Community Detection on Euclidean Random Graphs

Abishek Sankararaman[*], François Baccelli [†]

October 26, 2017

Abstract

Motivated by applications in online social networks, we introduce and study the problem of Community Detection on a new class of sparse *spatial* random graphs embedded in the Euclidean space. Our random graph is the planted-partition version of the classical random connection model studied in Stochastic Geometry. Roughly speaking, each node of our graph has an uniform i.i.d. $\{-1, +1\}$ valued community label and a $\mathbb{R}^d$ valued location label given by the support of a homogeneous Poisson point process of intensity λ. Conditional on the labels, edges are drawn independently at random depending both on the Euclidean distance between the nodes and the community labels on the nodes. We study the Community Detection problem on this random graph which consists in estimating the partition of nodes into communities, based on an observation of the random graph along with the spatial location labels on nodes.

We show that for $d = 1$, Community Detection is impossible for any parameters. For $d \geq 2$, we establish a phase-transition based on the intensity λ of the point process. In particular, we show that if the intensity λ is small, then no algorithm for community detection can beat a random guess for the partitions. We prove this by introducing and analyzing a new problem which we call 'Information Flow from Infinity'. On the positive side, we give a novel algorithm that performs Community Detection as long as the intensity λ is larger than a sufficiently high constant. Along the way, we establish a *distinguishability* result which says that one can always efficiently infer the existence of a partition given the graph and the spatial locations even when one cannot identify the partition better than at random. This is a surprising new phenomenon not observed thus far in any non-spatial Erdős-Rényi based planted-partition models.

[*]Department of ECE, The University of Texas at Austin. Email - abishek@utexas.edu.
[†]Department of Mathematics and ECE, The University of Texas at Austin. Email - baccelli@math.utexas.edu.

1 Introduction

Community Detection, also known as the graph clustering problem, is the task of grouping together nodes of a graph into representative clusters. This problem has several incarnations that have proven to be useful in various applications ([1]) such as social sciences ([2],[3]), image segmentation [4], recommendation systems ([5],[6]), web-page sorting [7], and biology ([8], [9]) to name a few. In the present paper, we introduce a new sparse *spatial* random graph and study the graph clustering problem on it. The random graph we consider is denoted by G_n, which has a random number of nodes denoted by N_n that is Poisson distributed with mean λn. In our formulation, n is a scaling parameter going to infinity and $\lambda > 0$ is a *fixed* constant that denotes the *intensity* parameter. Nodes are equipped with two i.i.d. labels, a uniform $\{-1, +1\}$ valued *community label* and a uniform $B_n := \left[-\frac{n^{1/d}}{2}, \frac{n^{1/d}}{2}\right]^d$, $d \in \mathbb{N}$ valued *location label*. Therefore, the average number of nodes having location labels in any subset of B_n of unit volume is λ, which explains why we call it the intensity parameter. Conditional on the node labels, two nodes at locations $x, y \in B_n$ and community labels $Z_x, Z_y \in \{-1, +1\}$ are connected by an edge in G_n independently of other edges with probability $f_{in}(||x-y||)$ if $Z_x = Z_y$ or with probability $f_{out}(||x-y||)$ if $Z_x \neq Z_y$, where $||\cdot||$ denotes the Euclidean norm. We assume $1 \geq f_{in}(r) \geq f_{out}(r) \geq 0$ for all $r \geq 0$ and therefore the graph is expected to have more edges within communities than across. Furthermore, we consider the *sparse* graph regime where $\int_{x\in\mathbb{R}^d} f_{out}(||x||)dx \leq \int_{x\in\mathbb{R}^d} f_{in}(||x||)dx < \infty$. Hence in this sparse regime, the average degree of any node in G_n is bounded above by $(\lambda/2)\int_{x\in\mathbb{R}^d}(f_{in}(||x||) + f_{out}(||x||))dx < \infty$ uniformly in n. Moreover, in this sparse regime, the boundary effects due to the finite window B_n will not matter for large n. We give a rigorous discussion of our model in the sequel in Section 2.

The Community Detection problem in our context refers to when and how one can estimate the partition of the nodes of G_n into communities better than at random, from an observation of the graph **and** the spatial location labels that generated the graph. We say Community Detection is solvable (made precise in Section 2.2) if there exists an algorithm such that the fraction of misclassified nodes is strictly smaller than a half as n goes to infinity, i.e., we asymptotically beat a random guess of the partition. Although this requirement on estimation is very weak, we see through our results that this is indeed the best one can hope for in the sparse graph setting considered here. For simplicity, we also assume that the algorithm has knowledge of the connection functions $f_{in}(\cdot), f_{out}(\cdot)$. The estimation of the connection functions from data in our spatial setup is an interesting research question in itself which is beyond the scope of this paper.

