Community Detection on an Euclidean Random Graph

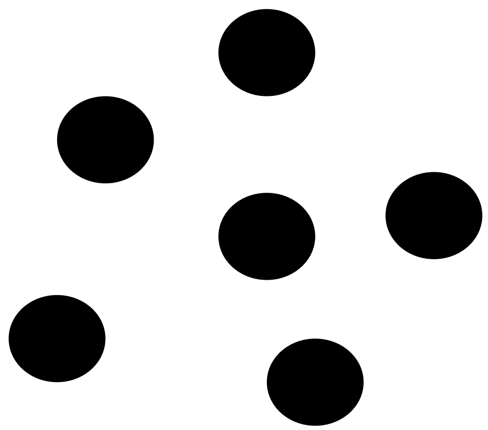
Abishek Sankararaman and François Baccelli
UT Austin

ACM-SIAM SODA 2018

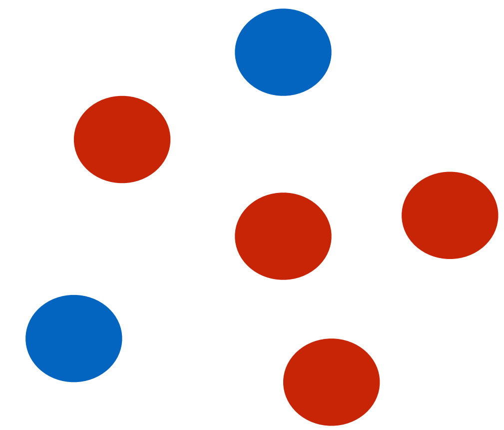
Community Detection - Abstract Definition

- Identifying ‘groups’ of objects in a population given ***indirect information*** on group memberships.

A population



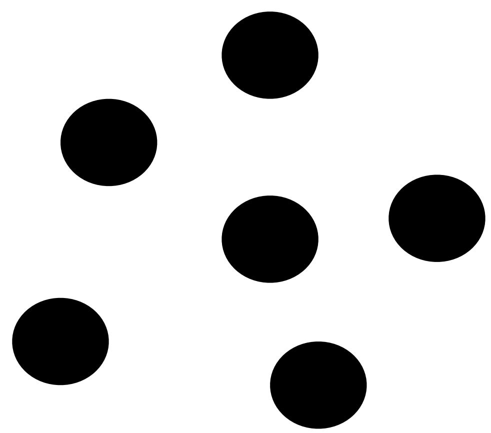
partitioned into groups



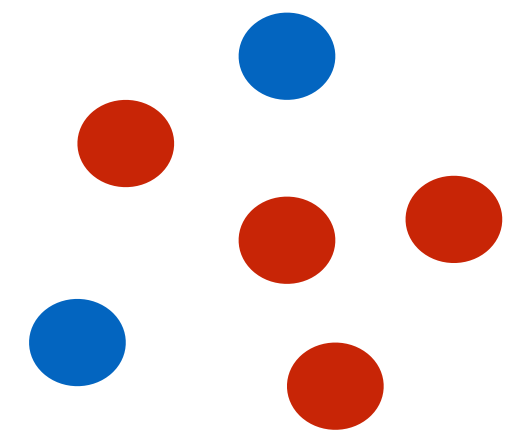
Community Detection - Examples

- Identifying ‘groups’ of objects in a population given ***indirect information*** on group memberships.

A population



partitioned into groups



1. People on an Online Social Network grouped according to whether or not they like or dislike a particular product or content.

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

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

Graph as Information

Useful sub-class of the general problem

The data is structured as follows -

Population - Represented as nodes of a graph.

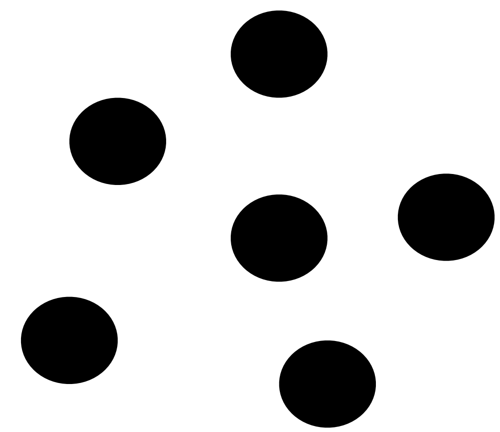
Membership Information - Encoded as labeled edges of the graph.

'Stochastic Block Model' - The simplest toy model to study this class of problems.

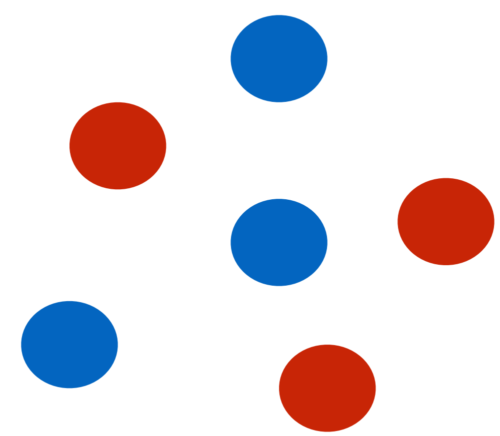
Stochastic Block Model (SBM)

The simplest case, SBM(n,a,b) $n \in \mathbb{N}$, $a, b \in [0, 1]$ is a random graph

Population of size n

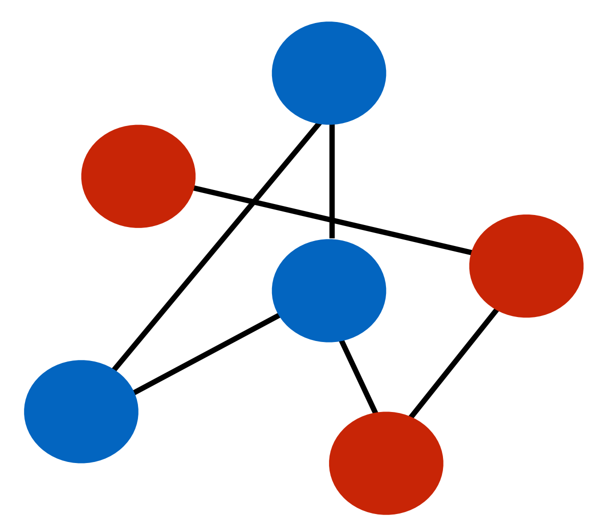


Color uniformly and independently



Conditional on the colors, draw an edge between two members with probability

- ***a*** if they have same colors.
- ***b*** if they have different colors.

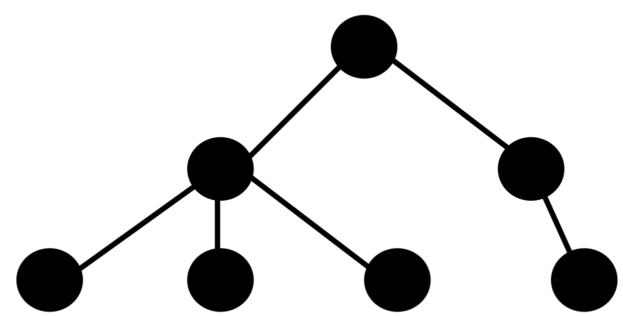


SBM for applications

The SBM is either

1. Sparse - (Finite Average Degree)

   The sparse SBM is 'Tree-Like' around any typical vertex !



[Mossel, Neeman, Sly '12]

2. Non-Sparse - Average Degree goes to infinity as $n \to \infty$.

   Not very convincing in practice.

Models for Social Network

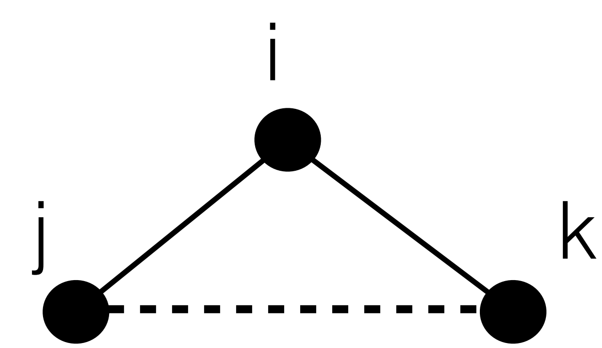
Social networks are Sparse and transitive

Sparsity - Dunbar's number :

An average human being cannot have more than 200 relationships at any point of time. This bound is a fundamental cognitive limitation, not a limitation of resources.

Transitivity

If i and j are friends, j and k are friends then i and k are likely to be friends



Latent Space Model

A *class* of models introduced by
[Hoff, Raftery, Handcock, 02], [Handcock, Raftery, Tantrum, 07].

1. The members of a social network are points in a ‘Latent Social Space’.

This is typically an unobservable abstract space, but in certain applications, it can be geographic or some feature space (age, income).

2. Conditional on the location in this latent space, edges are drawn independently at random ***depending on the Euclidean distance***.

Our Network Model - The simplest ‘planted version’ of the above.

Planted Partition Random Connection Model

Vertex Set - $\mathbb{N}$, i.e. countably infinite set.

Each node $i \in \mathbb{N}$ has two labels -

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

Model Parameters

$\lambda > 0$, $d \geq 2$, $f_{in}(\cdot), f_{out}(\cdot) : \mathbb{R}_+ \to [0,1]$ s.t $\forall r \geq 0$, $f_{in}(r) \geq f_{out}(r)$.

