse*

Algorithm Idea

Spatial graph - *Locally dense* but *globally sparse*



Consider the example $f_{in}(r) = a\mathbf{1}_{r \leq R}$, $f_{out}(r) = b\mathbf{1}_{r \leq R}$

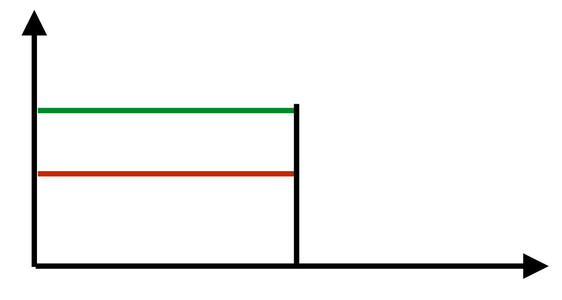
$$0 \leq b < a \leq 1$$

Algorithm Idea

Spatial graph - *Locally dense* but *globally sparse*

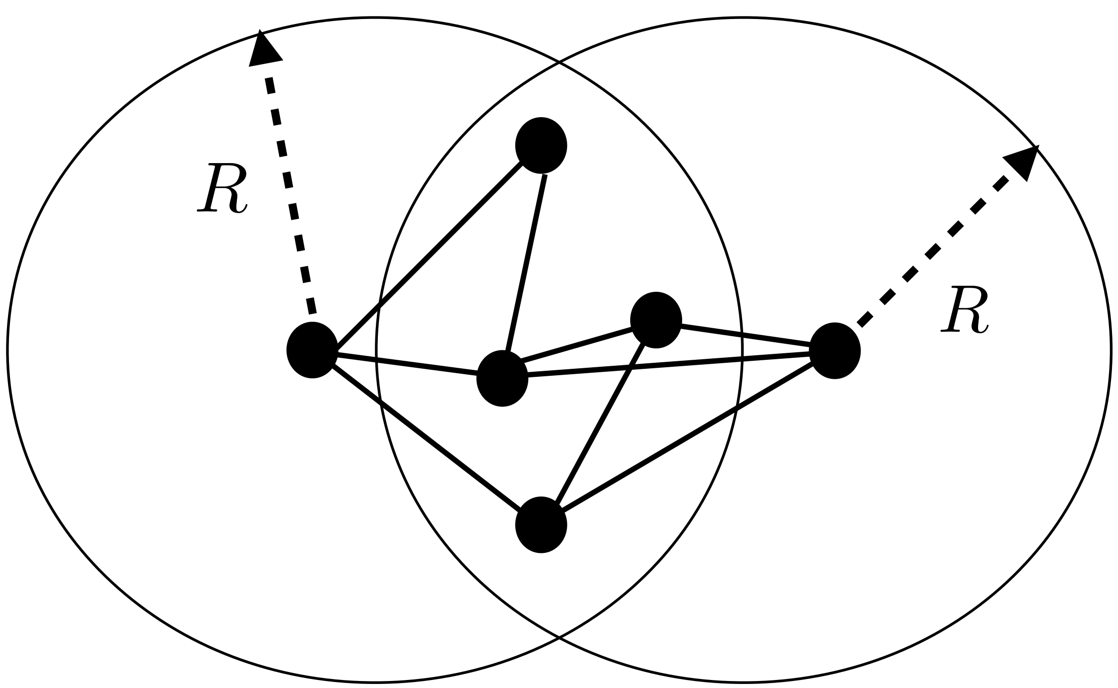
Consider the example $f_{in}(r) = a\mathbf{1}_{r \leq R}$, $f_{out}(r) = b\mathbf{1}_{r \leq R}$

$$0 \leq b < a \leq 1$$

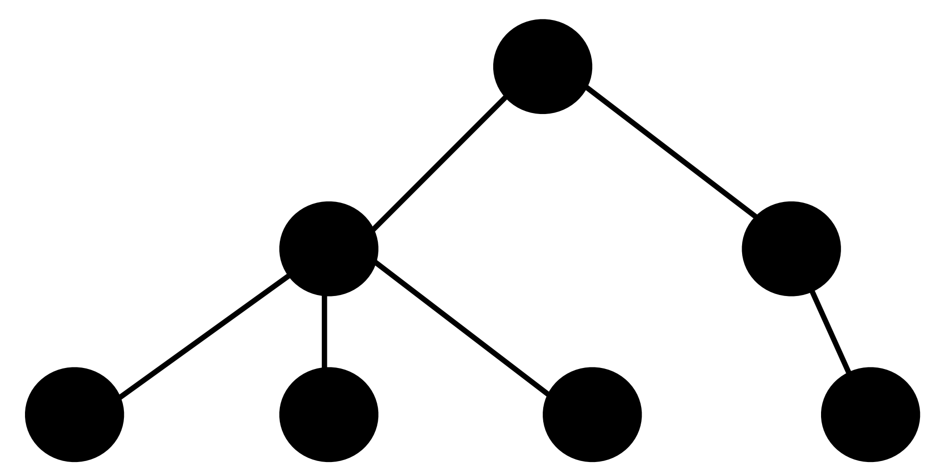


<u>*Locally Dense*</u> - '*Nearby*' nodes connect with *constant probability* independent of n