Motivations for a Spatial Model - The most widely studied model for Community Detection is the Stochastic Block Model (SBM), which is a multi-type Erdős-Rényi graph. In the simplest case, the two community symmetric SBM corresponds to a random graph with n nodes, with each node equipped with an i.i.d. uniform community label drawn from $\{-1, +1\}$. Conditionally on the labels, pairs of nodes are connected by an edge independently of other pairs with two different probabilities depending on whether the end points are in the same or different communities. Structurally, the sparse SBM is known to be locally tree-like ([10],[11]) while real social networks are observed to be *transitive* and sparse. Sparsity in social networks can be understood through 'Dunbar's number' [12], which concludes that an average human being can have only about 500 'relationships' (online and offline) at any point of time. Moreover, this is a fundamental cognitive limitation of the person and not that of access or resources, thereby justifying models where aver-

age node degree is independent of the population size. Social networks are transitive in the sense that any two agents that share a mutual common neighbor tend to have an edge among them as well, i.e., the graph has many triangles. These observations make the sparse SBM non realistic model and have led to the development of Latent Space Models ([13],[14]) in the social sciences literature. These are sparse *spatial* graphs in which the agents of the social network are assumed to be embedded in an abstract *social space* that is modeled as a finite dimensional Euclidean space and conditional on the embedding, edges of the graph are drawn independently at random as a non-increasing function of the distance between two nodes. Thanks to the properties of Euclidean geometry, these models are transitive and sparse, and have a better fit to data than any SBM ([13]).

Thus, one can view our model as the simplest planted-partition version of the Latent Space model, where the nodes are distributed uniformly in a large compact set B_n and conditional on the locations, edges are drawn depending on Euclidean distance through connection functions $f_{in}.(\cdot)$ and $f_{out}(\cdot)$. Although, our assumptions are not particularly tailored towards any real data-sets, our setting is the most challenging regime for the estimation problem as the location labels alone without the graph reveal no community membership information. However, a drawback in our formulation is that we assume the location labels on nodes are known exactly. In practice, it is likely that the locations labels are unknown (as in the original Latent Space models where the social space is unobservable) or are at-best estimated separately. Nonetheless, our formulation with known location labels forms a crucial first step towards more general models where the location labels are noisy or are missing. The problem with known spatial location labels is itself quite challenging as outlined in the sequel below and hence we decide to focus on this setting alone in the present paper. Another drawback of our formulation is that we assume the estimator has knowledge of the model parameters $f_{in}(\cdot)$ and $f_{out}(\cdot)$. In our spatial setup, the estimation of connection functions from data is an interesting research question in itself which is however beyond the scope of this paper.

Central Technical Challenges - The core technical challenge in studying our spatial graph model lies in the fact that it is not 'locally tree-like'. The spatial graph is locally dense (i.e. there are lots of triangles) which arises as a result of the constraints imposed by Euclidean Geometry, while it is globally sparse (i.e. the average degree is bounded above by a constant). The sparse SBM on the other hand, is locally 'tree-like' and has very few short cycles [10]. This comes from the fact that the connection probability in a sparse SBM scales as c/n for some $c > 0$. In contrast, the connection function in our model does not scale with n. From an algorithmic point of view however, most commonly used techniques (message passing, broadcast process on trees, convex relaxations, spectral methods etc) are not straight forward to apply in our setting since their analysis fundamentally relies on the locally tree-like structure of the graph (see [11] and references therein). This fundamental difficulty renders the problem quite challenging, even in the presence of known spatial location labels. We overcome this difficulty by proposing a novel clustering algorithm and lower bound technique. The key idea for our algorithm is to exploit the fact that our graph is locally quite dense, which can allows us to classify very accurately 'nearby' pair of nodes. We then use ideas from percolation theory to then piece together in a non-trivial fashion the different nearby estimators to produce a global clustering. To prove the lower bound, we develop a new coupling argument by connecting Community Detection with a problem we call 'Information Flow from Infinity', which is a new problem and is of independent interest.

The other speciality in our setting is the presence of location labels which are known exactly to the estimator. This provide some form of 'side-information' which any estimator must exploit. As an illustrative example to see this, consider the connection functions $f_{in}(\cdot)$ and $f_{out}(\cdot)$ to be of bounded support. In this case, the absence of an edge between two nearby nodes makes it likely that these two nodes belong to opposite communities but the lack of an edge between far-away nodes farther than the support of either connection functions does not give any community membership information. Thus the lack of an edge in this example has different interpretations depending on the location labels which needs to be exploited in a principled manner by the estimator. Nonetheless, the location labels alone without the graph provide no information on community membership as the community and location labels are independent of each other.

1.1 Overview of our Results and Techniques

The central result in this paper is the existence of a phase-transition for Community Detection on spatial graphs. We show that for every $f_{in}(\cdot), f_{out}(\cdot)$ and $d \geq 2$, there is a non-trivial[1] critical constant $\lambda_c \in (0, \infty)$, such that as λ increases beyond this critical value, the Community Detection problem shifts from being unsolvable to solvable. Our proof for this is in two parts.