Statistical Assumptions

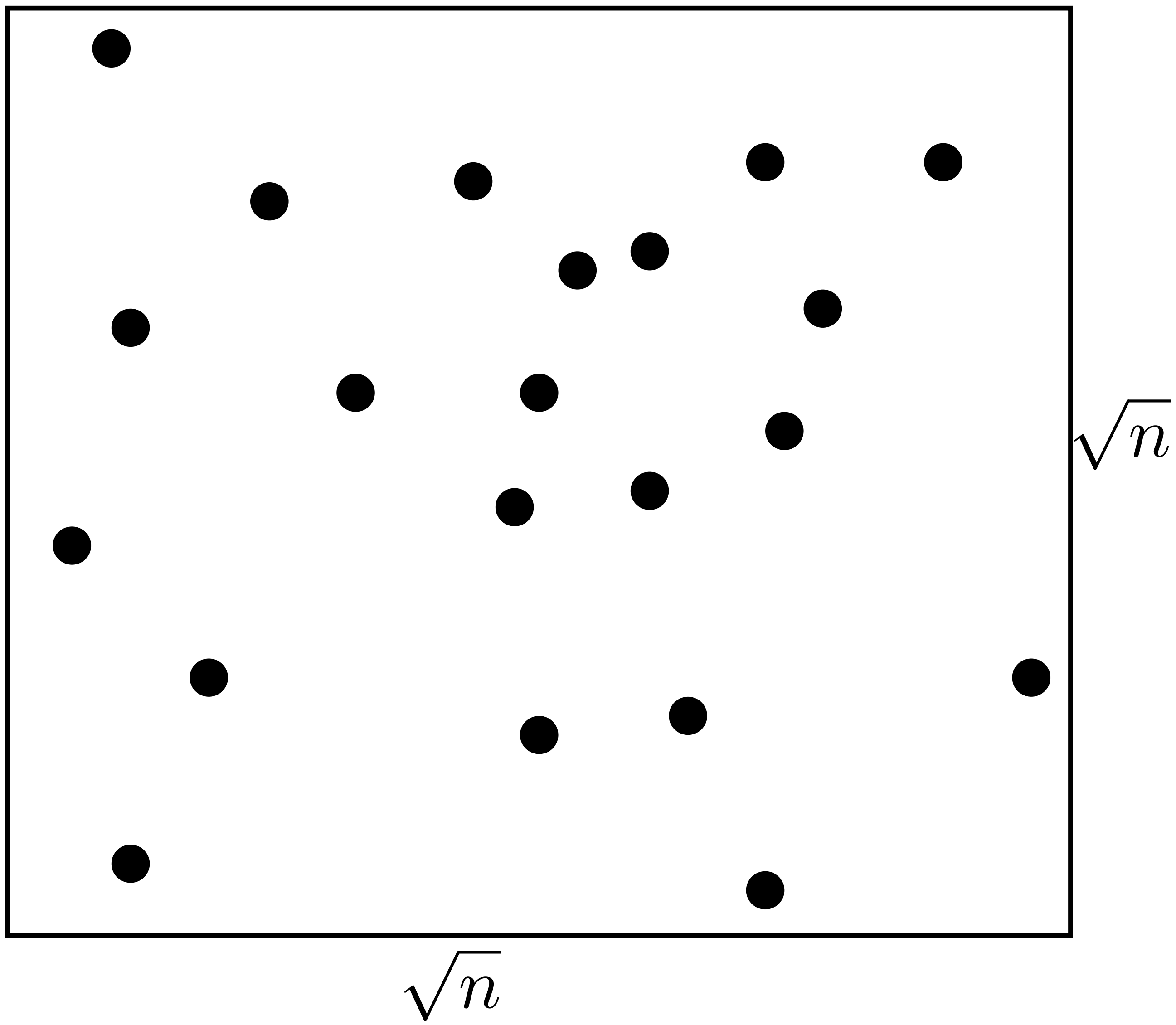
1. The locations $\{X_i\}_{i \in \mathbb{N}}$ form a **Poisson Point Process** of intensity λ on $\mathbb{R}^d$
2. $\{Z_i\}_{i \in \mathbb{N}}$ - i.i.d. sequence with each uniformly distributed on $\{-1, +1\}$
3. Conditional on node labels, edges are drawn independently at random.

   Two nodes $i \neq j \in \mathbb{N}$ are connected with probability

   $f_{in}(||X_i - X_j||)$ if $Z_i = Z_j$ or $f_{out}(||X_i - X_j||)$ if $Z_i \neq Z_j$

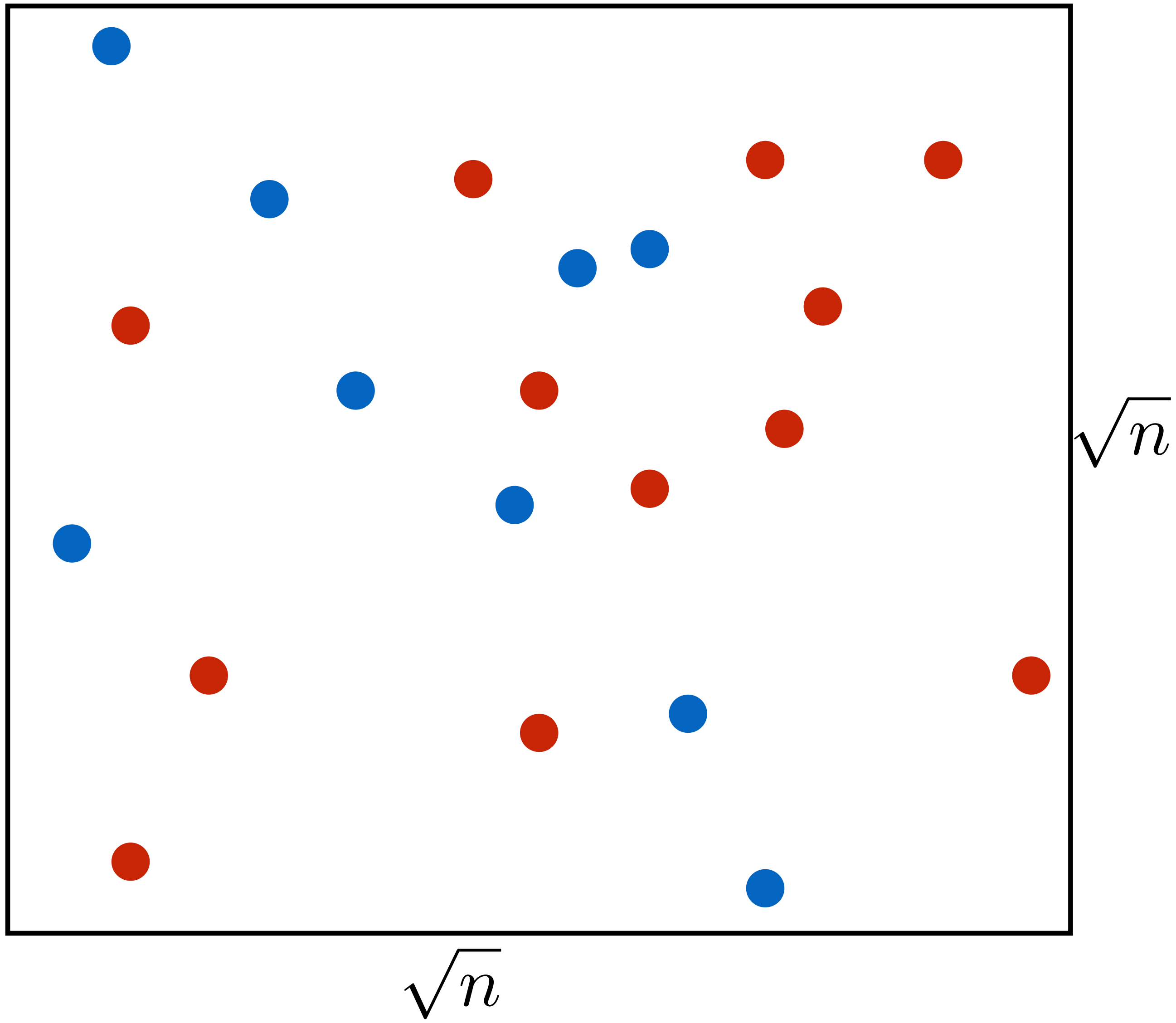
   (On average, more edges within communities than across)

Planted Partition Random Connection Model



Place $\mathrm{Poi}(\lambda n)$ points independently and uniformly in $\left[-\frac{n^{1/d}}{2}\frac{n^{1/d}}{2}\right]$

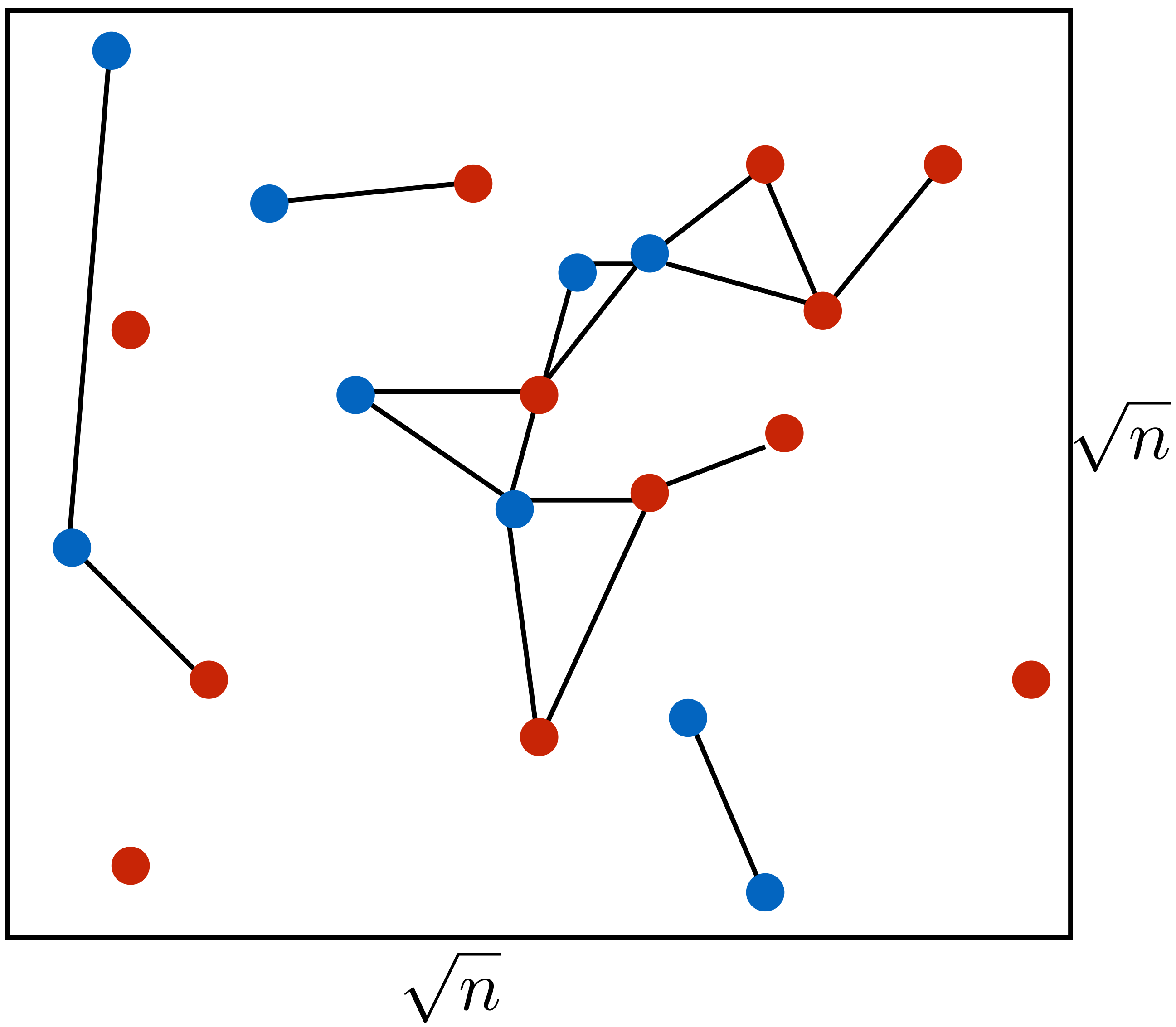
Planted Partition Random Connection Model



Place $\mathrm{Poi}(\lambda n)$ points independently and uniformly in $\left[-\frac{n^{1/d}}{2}, \frac{n^{1/d}}{2}\right]$

Color uniformly and independently

Planted Partition Random Connection Model



Place $\mathrm{Poi}(\lambda n)$ points independently and uniformly in $\left[-\frac{n^{1/d}}{2}\frac{n^{1/d}}{2}\right]$

Color uniformly and independently

Conditional on the location and colors, draw edges independently.

Two points at distance r are connected with probability

- $f_{in}(r)$ - if they have **same** colors.
- $f_{out}(r)$ - if they have **opposite** colors.