<u>*Globally Sparse*</u> - Order n edges in total

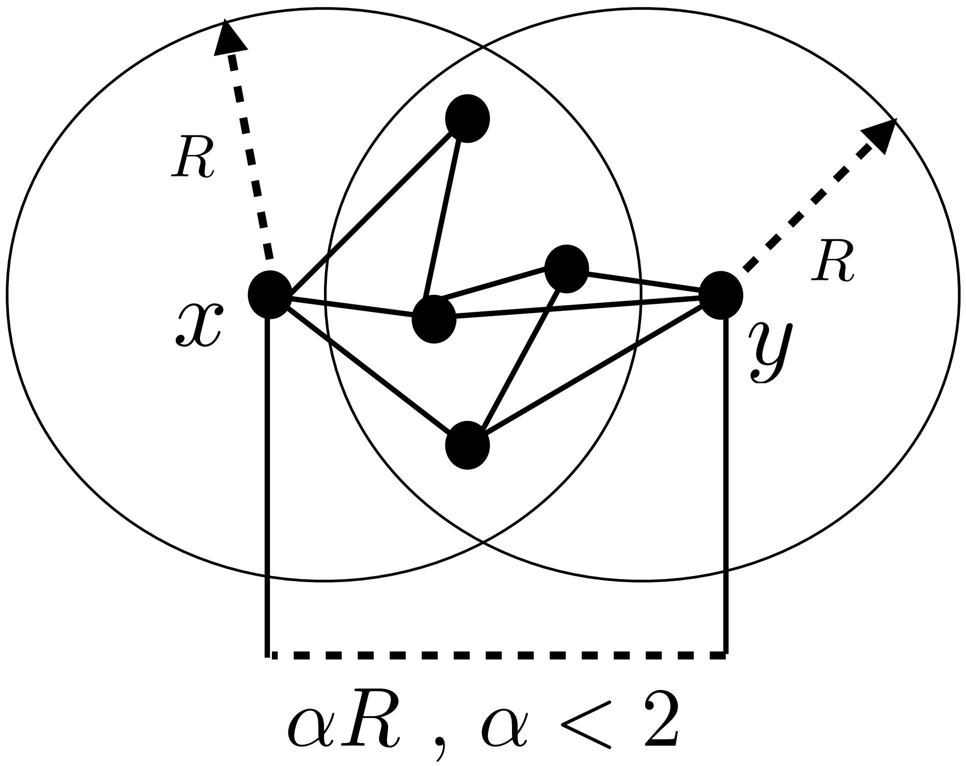


Spatial Graph



SBM

Algorithm Idea

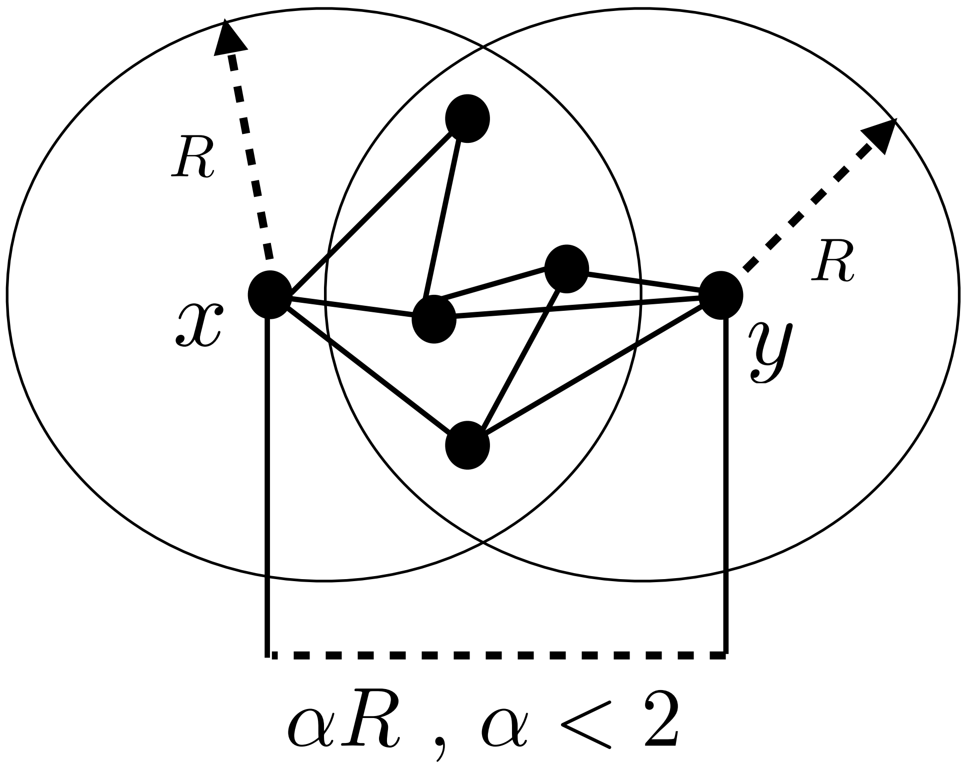


\# common neighbors is Poisson with mean

Same community - $\lambda c(\alpha) R^d \left(\frac{a^2 + b^2}{2} \right)$

Opposite communities - $\lambda c(\alpha) R^d ab$

Algorithm Idea



\# common neighbors is Poisson with mean

Same community - $\lambda c(\alpha) R^d \left(\frac{a^2 + b^2}{2} \right)$

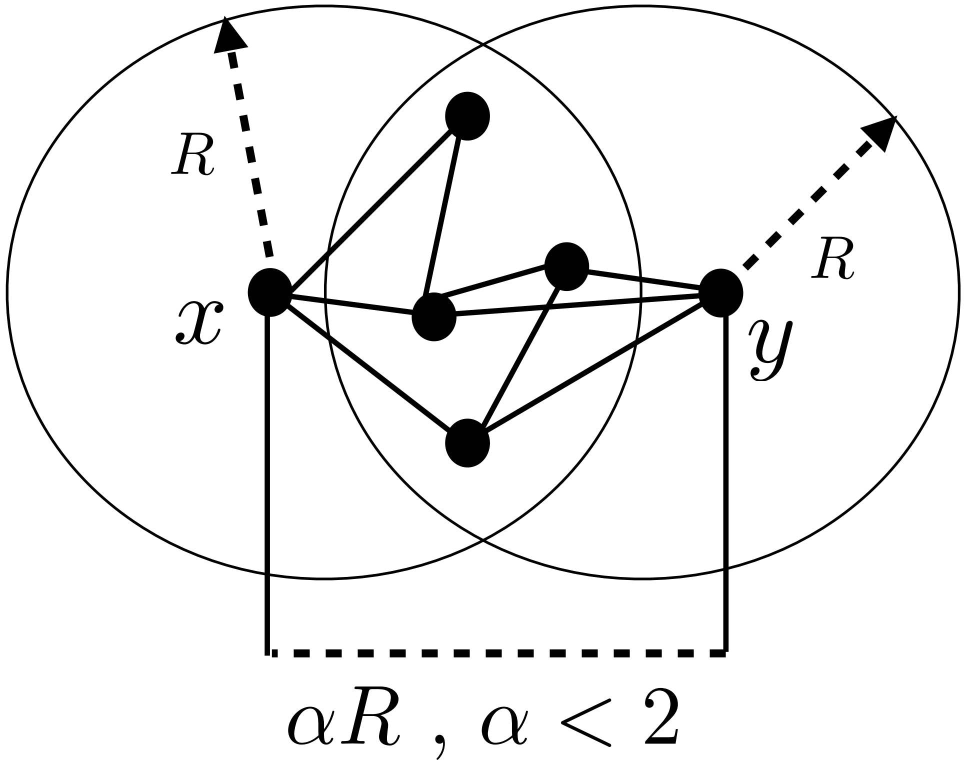
Opposite communities - $\lambda c(\alpha) R^d ab$

Set threshold - $T(\alpha) = c(\alpha) R^d \lambda \left(\frac{a+b}{2} \right)^2$

Pairwise-Classify(x,y)

- IF # (common neighbors) < $T(\alpha)$, *DECLARE* community(x) $\neq$ community(y)
- ELSE *DECLARE* community(x) $=$ community(y)

Algorithm Idea



common neighbors is Poisson with mean

Same community - $\lambda c(\alpha) R^d \left(\frac{a^2+b^2}{2}\right)$

Opposite communities - $\lambda c(\alpha) R^d ab$

Set threshold - $T(\alpha) = c(\alpha) R^d \lambda \left(\frac{a+b}{2}\right)^2$

Pairwise-Classify(x,y)

- IF # (common neighbors) < $T(\alpha)$, *DECLARE* community(x) $\neq$ community(y)
- ELSE *DECLARE* community(x) $=$ community(y)

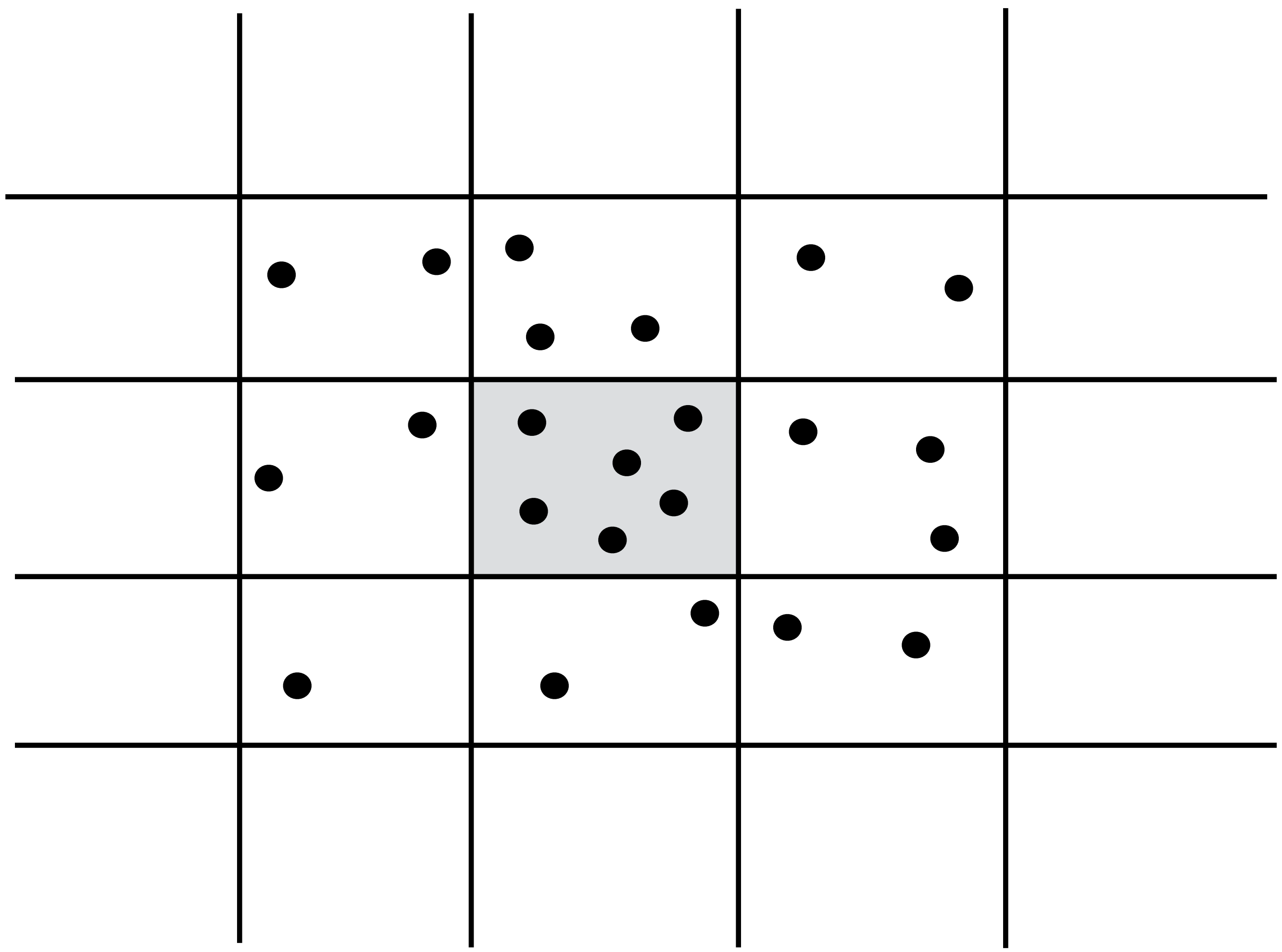
Chernoff bound -

P(Mis-classifying a given pair of nodes at distance αR) $\leq e^{-\lambda c'(\alpha) R}$

Algorithm Idea

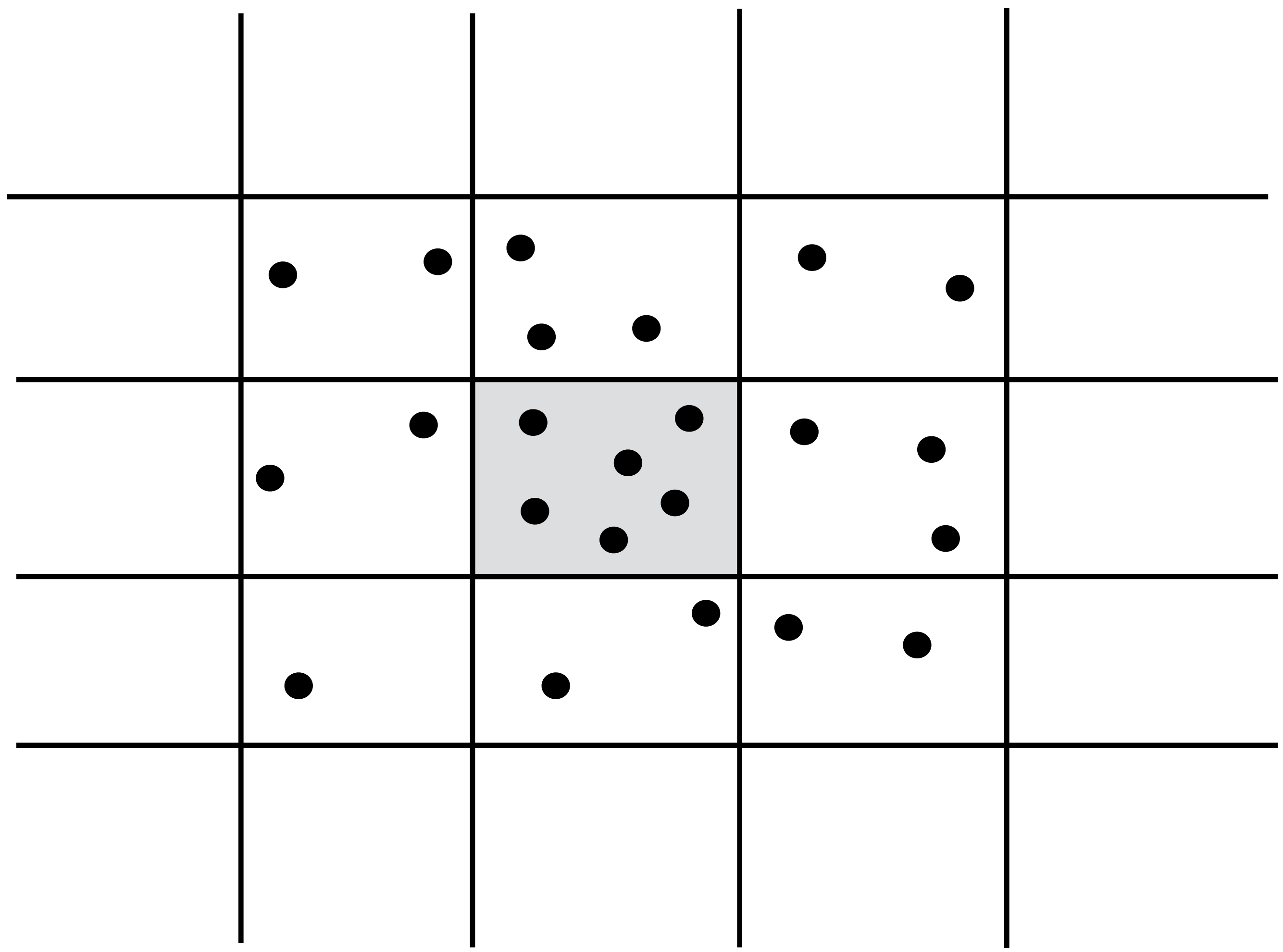
Tesselate $\mathbb{R}^d$ into grids of side $R/4$

Classify cells to be Good or Bad



Algorithm Idea

Tesselate $\mathbb{R}^d$ into grids of side $R/4$

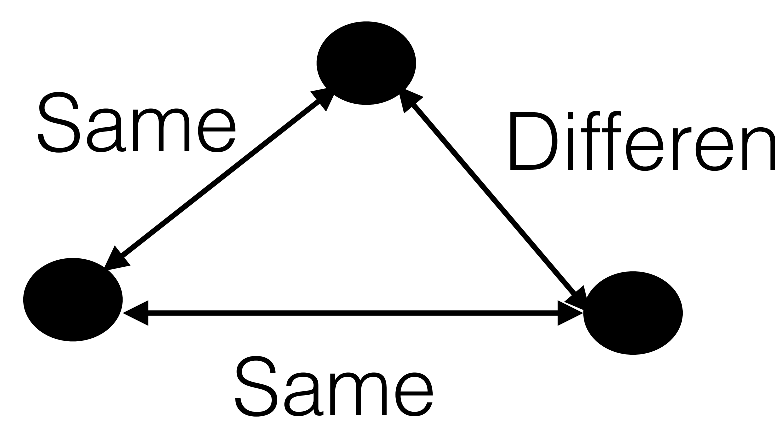


Classify cells to be Good or Bad

Cell ***<u>Good</u>*** if

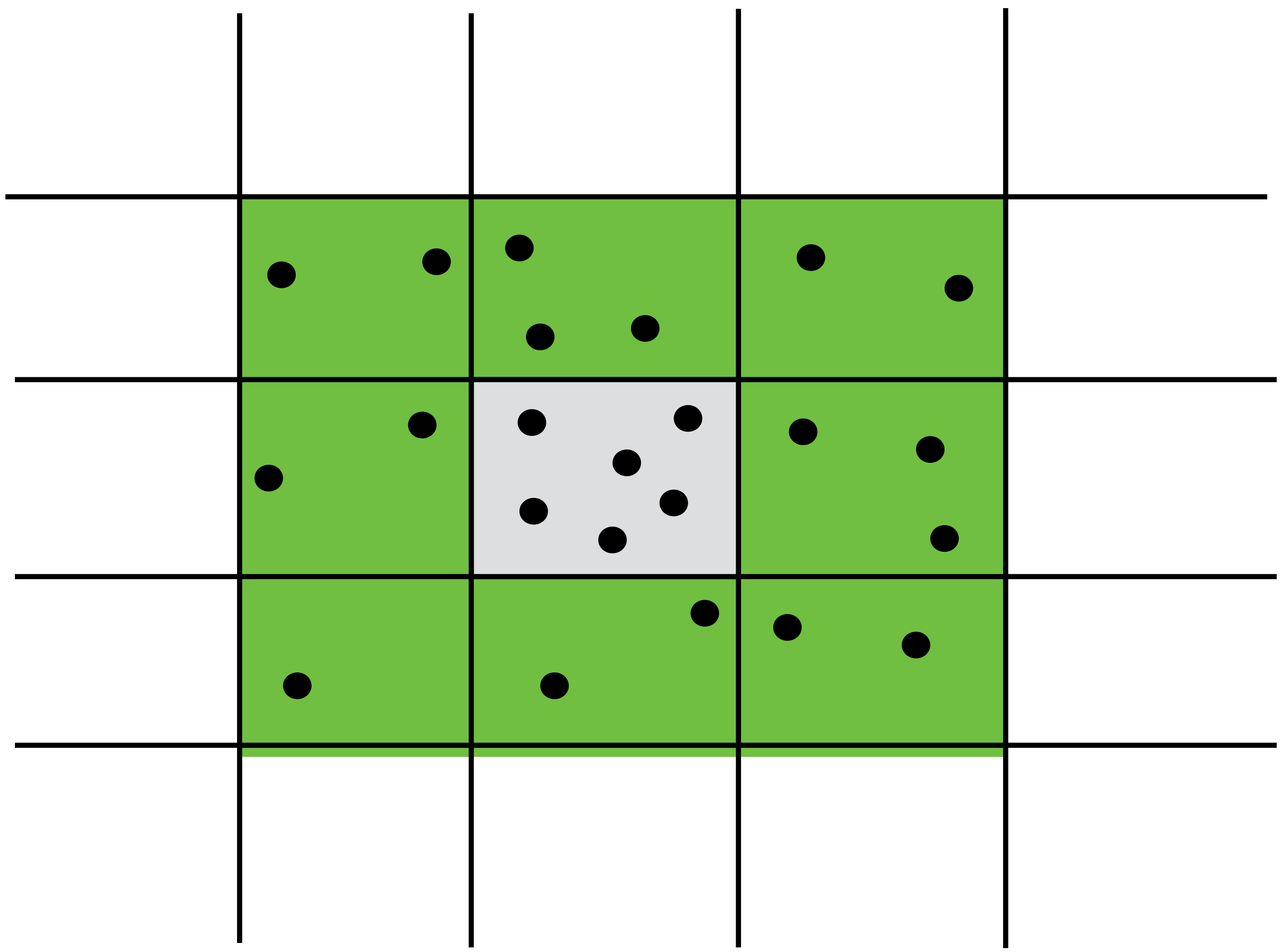
1. At-least $(1-\epsilon)$ Mean # of nodes
2. No ***inconsistencies*** in pairwise checks *with all neighboring cells*