Ignore Edge Effects

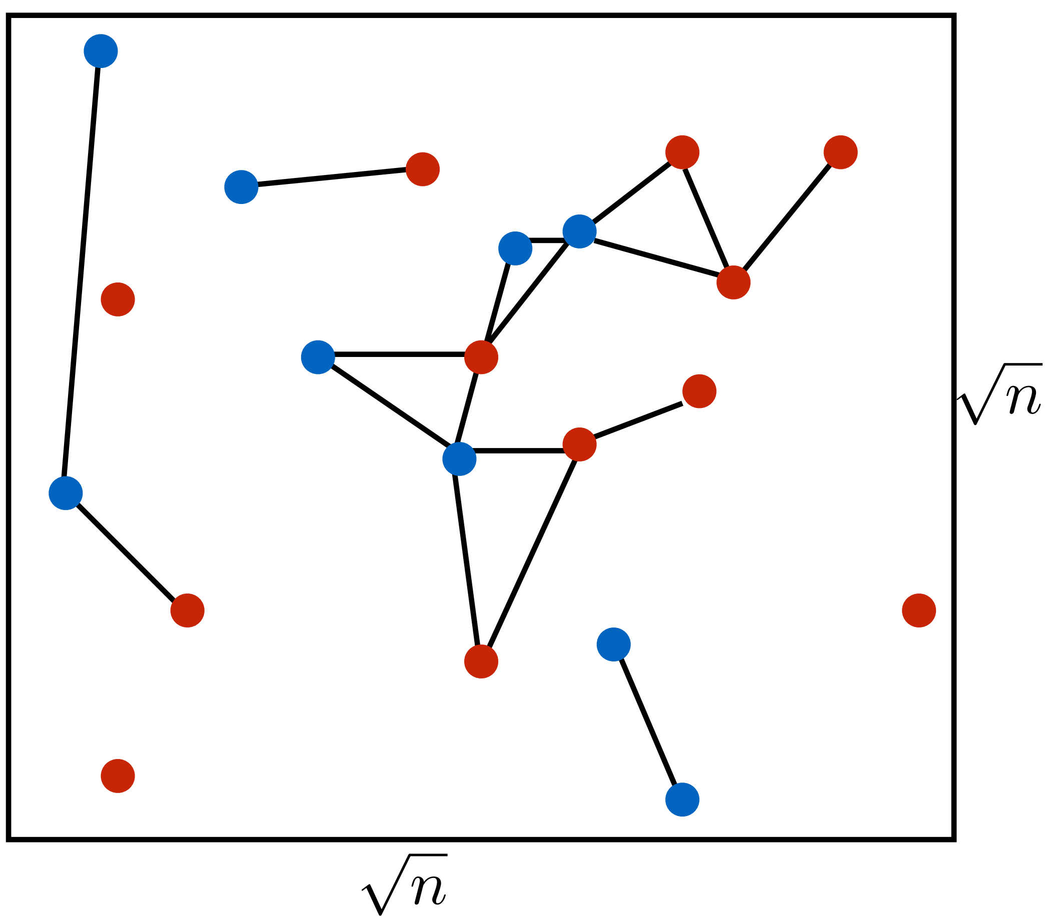
Denote by graph by G_n and its "limit" as $n \to \infty$ by G

Planted Partition Random Connection Model

Nodes are indexed in increasing l_∞ distance of its location labels, i.e. $||X_i||_\infty < ||X_{i+1}||_\infty \ \forall i \in \mathbb{N}$.

N_n denotes the number of nodes in G_n

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

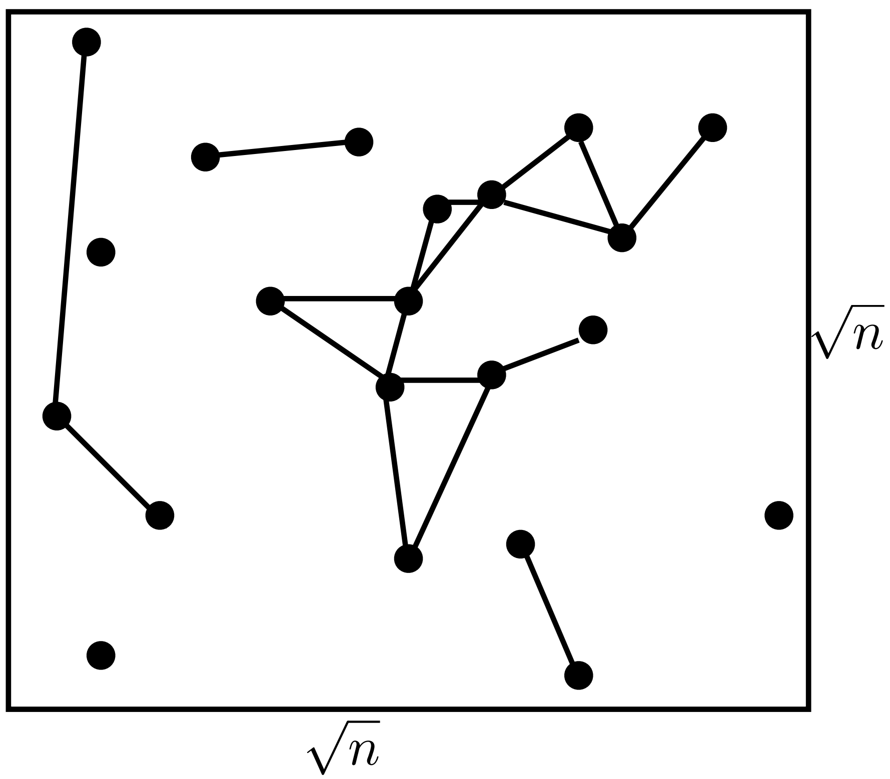


Average number of neighbors in the

same community $\int_{x\in\mathbb{R}^d} f_{in}(||x||)dx - o(1)$

opposite community $\int_{x\in\mathbb{R}^d} f_{out}(||x||)dx - o(1)$

Community Detection Problem



Given G_n and $\{X_i\}_{i \in [0,N_n]}$, can one produce an estimate $\{\tau_i\}_{i \in [0,N_n]}$ of the community labels ?

Will assume $\lambda, d, f_{in}(\cdot), f_{out}(\cdot)$ to be known

Community Detection ***solvable*** if $\exists \gamma > 0$ and $\{\tau_i\}_{i \in [0,N_n]}$ which are measurable functions of $(G_n, \{X_i\}_{i \in [0,N_n]})$ such that

$$\lim_{n \to \infty} \mathbb{P}\left[\left| \sum_{i=1}^{N_n} \frac{\tau_i Z_i}{N_n} \right| > \gamma\right] = 1$$

(Asymptotically beating a random guess)

Monotonicity

$\forall f_{in}(\cdot), f_{out}(\cdot), d \geq 2, \exists \lambda_c \in [0, \infty]$ such that -

$\lambda < \lambda_c \implies$ Community Detection is not solvable.

$\lambda > \lambda_c \implies$ Community Detection is solvable

Proof - Independently deleting nodes from the (planted partition) random connection model yields another (planted partition) random connection model.

However not satisfying -

λ_c could be either 0 or ∞

No insight into designing efficient algorithms

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.

<u>Proposition -</u> In certain special cases, we find $\lambda_1 = \lambda_2$, i.e. characterize the exact phase-transition point.

$f_{in}(r) = \mathbf{1}_{r \leq R_1}$, $f_{out}(r) = \mathbf{1}_{r \leq R_2}$ with $0 \leq R_2 < R_1$

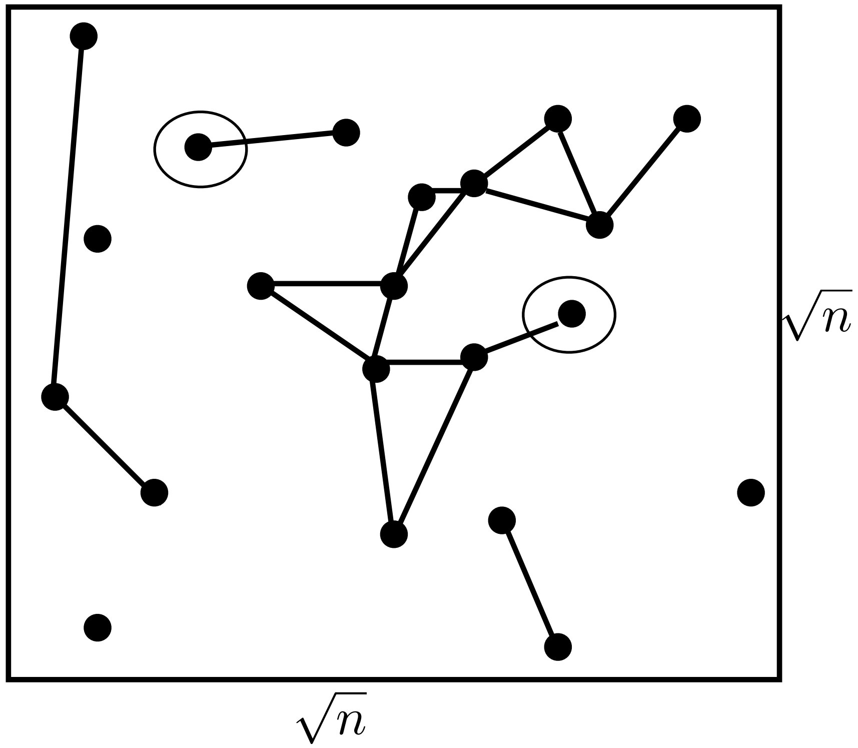
In general, characterizing the exact-phase transition is hard -

The location of the phase-transition for percolation in random connection models is itself unknown.

Impossibility

Consider the following *easier* problem -

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



If Community Detection is solvable, i.e. if

$$\left|\frac{\sum_{i=1}^{N_n} Z_i\tau_i}{N_n}\right| \geq \gamma$$

, then the above can be solved with success probability at-least $\frac{1+\gamma}{2}$

(Cluster the whole graph and then answer)

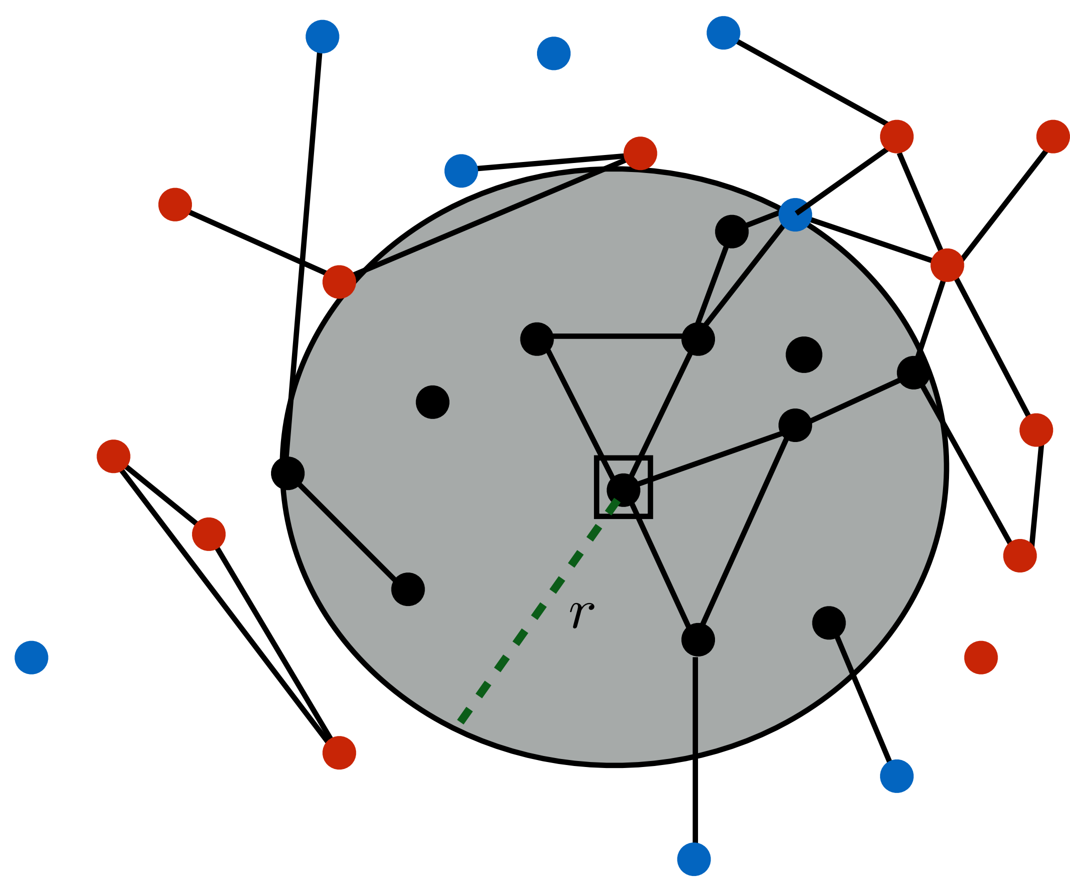
Will prove that the question above is not solvable for sufficiently small λ

Impossibility

W.h.p, the distance between the two chosen nodes is '***large***'
— in particular larger than any constant r

An easier problem

Can you estimate better than chance, the community label of a random node in G_n given the infinite graph G, $\{X_i\}_{i \in \mathbb{N}}$ all locations and all community labels of nodes that are at a distance r or more from this chosen node.

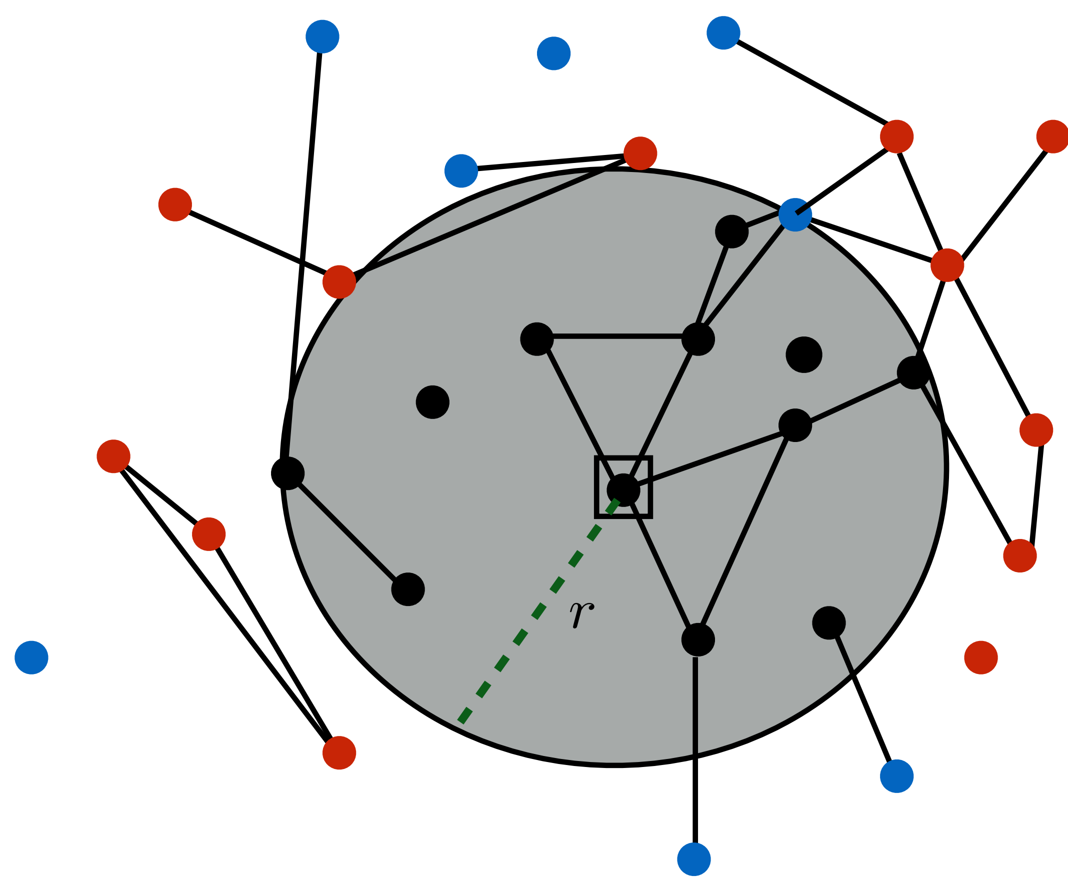


Information Flow from Infinity Problem

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 $\liminf_{r \to \infty} \mathbb{P}^0[\tau_0^{'} = Z_0] \geq \frac{1}{2} + \gamma^{'}$?

$\mathbb{P}^0$ Palm Probability measure - Place a fictitious node at origin with an independent community label Z_0 and independent edges to G

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



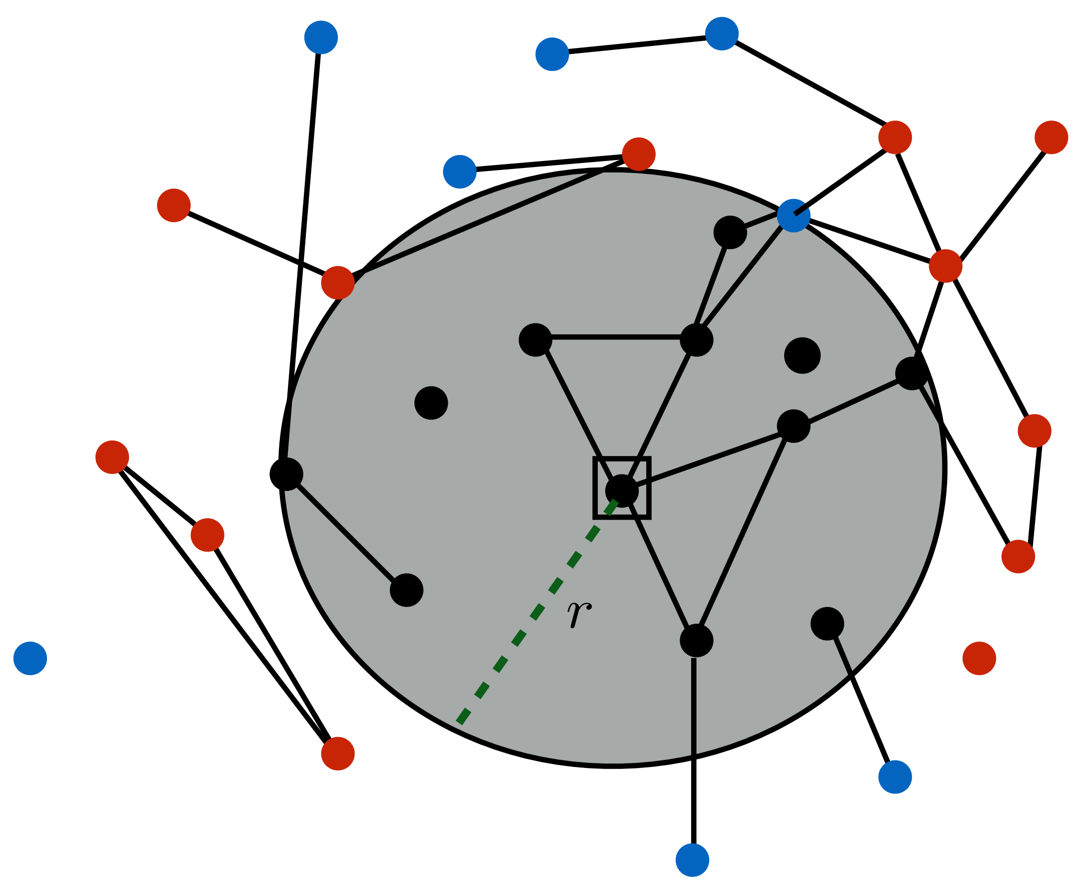
Information Flow from Infinity Problem

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 connection model on a PPP of intensity λ and connection function $f_{in}(\cdot) - f_{out}(\cdot)$ ***does not percolate***, then the answer to the above question is NO.

Corollaries

1. If $d = 1$, then community detection is not solvable for any $\lambda, f_{in}(\cdot), f_{out}(\cdot)$.
2. If $\lambda \int_{x \in \mathbb{R}^d} (f_{in}(||x||) - f_{out}(||x||)) dx \leq 1$, then community detection is not solvable.



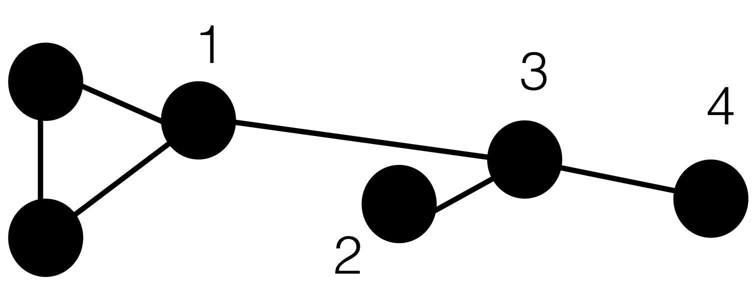
Information Flow from Infinity Problem

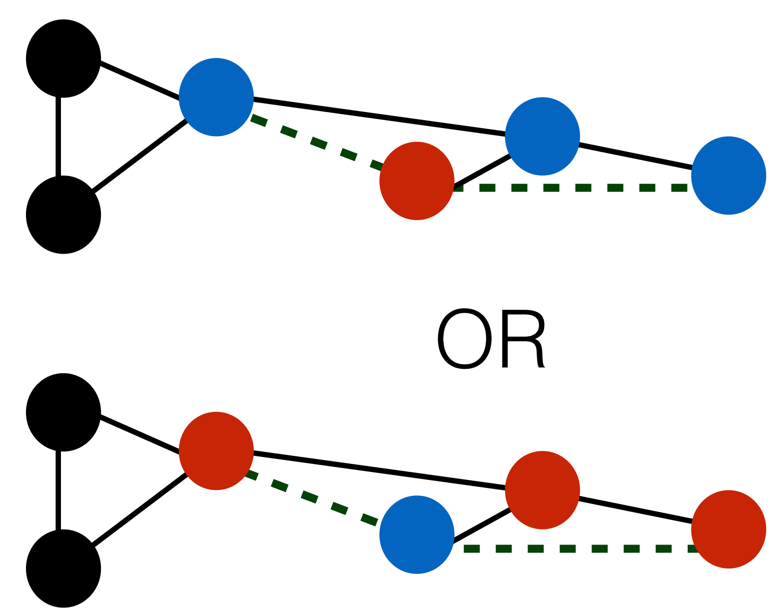
Consider an illustrative example.