Example of Inconsistent output



Algorithm Idea

Tesselate $\mathbb{R}^d$ into grids of side $R/4$

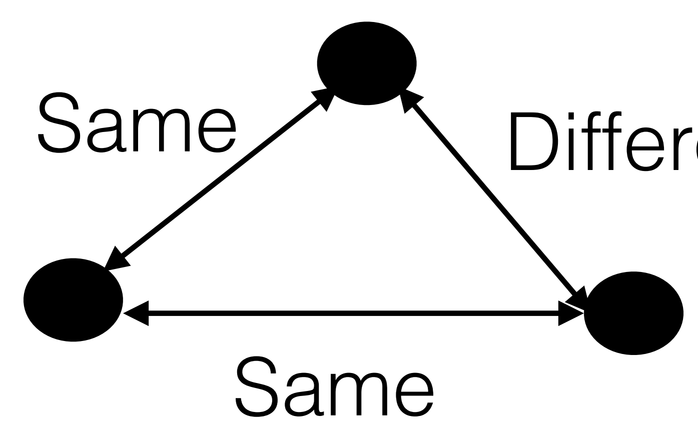


Classify cells to be Good or Bad

Cell ***<u>Good</u>*** if

1. At-least $(1-\epsilon)$ Mean # of nodes
2. No ***inconsistencies*** in pairwise checks *with all neighboring cells*

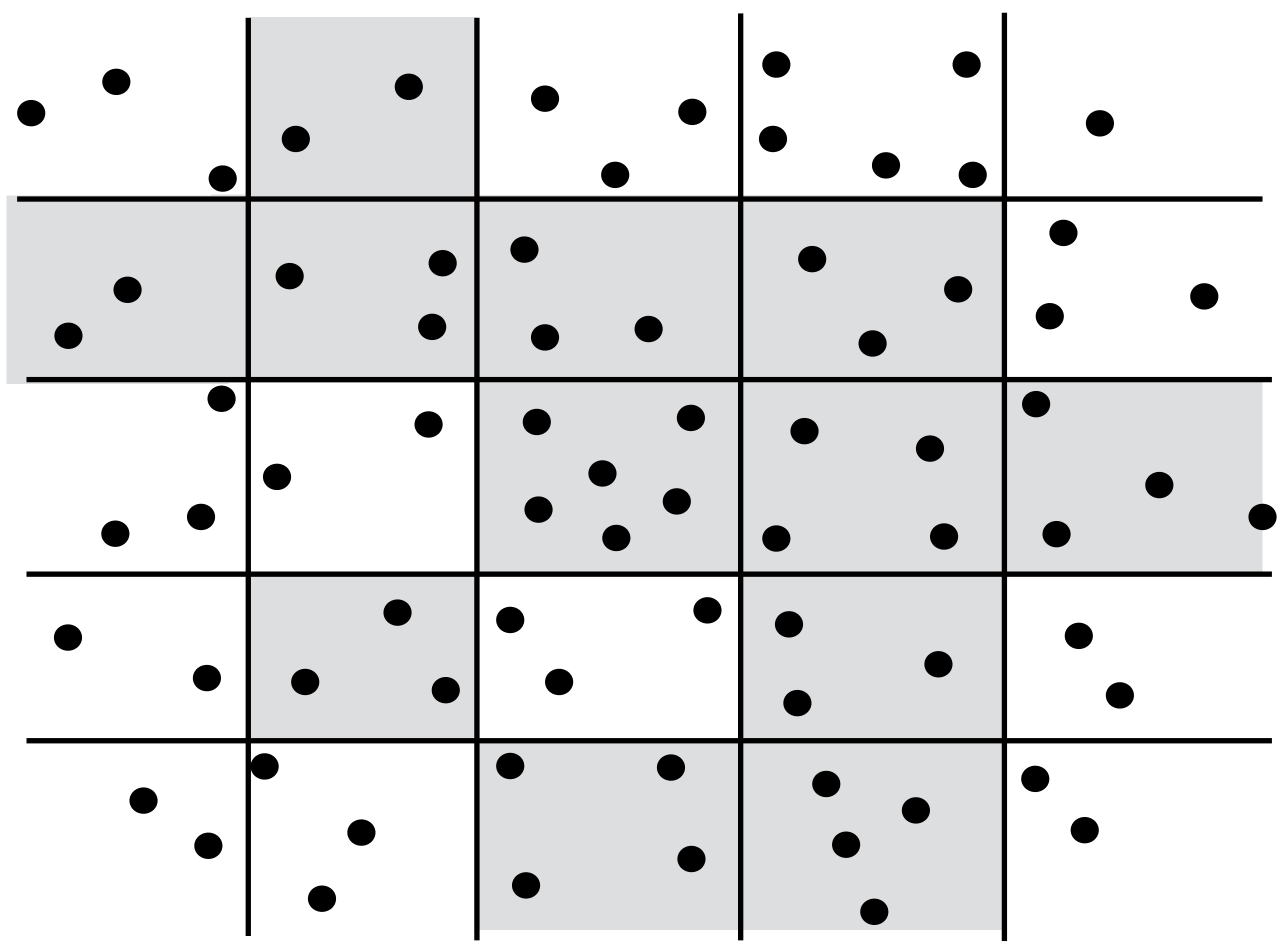
Example of Inconsistent output



Algorithm Idea

Main Routine

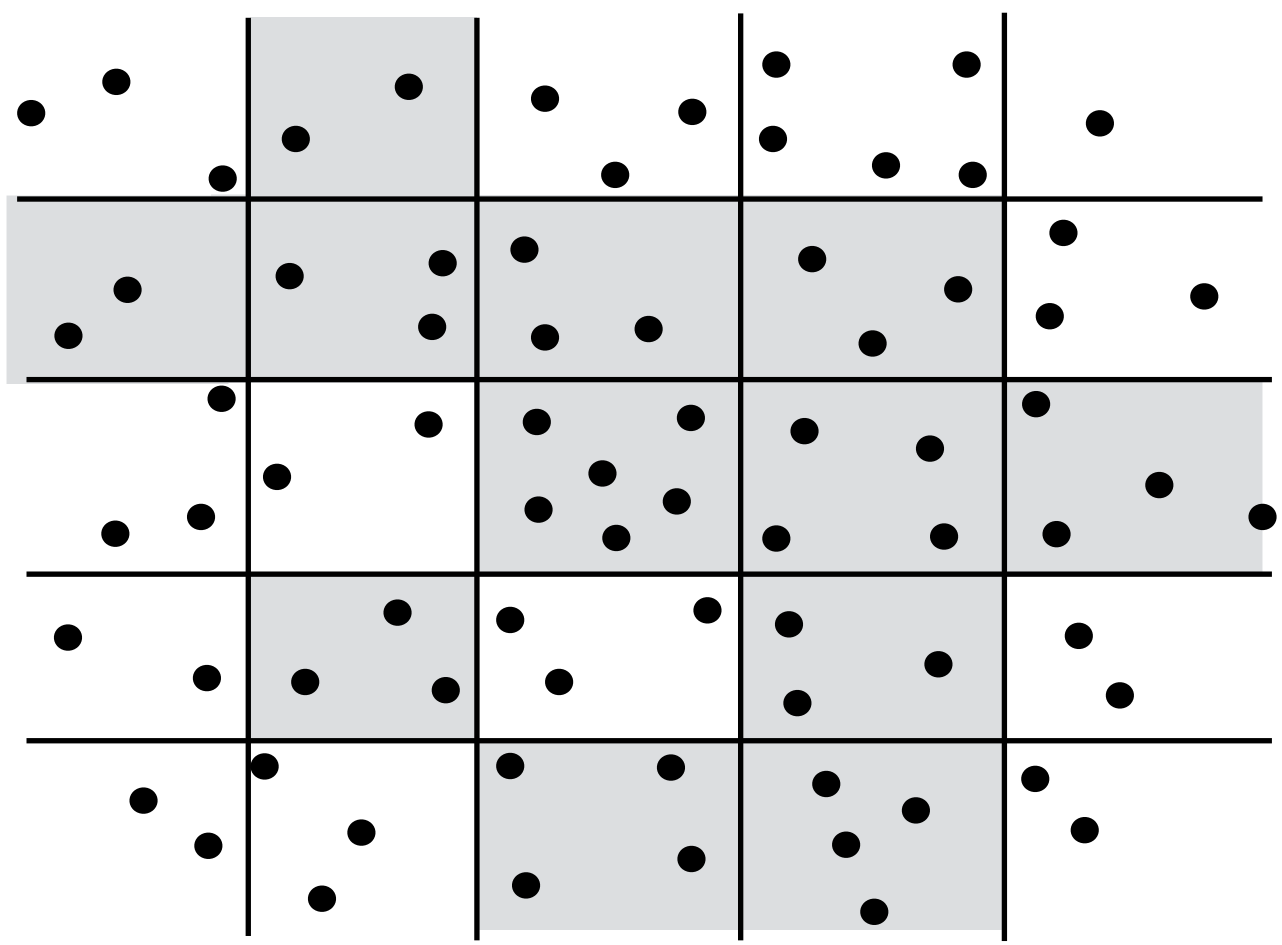
- Partition each good component with Pairwise-Classify
- Output +1 estimate to all nodes in bad cells



Algorithm Idea

Main Routine

- Partition each good component with Pairwise-Classify
- Output +1 estimate to all nodes in bad cells



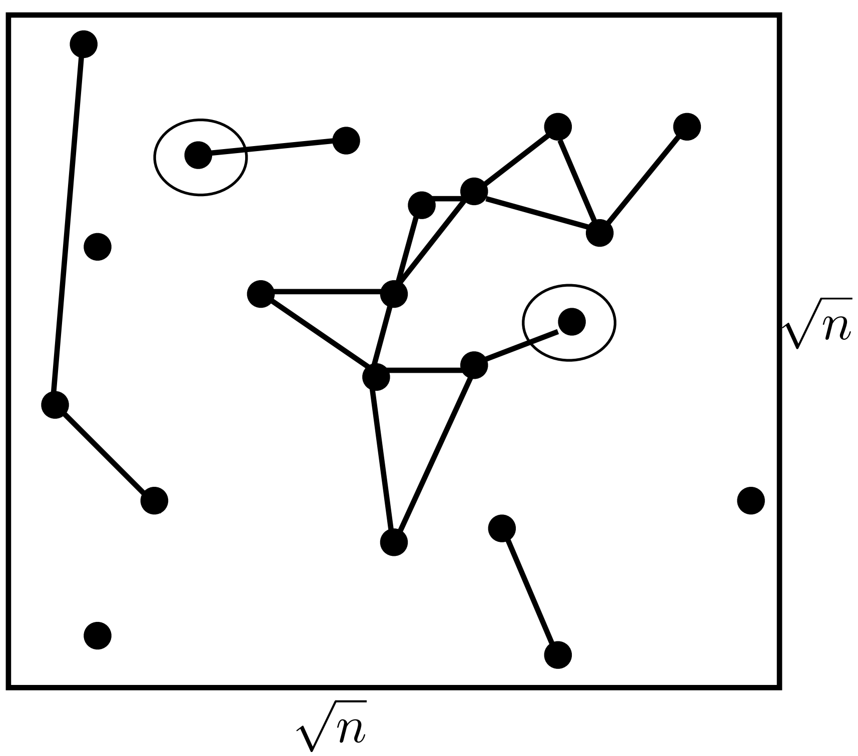
Algorithm succeeds if a “large” connected component of “gray” cells is present

A k-Dependent Percolation Process. [Liggett, Schonmann, Stacey, ’97]

Impossibility

Easier problem -

Given the data $(G, \{X_i\}_{i \in [1, N_n]})$, can you classify ***any two randomly chosen nodes*** better than chance.



Community Detection is solvable if the above

can be solved with success probability at-least

$$\frac{1+\gamma}{2}$$

(Cluster the whole graph and then answer)

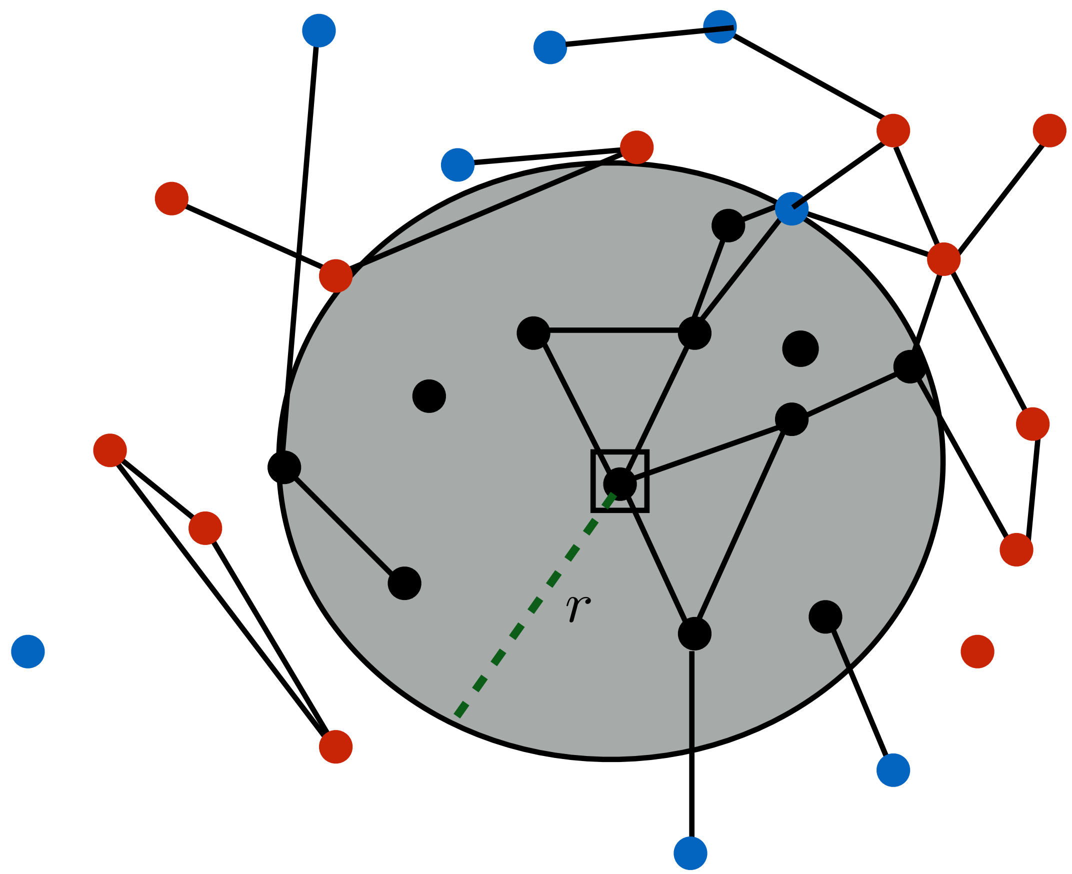
Will prove that the above is not solvable for small λ

Impossibility

W.h.p. - distance between the two chosen nodes is '***large***'

An easier problem

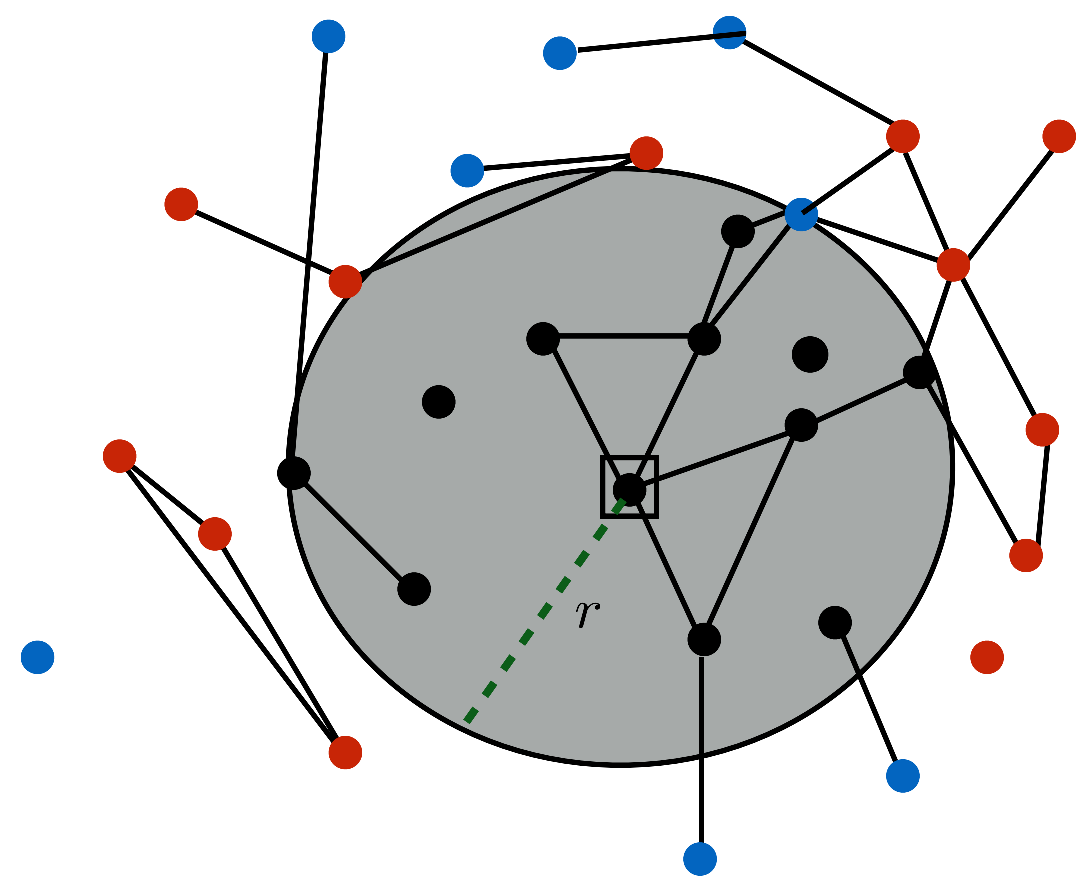
Estimate better than chance, the community label of a random node given community labels of all "far away" nodes.



Information Flow from Infinity Problem

Does $\exists \gamma^{'} > 0$ and $\tau_0^{'} \in \{-1, +1\}$ as a measurable function of $G, \{X_i\}_{i \in \mathbb{N}}, \{Z_i : ||X_i|| > r\}$ such that $\liminf_{r \to \infty} \mathbb{P}^0[\tau_0^{'} = Z_0] \geq \frac{1}{2} + \gamma^{'}$?

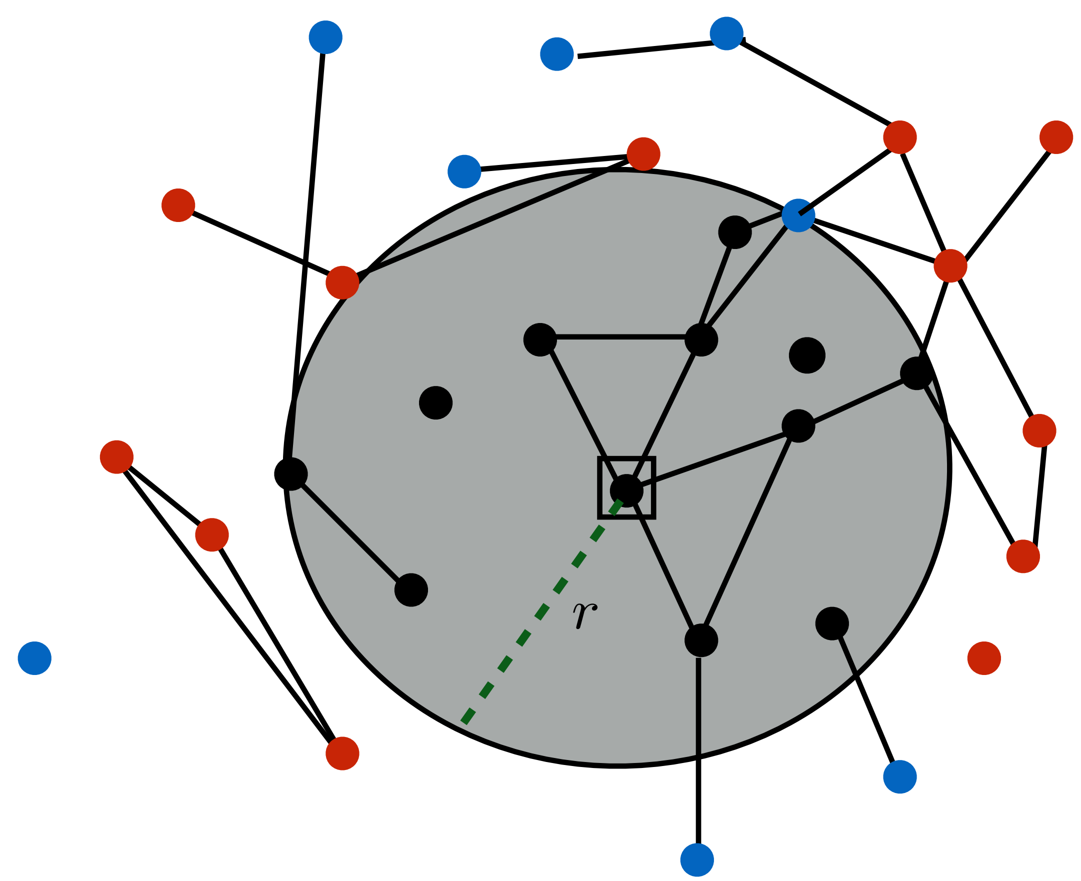
If answer above is NO, then by classical ergodic arguments Community Detection is not solvable.