Let $f_{in}(r) = \mathbf{1}_{r \leq R_{in}}$ and $f_{out}(r) = \mathbf{1}_{r \leq R_{in}}$ where $R_{in} > R_{out}$

Therefore in G , any two nodes will have

1) An edge if they are within a distance of R_{out}
2) No edge if they are more than a distance of R_{in}
3) An edge only if they belong to same community and are at a distance of $(R_{out}, R_{in}]$





$$||X_1 - X_2||, ||X_2 - X_4||, ||X_1 - X_3|| \in (R_{out}, R_{in}]$$

Information Flow from Infinity Problem

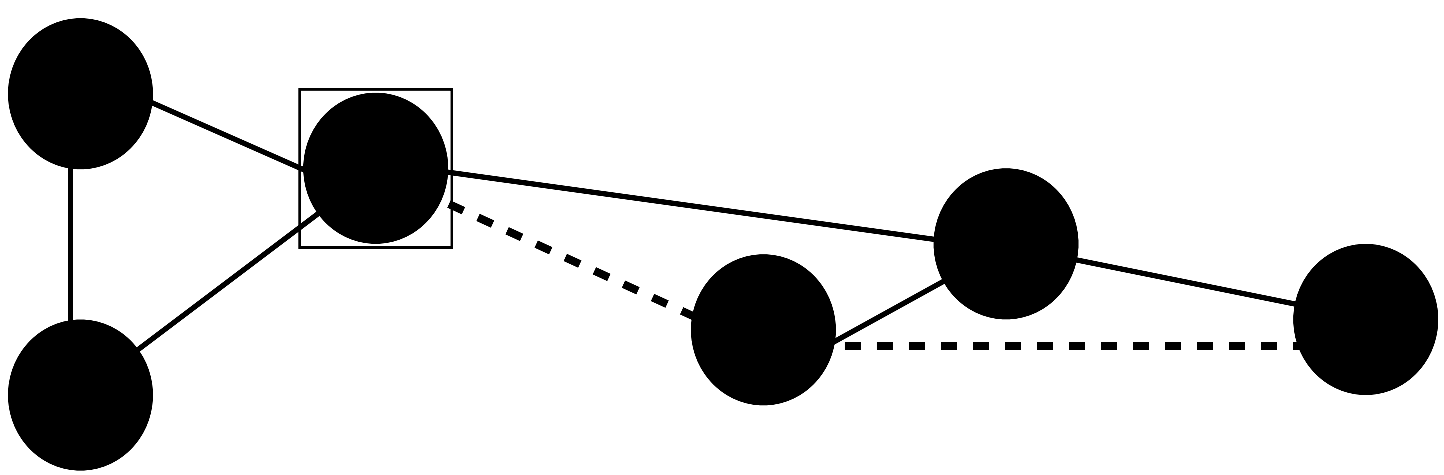
Consider.

$$f_{in}(r) = \mathbf{1}_{r \leq R_{in}} \;,\; f_{out}(r) = \mathbf{1}_{r \leq R_{out}} \qquad R_{in} > R_{out}$$

A natural strategy

If $\exists\ 0 := X_0, X_1, \cdots X_k \in \phi$, $||X_i - X_{i+1}|| \in (R_{out}, R_{in}]\ \forall\ i \in [0, k-1]$

If Z_k is known , then we can '*propagate it to infer* Z_0.



Information Flow from Infinity Problem

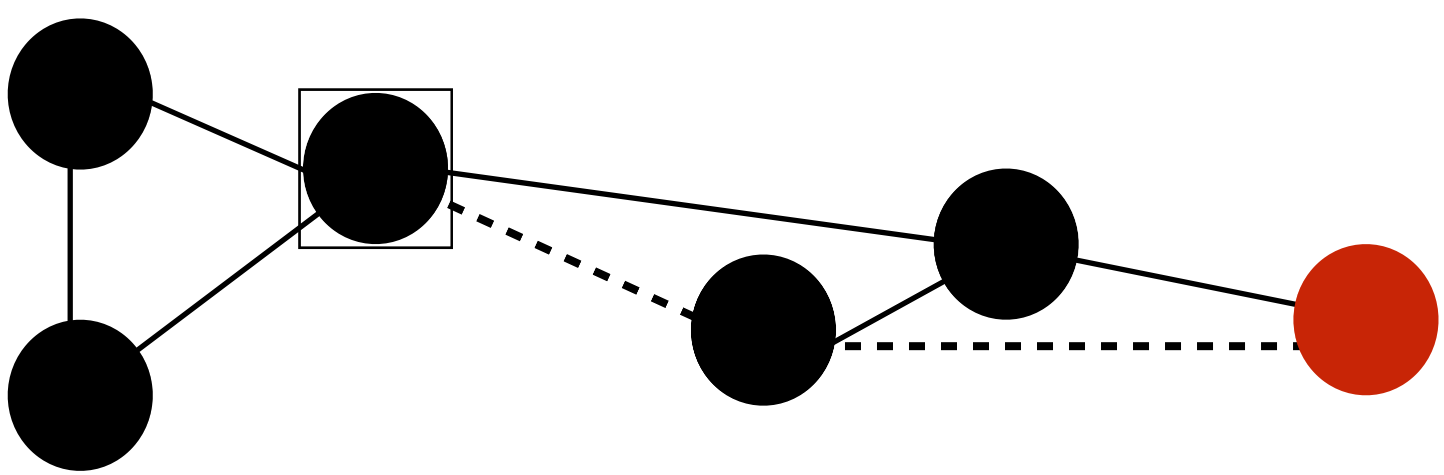
Consider.

$$f_{in}(r) = \mathbf{1}_{r \leq R_{in}} \ , \ f_{out}(r) = \mathbf{1}_{r \leq R_{out}} \qquad R_{in} > R_{out}$$

<u>A natural strategy</u>

If $\exists \ 0 := X_0, X_1, \cdots X_k \in \phi$, $||X_i - X_{i+1}|| \in (R_{out}, R_{in}] \ \forall \ i \in [0, k-1]$

If Z_k is known , then we can '*propagate it to infer* Z_0*.*



Information Flow from Infinity Problem

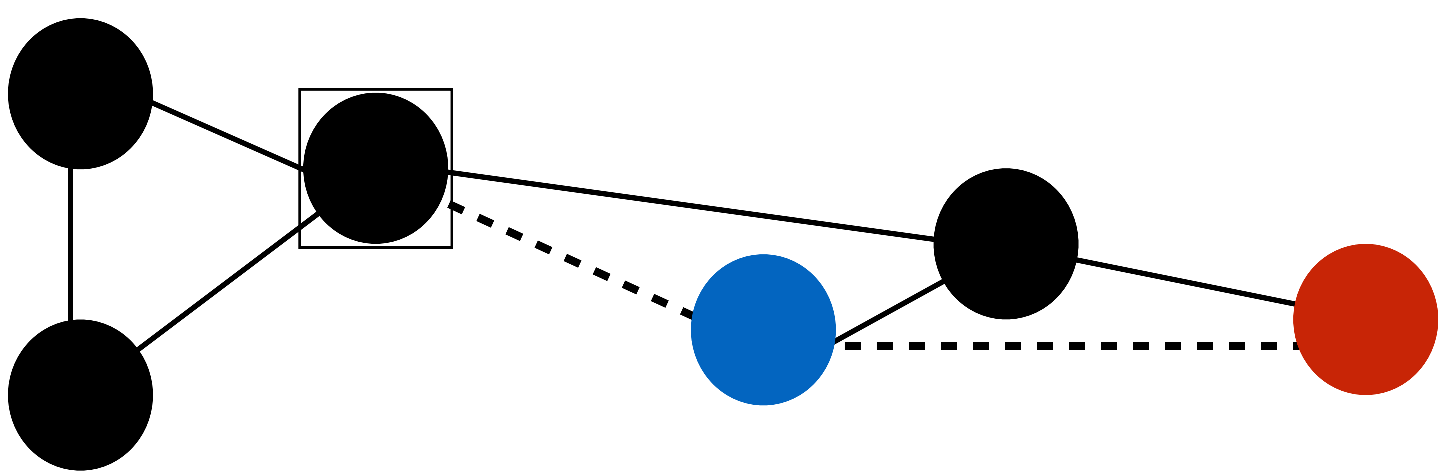
Consider.

$$f_{in}(r) = \mathbf{1}_{r \leq R_{in}} \;, \; f_{out}(r) = \mathbf{1}_{r \leq R_{out}} \qquad R_{in} > R_{out}$$

<u>A natural strategy</u>

If $\exists\ 0 := X_0, X_1, \cdots X_k \in \phi$, $||X_i - X_{i+1}|| \in (R_{out}, R_{in}]\ \forall\ i \in [0, k-1]$

If Z_k is known , then we can '*propagate it to infer* Z_0.



Information Flow from Infinity Problem

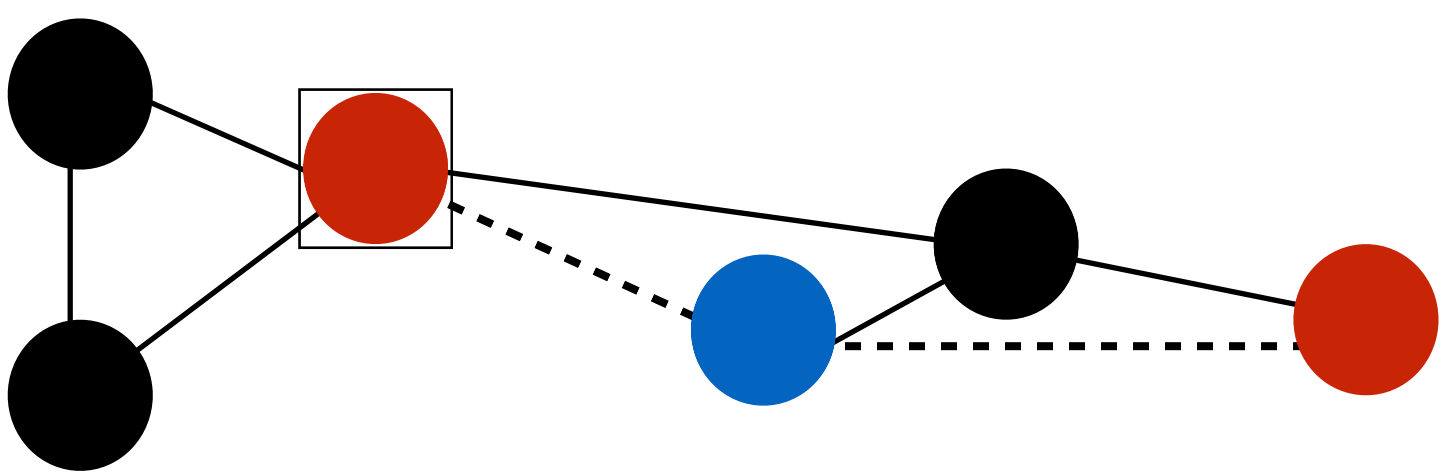
Consider.

$$f_{in}(r) = \mathbf{1}_{r \le R_{in}} \ , \ f_{out}(r) = \mathbf{1}_{r \le R_{out}} \qquad R_{in} > R_{out}$$

<u>A natural strategy</u>

If $\exists \ 0 := X_0, X_1, \cdots X_k \in \phi$, $||X_i - X_{i+1}|| \in (R_{out}, R_{in}] \ \forall \ i \in [0, k-1]$

If Z_k is known , then we can '*propagate it to infer* Z_0.