Information Flow from Infinity Problem

Does $\exists \gamma^{'} > 0$ and $\tau_0^{'} \in \{-1, +1\}$ as a measurable function of $G, \{X_i\}_{i \in \mathbb{N}}, \{Z_i : ||X_i|| > r\}$ such that $\liminf_{r \to \infty} \mathbb{P}^0[\tau_0^{'} = Z_0] \geq \frac{1}{2} + \gamma^{'}$?

Theorem - If the random spatial graph with intensity λ and connection function $f_{in}(\cdot) - f_{out}(\cdot)$ ***does not percolate***, then the answer to the above question is NO.



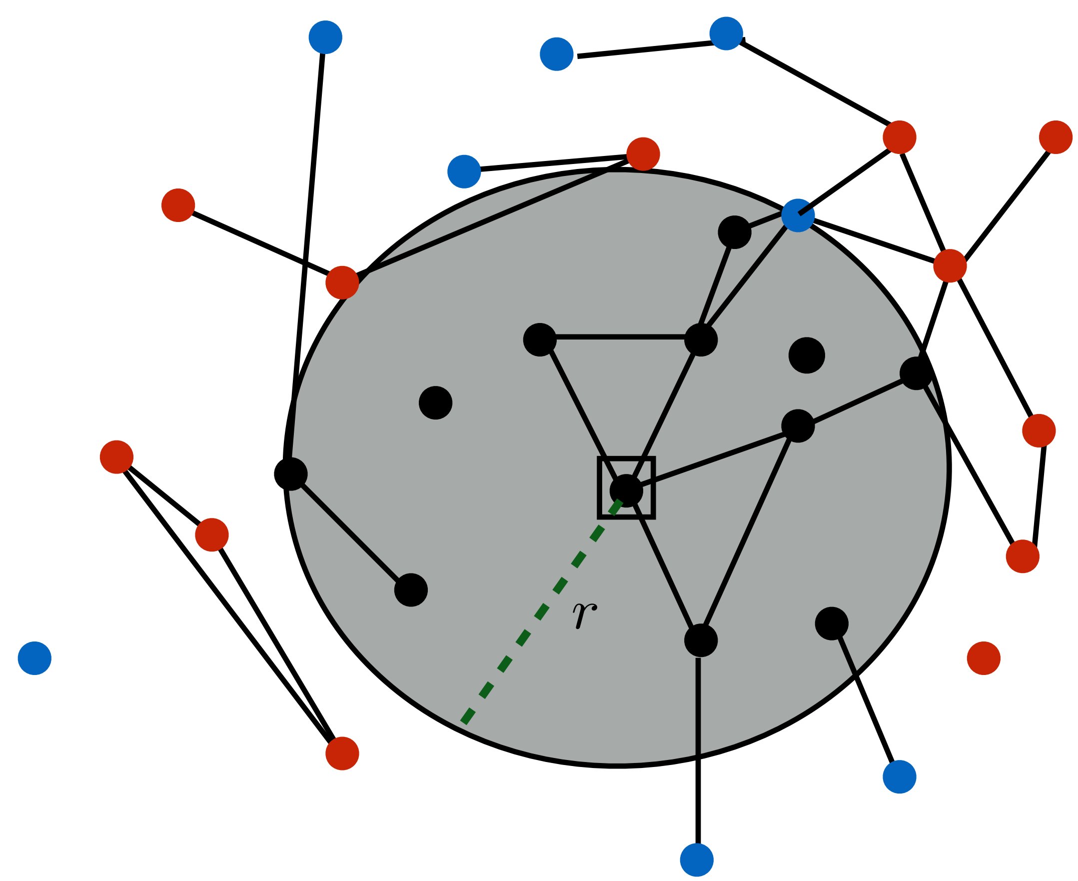
Information Flow from Infinity Problem

Does $\exists \gamma^{'} > 0$ and $\tau_0^{'} \in \{-1,+1\}$ as a measurable function of $G, \{X_i\}_{i \in \mathbb{N}}, \{Z_i : ||X_i|| > r\}$ such that $\liminf_{r \to \infty} \mathbb{P}^0[\tau_0^{'} = Z_0] \geq \frac{1}{2} + \gamma^{'}$?

Theorem - If the random spatial graph with intensity λ and connection function $f_{in}(\cdot) - f_{out}(\cdot)$ ***does not percolate***, then the answer to the above question is NO.

Corollary

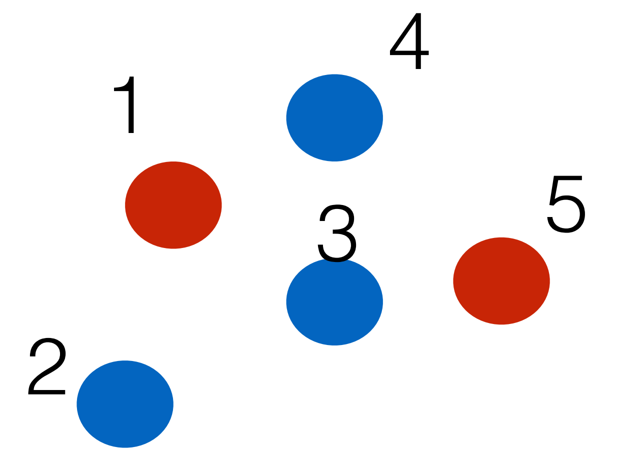
1. If $d = 1$, then community detection is not solvable for any $\lambda, f_{in}(\cdot), f_{out}(\cdot)$.



Information Flow from Infinity Problem

Enriched probability space.

1) Sample the location labels and community labels as before.

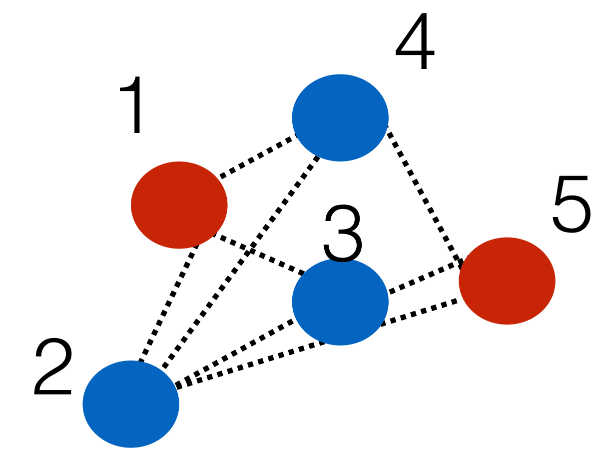


Information Flow from Infinity Problem

Enriched probability space.

1) Sample the location labels and community labels as before.

2) $\{U_{ij}\}_{i<j\in\mathbb{N}}$ - i.i.d. $U[0,1]$ RVs.
every pair $i<j\in\mathbb{N}$ nodes, marked with U_{ij}

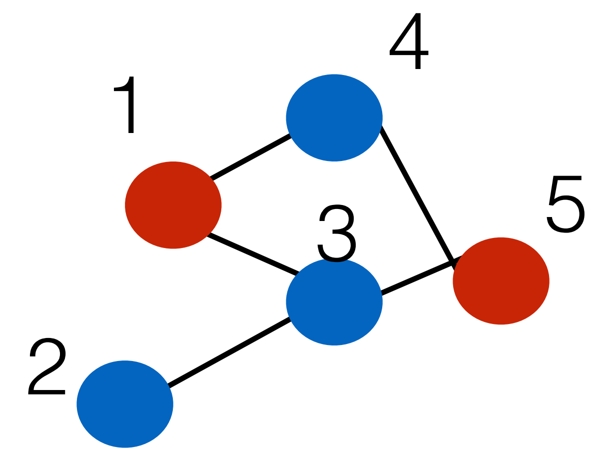


Information Flow from Infinity Problem

Enriched probability space.

1) Sample the location labels and community labels as before.

2) $\{U_{ij}\}_{i<j\in\mathbb{N}}$ - i.i.d. $U[0,1]$ RVs.
   every pair $i < j \in \mathbb{N}$ nodes, marked with U_{ij}



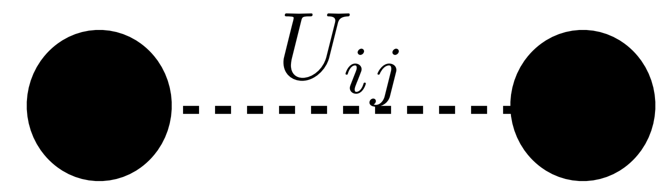
3) An edge between $i < j \in \mathbb{N}$ iff

$$U_{ij} \leq f_{in}(||X_i - X_j||)\mathbf{1}_{Z_i=Z_j} + f_{out}(||X_i - X_j||)\mathbf{1}_{Z_i\neq Z_j}$$

Information Flow from Infinity Problem

$\{U_{ij}\}_{i<j\in\mathbb{N}}$, -i.i.d. $U[0,1]$ sequence.
Edge between $i<j\in\mathbb{N}$ iff

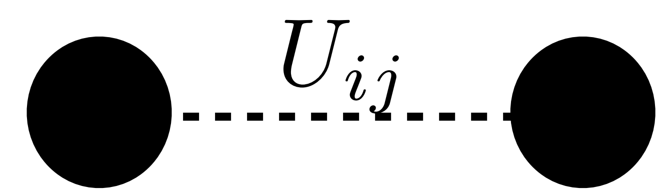
$$U_{ij} \le f_{in}(||X_i - X_j||)\mathbf{1}_{Z_i=Z_j} + f_{out}(||X_i - X_j||)\mathbf{1}_{Z_i\neq Z_j}$$