Information Flow from Infinity Problem

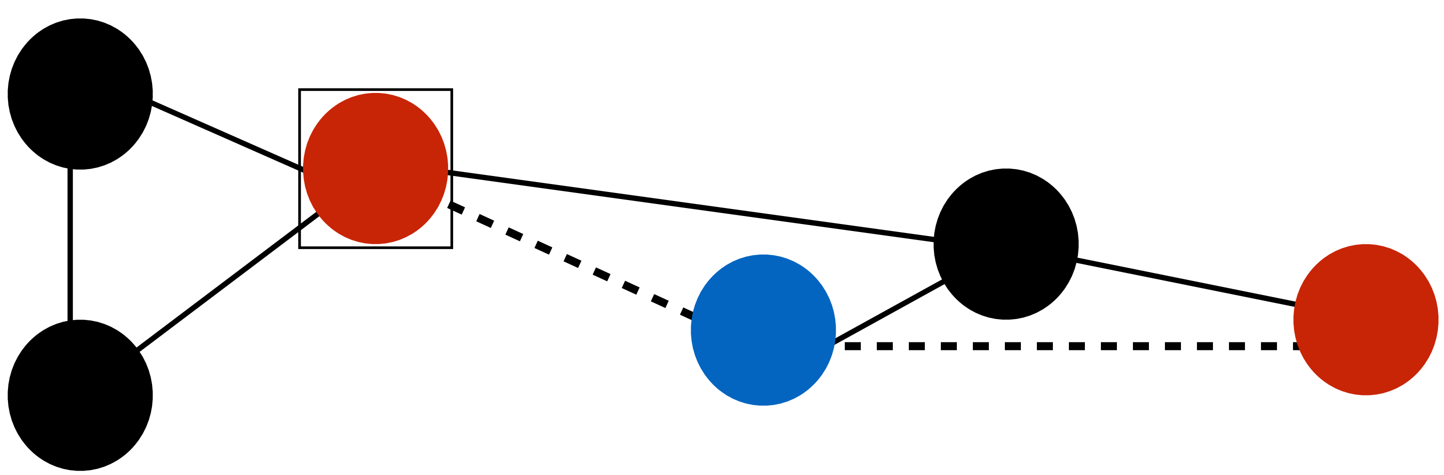
Consider.

$$f_{in}(r) = \mathbf{1}_{r \leq R_{in}} \text{ , } f_{out}(r) = \mathbf{1}_{r \leq R_{out}} \qquad R_{in} > R_{out}$$

A natural strategy

If $\exists \ 0 := X_0, X_1, \cdots X_k \in \phi$, $||X_i - X_{i+1}|| \in (R_{out}, R_{in}] \ \forall \ i \in [0, k-1]$

If Z_k is known , then we can '*propagate it to infer* Z_0*.*



Our result - If no such path exists, then cannot determine the label at 0.

Information Flow from Infinity Problem

Enriched probability space with marks on pairs of nodes.

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

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

3) An edge between nodes $i<j\in\mathbb{N}$ if and only if

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

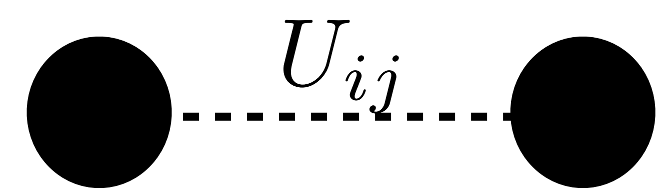
Thus the graph G is a deterministic function of node labels $\{(X_i, Z_i)\}_{i\in\mathbb{N}}$

and the edge labels $\{U_{ij}\}_{i<j\in\mathbb{N}}$.

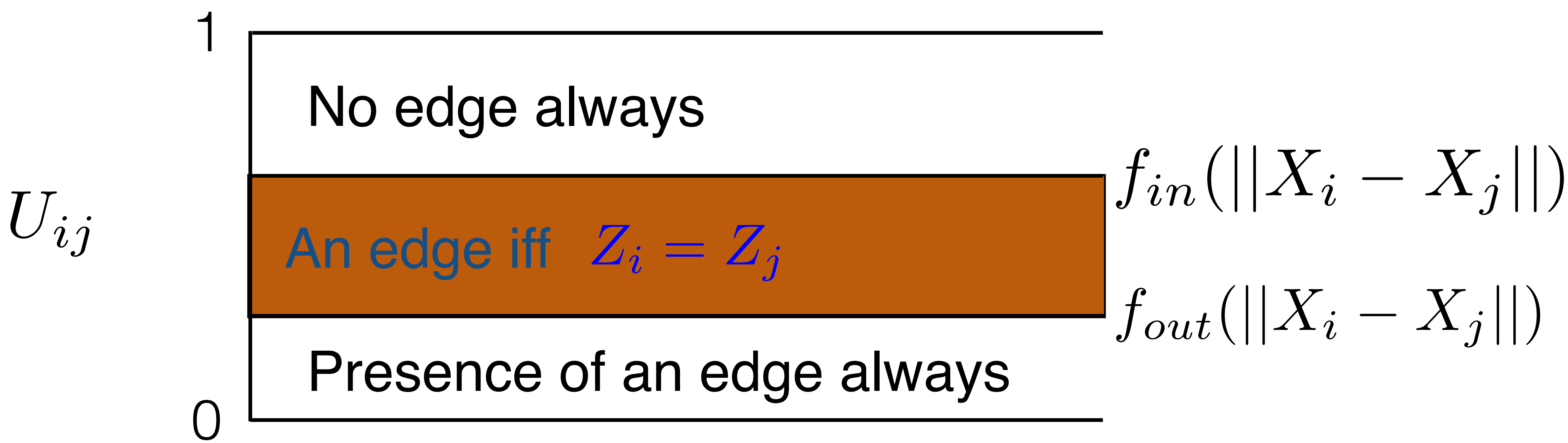
Information Flow from Infinity Problem

$\{U_{ij}\}_{i<j\in\mathbb{N}}$, -i.i.d. $U[0,1]$ sequence, one for each pair of nodes.
An edge between nodes $i < j \in \mathbb{N}$ if and only if

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

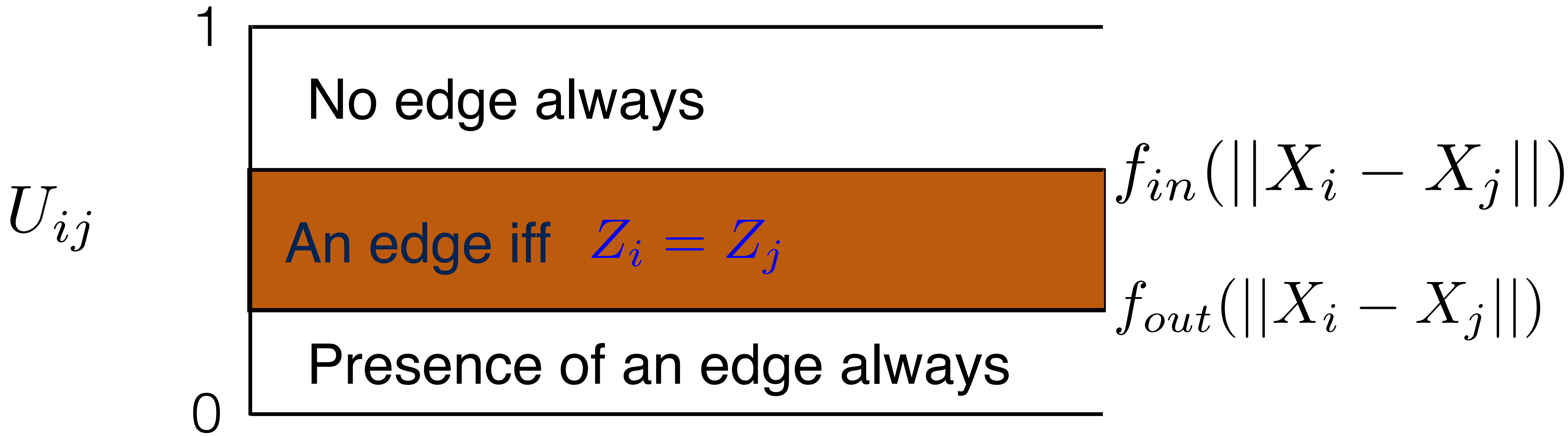


Only certain edges are *Informative*



The presence or absence of an edge is only informative when
$U_{ij} \in (f_{out}(||X_i - X_j||), f_{in}(||X_i - X_j||)]$

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

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

$V_I(0) \subset \mathbb{N}$ - 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.

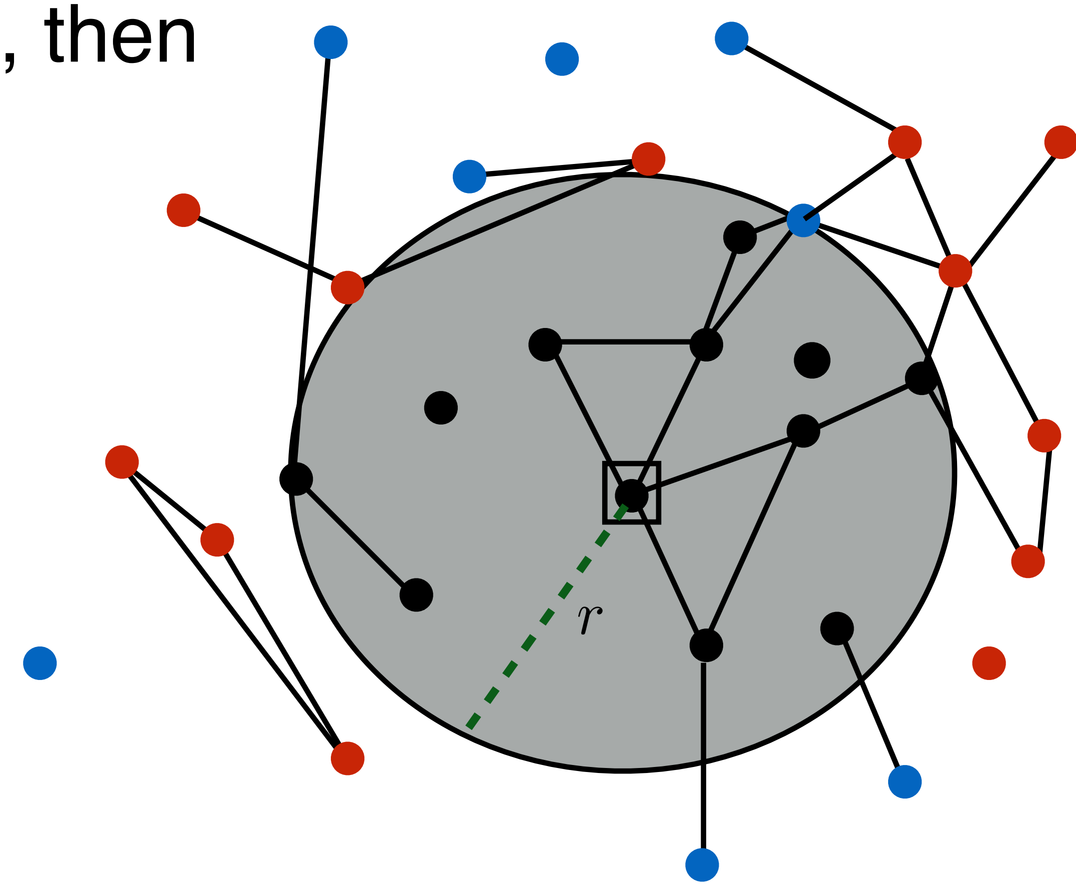
Information Flow from Infinity Problem

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 $\liminf_{r \to \infty} \mathbb{P}^0[\tau_0^{'} = Z_0] \geq \frac{1}{2} + \gamma^{'}$?