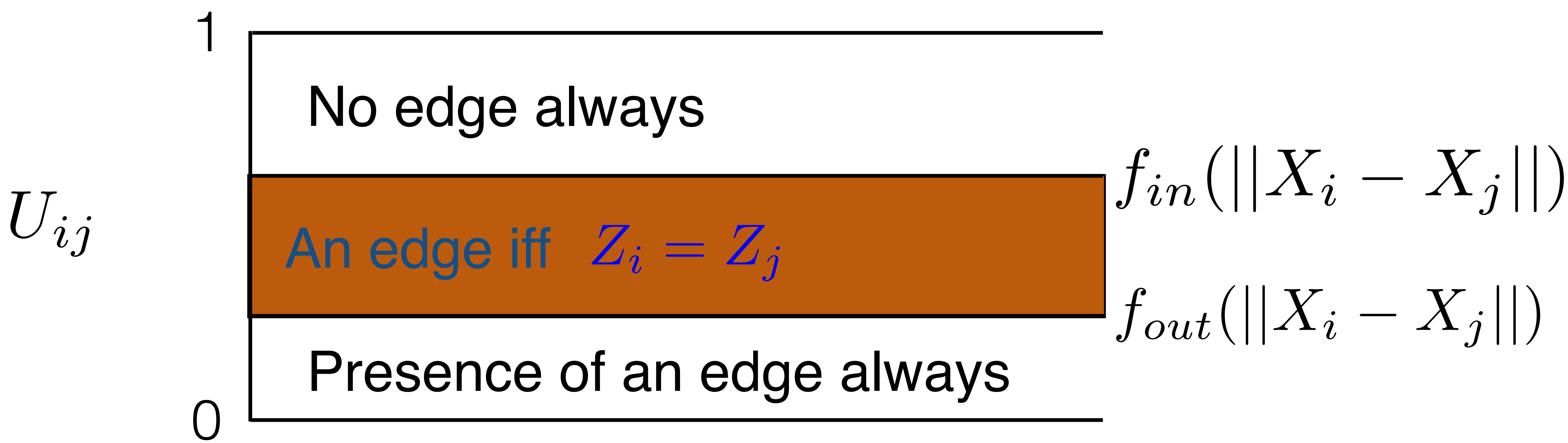
Information Flow from Infinity Problem

$\{U_{ij}\}_{i<j\in\mathbb{N}}$, -i.i.d. $U[0,1]$ sequence.
Edge between $i<j\in\mathbb{N}$ iff

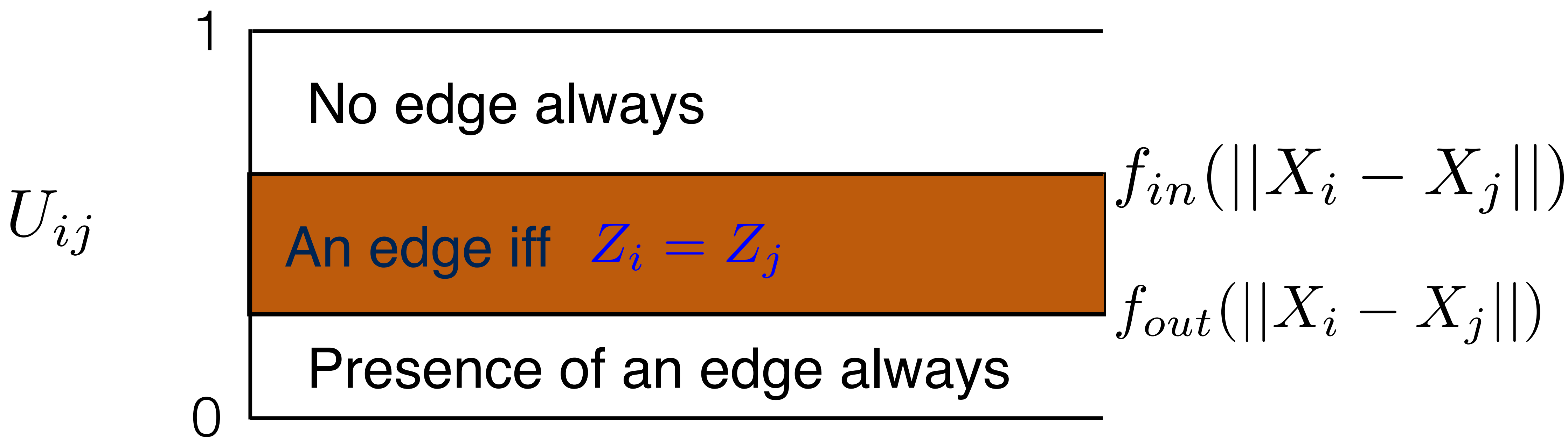
$$U_{ij} \leq f_{in}(||X_i - X_j||)\mathbf{1}_{Z_i=Z_j} + f_{out}(||X_i - X_j||)\mathbf{1}_{Z_i\neq Z_j}$$



<u>Only certain edges are *Informative*</u>



Information Flow from Infinity Problem



Create an *Information Graph* I from $\{X_i\}_{i \in \mathbb{N}}$ and $\{U_{ij}\}_{i<j \in \mathbb{N}}$

$$i \sim_I j \iff f_{out}(||X_i - X_j||) < U_{ij} \leq f_{in}(||X_i - X_j||)$$

<u>Structural Lemma -</u>

$$i \sim_I j, i \sim_G j \implies Z_i = Z_j$$

$$i \sim_I j, i \nsim_G j \implies Z_i \neq Z_j$$

Extend to connected components of I instead of just edges.

Information Flow from Infinity Problem

$V_I(0) \subset \mathbb{N}$ - Set of nodes in the connected component of origin in I.

<u>Lemma</u> - On the event $|V_I(0)| < \infty$,

$$\mathbb{P}^0 \left[Z_0 = +1 \middle| G, \{U_{ij}\}_{i<j}, \{X_i\}_{i \in \mathbb{N}}, \{Z_k\}_{k \in V_I^{\complement}(0)} \right] = \frac{1}{2} \text{ a.s.}$$

Community labels on disconnected components of I are independent.

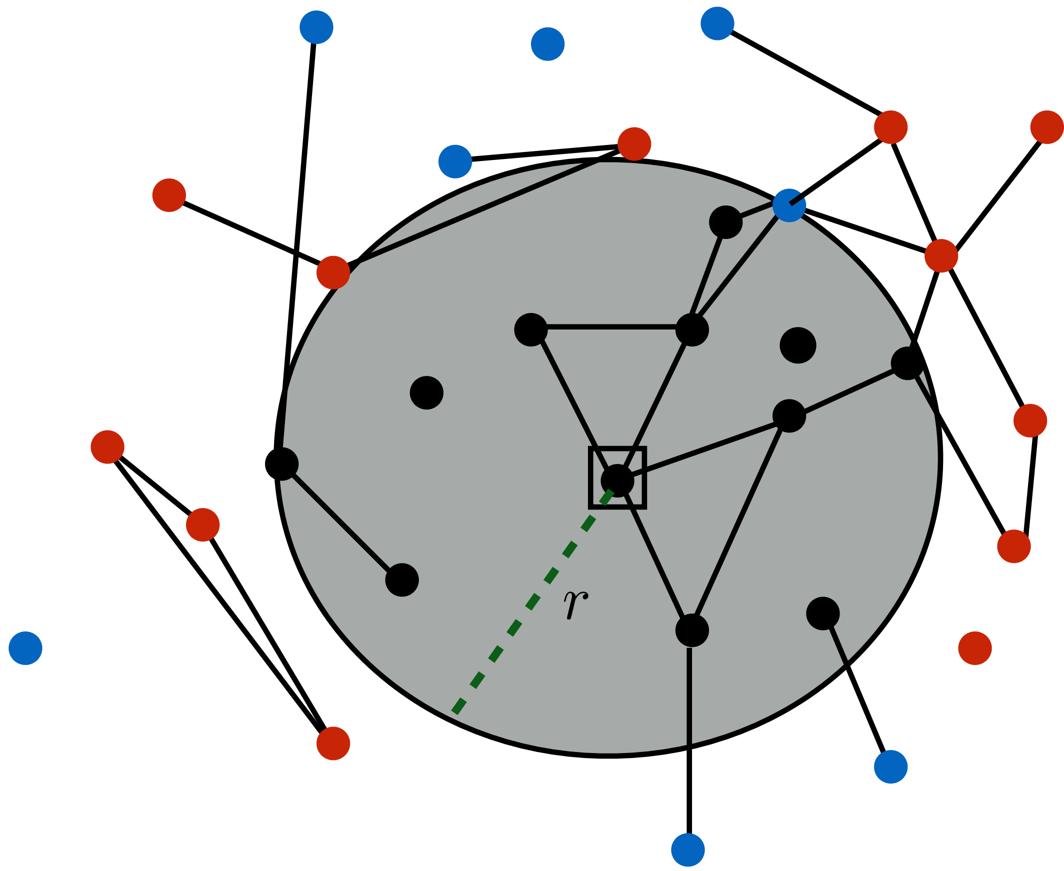
Proof - Bayes' rule along with the previous structural observation.

Information Flow from Infinity Problem

On the event $|V_I(0)| < \infty$, no estimator for the community label at origin can beat a random guess for large enough r.

<u>Corollary</u>

If $|V_I(0)| < \infty$ a.s. , i.e. if I does not percolate, then cannot solve the Information Flow from Infinity Problem.