From previous lemma, 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 the answer above is no.



Algorithm Idea

Our spatial graph - *locally dense* but globally *sparse*

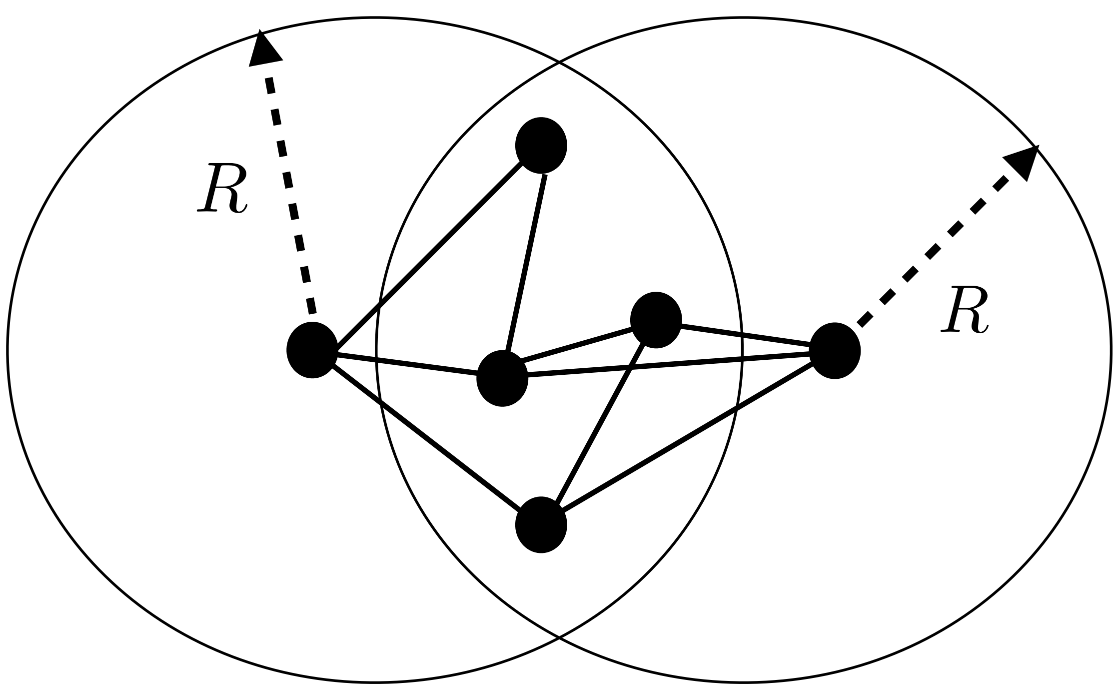
Consider the example $f_{in}(r) = a\mathbf{1}_{r \leq R}$, $f_{out}(r) = b\mathbf{1}_{r \leq R}$

Locally Dense - '*Nearby*' nodes connect with *constant probability* independent of n

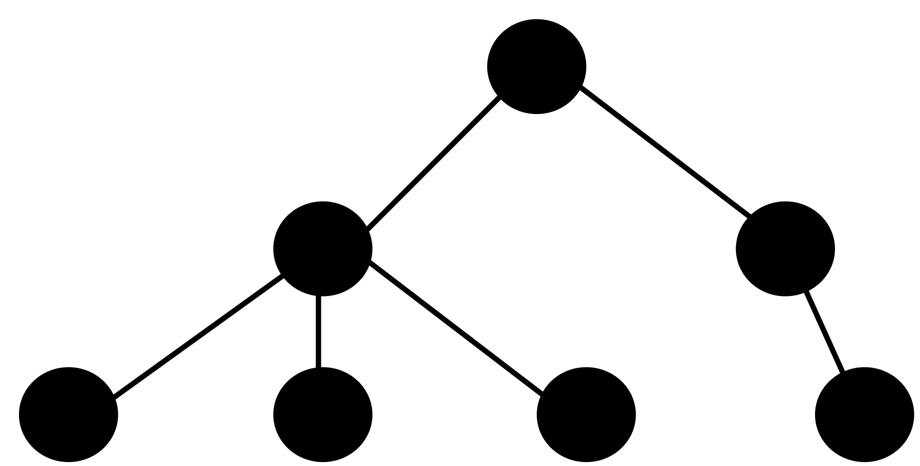
Globally Sparse - Order n edges in total.

The sparse SBM is locally tree like.

Every node connects with each other with probability tending to 0.



Spatial Graph



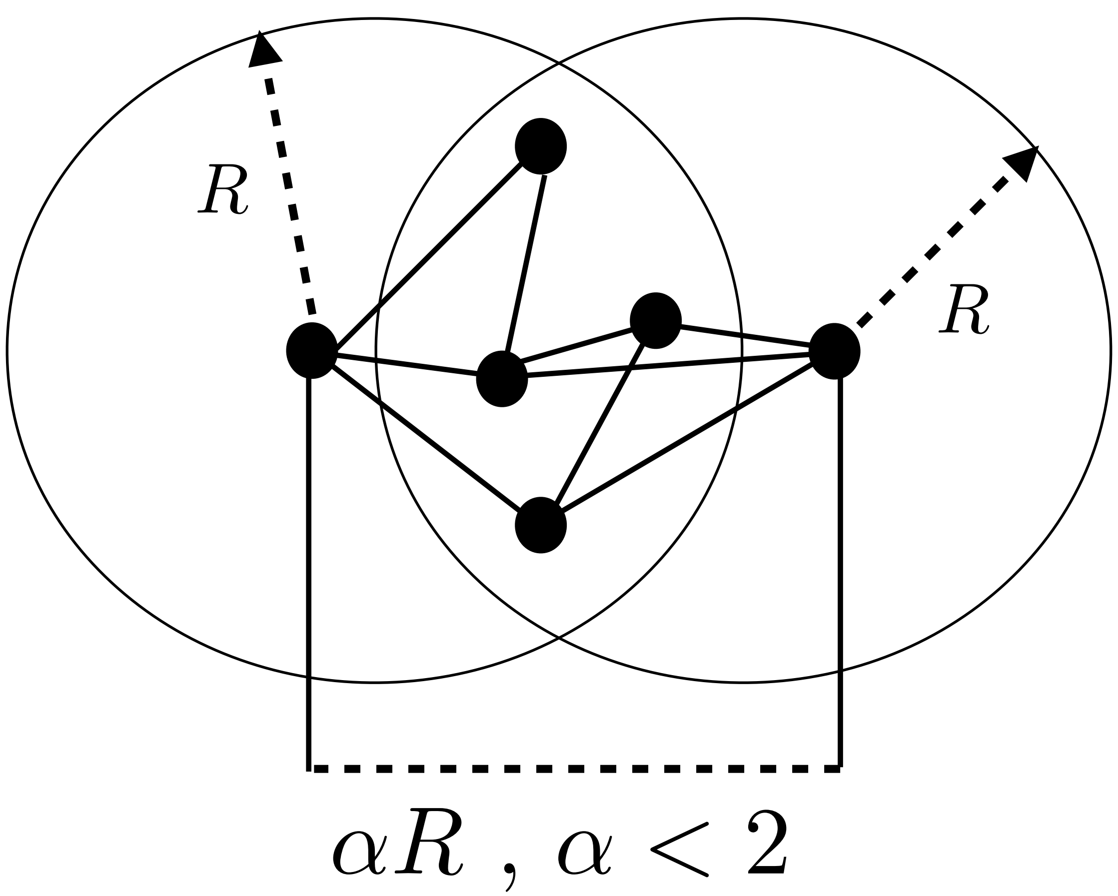
SBM

Algorithm Idea

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

Locally Dense - Geometry around ‘nearby’ nodes have lot of information.

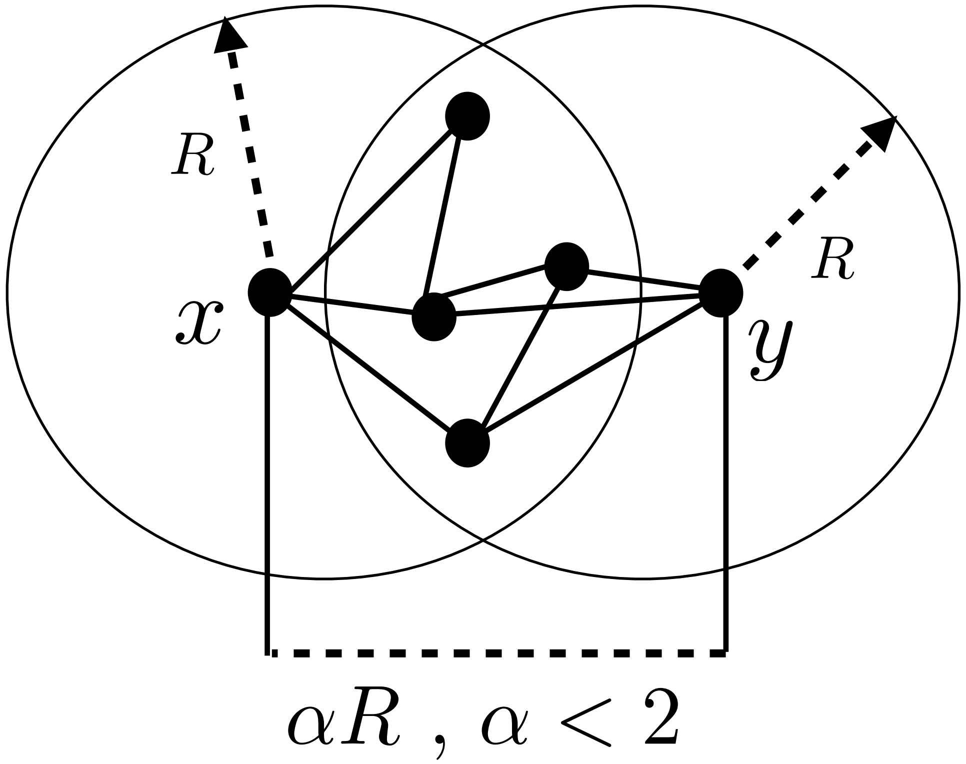
The number of common neighbors of two nodes is Poisson with mean



- $\lambda c(\alpha) R^d \left(\frac{a^2 + b^2}{2} \right)$ if they belong to same community.
- $\lambda c(\alpha) R^d ab$ if they belong to different communities.