Information Flow from Infinity Problem

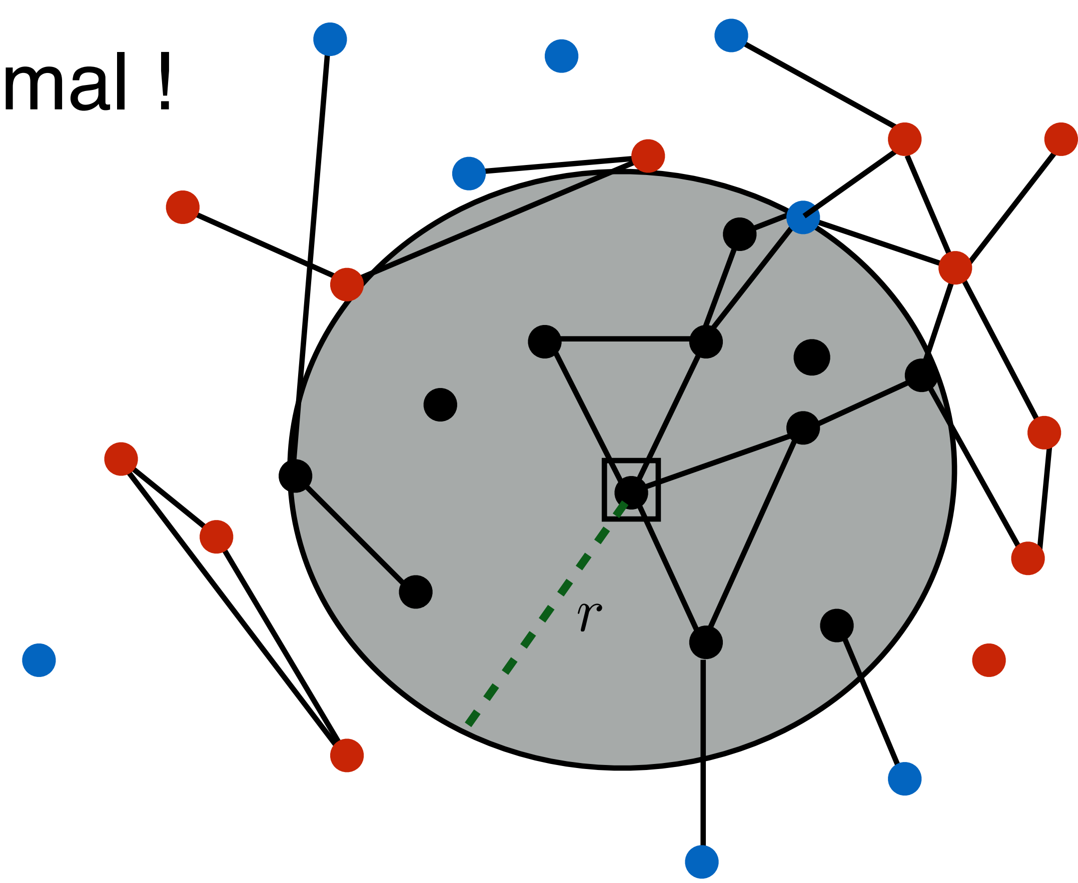
The Key Idea -

Reduce to a percolation criteria.

Labels on different components are independent.

[Mossel, ’00],[Lubetzky, Sly, ’14],
[Abbe,Massoulié,Montanari,Sly,Srivastava,’17]

Drawbacks   Our method is provably sub-optimal !



Recent methods that improve this technique.

[Polyanskiy, Wu, ’18][Abbe, Boix, ‘18]

Distinguishability - Are there communities ?

Determine whether the data $\{X_i\}_{i\in\mathbb{N}}, G$ is sampled from

1) The planted model with connection functions $f_{in}(\cdot)$ and $f_{out}(\cdot)$

2) $H_{\lambda,g(\cdot),d}$ - a model without planted communities.

Distinguishability - Are there communities ?

Determine whether the data $\{X_i\}_{i \in \mathbb{N}}, G$ is sampled from

1) The planted model with connection functions $f_{in}(\cdot)$ and $f_{out}(\cdot)$

2) $H_{\lambda, g(\cdot), d}$ - a model without planted communities.

<u>Theorem</u> - The induced measure by $H_{\lambda, g(\cdot), d}$ is mutually singular with respect to that by G for any λ , $f_{in}(\cdot), f_{out}(\cdot)$ and $g(\cdot)$ where $f_{in} \neq f_{out}$ a.e.

Distinguishability - Are there communities ?

Determine whether the data $\{X_i\}_{i \in \mathbb{N}}, G$ is sampled from

1) The planted model with connection functions $f_{in}(\cdot)$ and $f_{out}(\cdot)$

2) $H_{\lambda, g(\cdot), d}$ - a model without planted communities.

Theorem - The induced measure by $H_{\lambda, g(\cdot), d}$ is mutually singular with respect to that by G for any λ , $f_{in}(\cdot), f_{out}(\cdot)$ and $g(\cdot)$ where $f_{in} \neq f_{out}$ a.e.

Can learn the ***presence*** of a partition, even though in some cases cannot find it better than a random guess !

Distinguishability

Theorem - The induced measure by $H_{\lambda,g(\cdot),d}$ is mutually singular with respect to that by G for any λ ,$f_{in}(\cdot), f_{out}(\cdot)$ and $g(\cdot)$ where $f_{in} \neq f_{out}$ a.e.

Proof - ***Triangle profiles are different in the two models.***

Let L be a large constant. Define $h(x,y) = \mathbf{1}_{||x|| \leq L, ||y|| \leq L, ||x-y|| \leq L}$

At each node $\tilde{h}(X_i) = \sum_{j,k \in \mathbb{N}, j \neq k \neq i} h(X_j - X_i, X_k - X_i) \mathbf{1}_{i \sim_G j, i \sim_G k, j \sim_G k}$

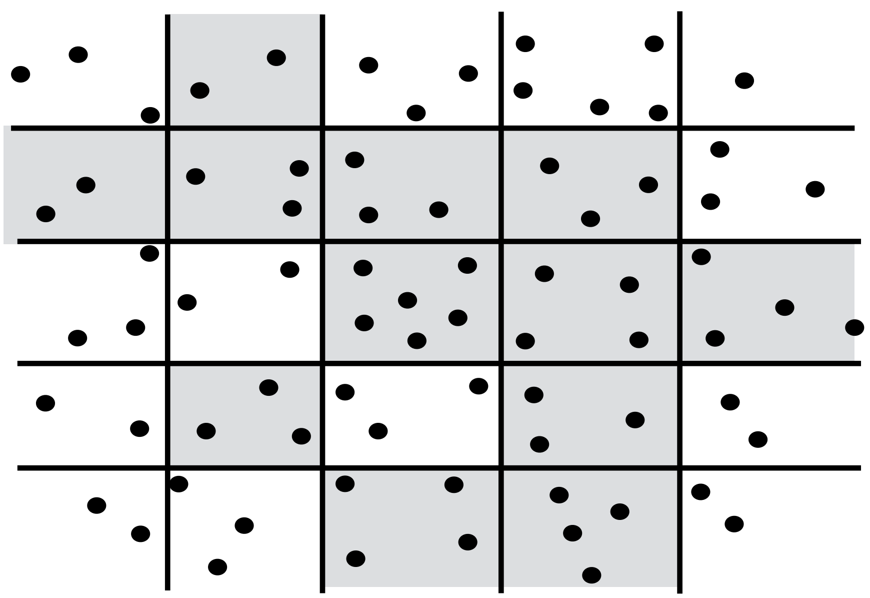
Ergodicity and moment measure expansion implies the empirical average

$\lim_{T \to \infty} \frac{\sum_{i \in \mathbb{N}} \mathbf{1}_{||X_i|| \leq T} \tilde{h}(X_i)}{\sum_{i \in \mathbb{N}} \mathbf{1}_{||X_i|| \leq T}}$ is a.s. finite and different in the two models.

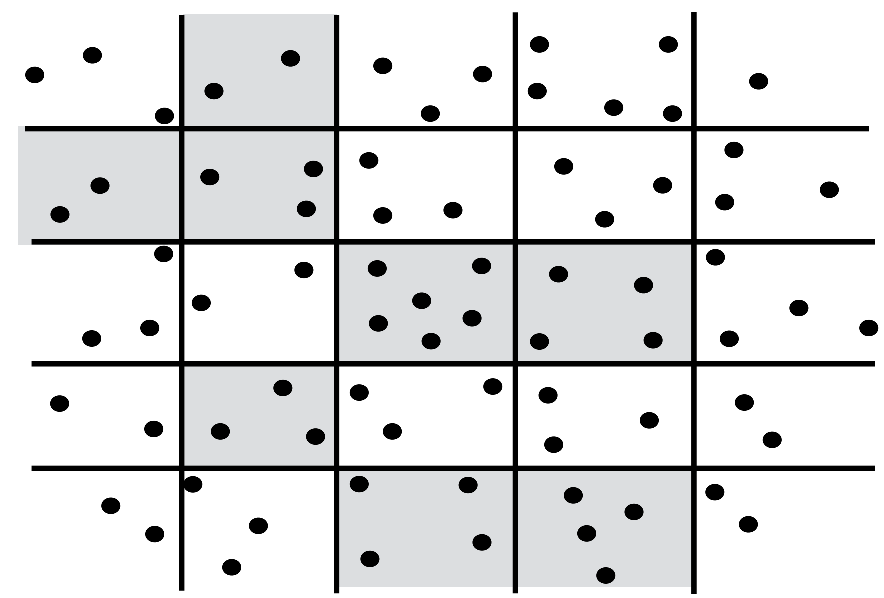
Proof gives a linear time algorithm to test between the two models.

Distinguishability Problem

Can cluster spatially locally, but no way to "synchronize" them.



Connected component to perform Community Detection.



Distinguishability only requires large number of "gray" cells.
True by SLLN for all parameters

New Phenomena - [Mossel, Neeman, Sly, '15] show that the SBM is distinguishable from the Erdos-Renyi graph iff Community Detection is solvable on the SBM.

Conclusions

- Spatial graphs are ‘locally-dense’ - basis for algorithms and analysis.
- Community Detection in the case with spatial labels has a non-trivial phase transition.
- Can always identify the presence of a partition, i.e. no phase-transition for the distinguishability problem.

Future Work

- Relax the assumption that spatial locations are known.
    - Either known noisily or are missing completely.

Thank You

https://arxiv.org/abs/1706.09942