Both are of order λ

Algorithm Idea



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

<u>Pairwise-Classify(x,y)</u>

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

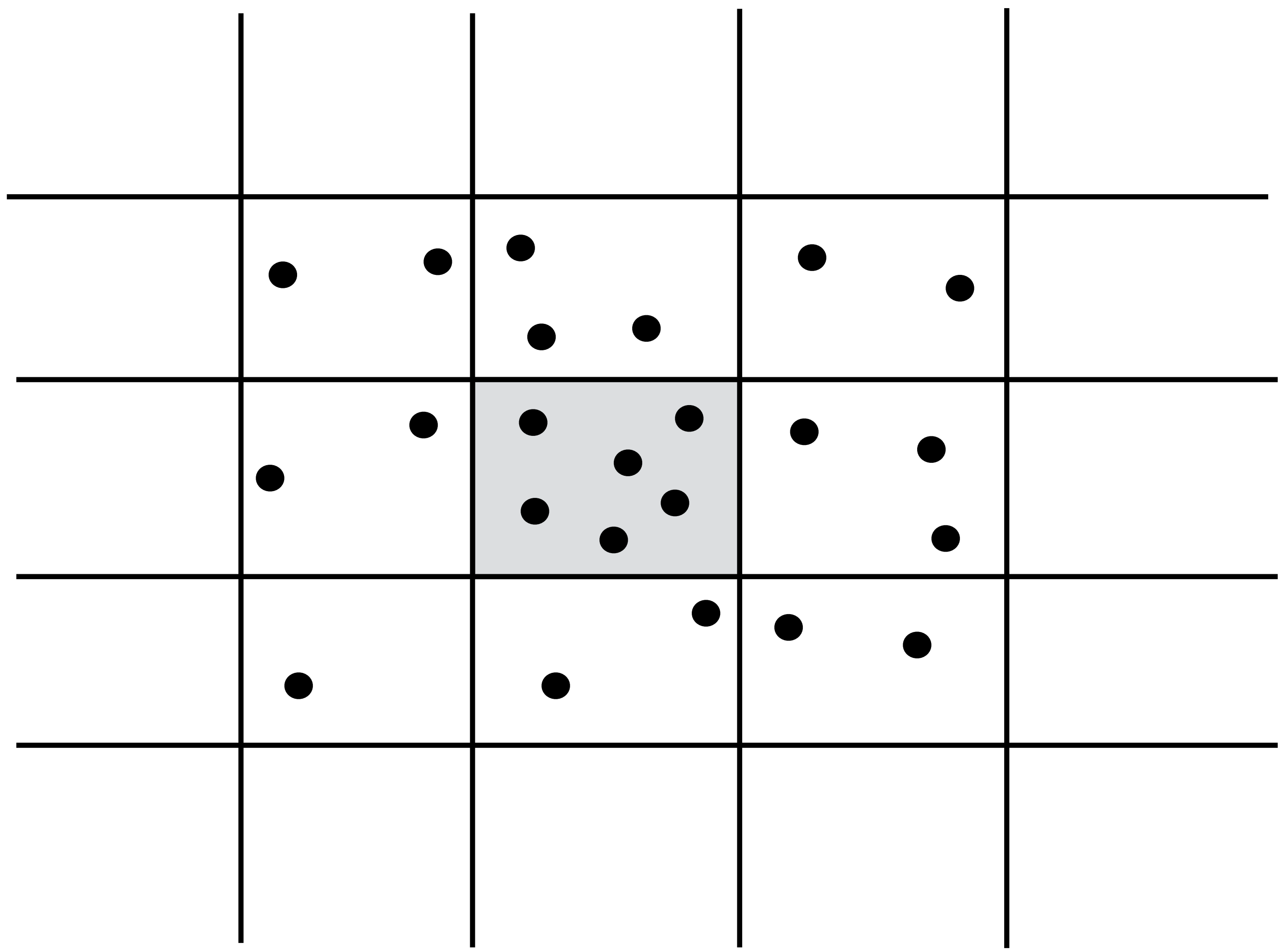
Simple Chernoff bound -

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

$$\alpha < 2 \implies c'(\alpha) > 0$$

Algorithm Idea

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

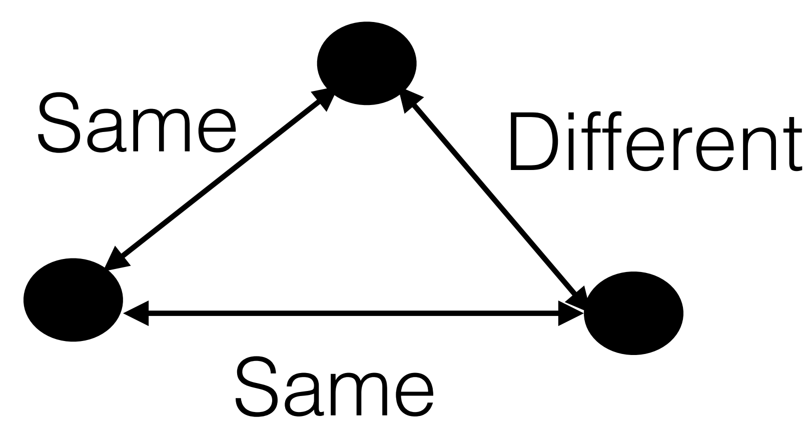


Classify cells to be Good or Bad.

Cell ***Good*** if

1. At-least $1 - \epsilon$ the mean number of points
2. No ***inconsistencies*** in pairwise checks *with all neighboring cells*

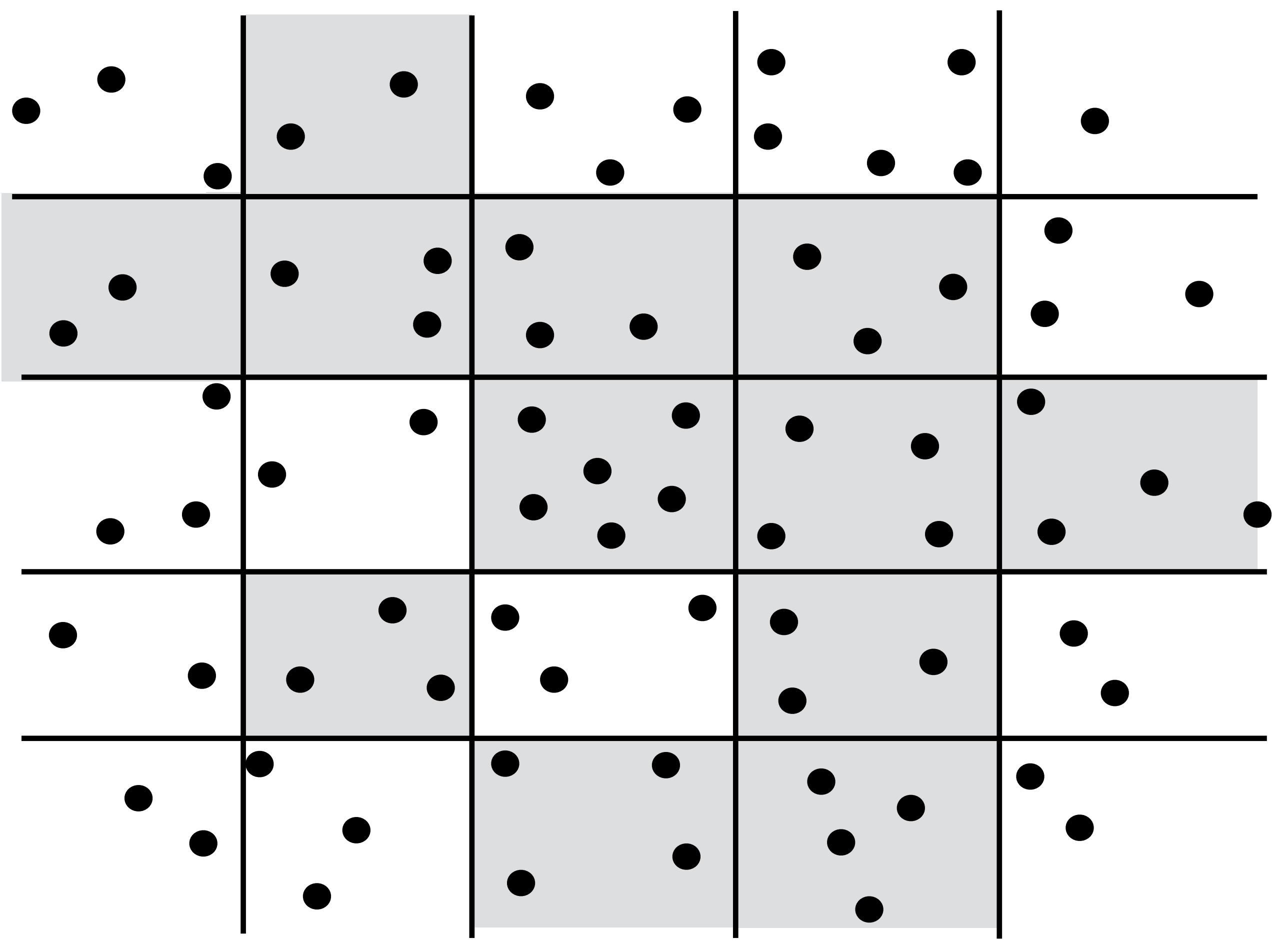
Example of Inconsistent output of Pairwise-Classify



Algorithm Idea

__Main Routine__

- Create a partition of each good component.
  Unique partition of the nodes in good component compatible with
  Pairwise-Classify
- Output +1 estimate to all nodes in bad cells



For any $\gamma \in [0,1)$, $\exists \lambda_0(\gamma) < \infty$, such that $\forall \lambda \geq \lambda_0(\gamma)$ the algorithm will succeed, i.e.

$$\lim_{n \to \infty} \mathbb{P}\left[\left| \sum_{i=1}^{N_n} \frac{\tau_i Z_i}{N_n} \right| > \gamma \right] = 1$$

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

Distinguishability - Are there communities ?

$H_{\lambda,g(\cdot),d}$ Random Connection Model on a PPP of intensity λ and connection function $g(\cdot)$.

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.

Are there communities at all

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 says we can answer this **always**. No phase-transition.*

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

Distinguishability Problem

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 - Then 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.

An algorithm to test between the two models.

Conclusions

- A new model of random graph with planted communities.
- Spatial graphs are ‘locally-dense’ - basis for algorithms and analysis.

- Community Detection in the case with spatial labels has a non-trivial phase transition.
- However can always identify the presence of a partition, i.e. no phase-transition for the distinguishability problem.

<u>Future Work</u>

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

- Sharp Phase-Transitions in some regimes of the problem.
  - Help characterize and design ‘optimal’ algorithms.

Full paper on Arxiv - https://arxiv.org/abs/1706.09942

Thank You