1. We establish in Theorem 4 that no Community Detection algorithm (polynomial or exponential time) can beat a random guess of the partition if the intensity λ is small (with an explicit estimate provided for this constant). To show this, we first argue that Community Detection is easier than determining whether any two uniformly randomly chosen nodes of G_n belong to the same or different communities with success probability larger than a half. Now since the nodes are uniformly distributed in the set B_n which has volume n, with high probability, the chosen nodes will be 'far-away'; in particular the distance between the two nodes will exceed any constant r. We then make this pair-wise problem easier by revealing the true label of all nodes at distance r or more from one of the two chosen nodes and asking how well can one estimate the community label on this chosen node. We call this problem 'Information Flow from Infinity', since we want to understand whether one can estimate the community label of a uniformly randomly chosen node of G_n, given the true community labels of all nodes 'far-away' (at distance r or more). To show impossibility, we construct 'information percolation' clusters and establish that when these clusters are 'small', then the information from the revealed community labels does not 'flow', thereby one cannot solve the pairwise classification problem.

2. On the positive side, we give a novel Community Detection algorithm called `GBG` in Section 4, which solves Community Detection if the intensity λ is a sufficiently high constant (with an explicit estimate of the constant). Our algorithm has running time complexity of order n^2 and storage complexity of order n. The key idea is to observe that for any two nodes that are 'near-by', we can classify them correctly with exponentially (in λ) small probability of error. However, this is still not sufficient to produce a global clustering since there will be certain pairs incorrectly classified that need to be identified and corrected. We establish this by embedding 'consistency checks' into our algorithm to correct some of the mis-classified pairs by partitioning the space B_n into 'good' and 'bad' regions. We then couple the partitioning of space with another

[1]The $d = 1$ case is in a sense trivial as seen in Corollary 5

percolation process to prove that our algorithm will mis-classify a fraction strictly smaller than half the nodes if λ is sufficiently high, with an explicit estimate of the constant. Furthermore, in certain special instances of connection functions $f_{in}(\cdot)$ and $f_{out}(\cdot)$, our lower bound and upper bound match to give a sharp phase-transition and we can characterize the critical λ for these cases.

Along the way, in Theorem 10 we also consider a problem which is known as the *distinguishability* problem. which asks how well one can solve a hypothesis testing problem between our graph and an appropriate null model (a plain random connection model with connection function $(f_{in}(\cdot) + f_{out}(\cdot))/2)$ without communities but having the same average degree and distribution for spatial locations. We show that for *all* parameters, we can solve the distinguishability problem with success probability $1 - o_n(1)$, even if we cannot learn the partition better than a random guess. The key technical contribution lies in identifying suitable graph 'triangle profiles' and show that they are different in the planted partition model and the null model. We are able to infer the existence of the partition but do not learn anything about it. This is because we can 'see' the partition 'locally' in space, but there is no way to consistently piece together the small partitions in different regions of space into one coherent partition of the graph. This phenomenon is new and starkly different from what is observed in the classical Erdős-Rényi based symmetric SBM with two communitities where the moment one can identify the presence of a partition, one can also recover the partition better than a random guess ([10]). Moreover, such phenomena where one can infer the existence of a partition but not identify it better than random are conjectured not to occur in the SBM even with many communities [15],[16].

1.2 Related Work

Community Detection on sparse graphs has mostly been studied on the SBM as the random graph model. The study of SBM has a long history in the different literatures of statistical physics (see [15] and references therein), mathematics (for ex. [17],[10]) and computer science (for ex. [18], [19],[20]). We refer the reader to the comprehensive survey [11] for a complete mathematical and algorithmic treatment of the Community Detection problem on the SBM. The survey of [21] gives a complete treatment of the SBM from a statistical physics view point. There has been renewed interest in the sparse regime of the SBM following the paper of [15], which made number of striking conjectures on phase-transitions. Subsequently, some of them have been established with the most notable and relevant achievements to ours being that of [10],[22], [23] and [24]. These papers prove that both Community Detection and the distinguishability problem for the two community sparse SBM undergo a phase-transition at the same point which they characterize explicitly. These results for the SBM motivates the investigation of the phase-transitions for Community Detection and distinguishability in our model. However, the tools needed for our model are very different from those used to study the SBM. Most prior work on sparse SBM relies on the locally tree-like property of the random graph in some way or the other ([11] has a nice account of the methods) and are hence not directly applicable in our setting. Thus, the key ideas for all of our results come from different problems, mostly those studied in the theory of percolation [25] and stochastic geometry [26]. Some of the key ideas for our algorithm are motivated by similar, albeit different ideas that appeared in the Interacting Particle Systems literature (for ex. [26] [27],[28],[29]). The papers there developed renormalization and percolation based ideas to study different particle systems arising in statistical mechanics. However all those papers are not algorithmic and ours is the first paper to consider such spatial percolation techniques and apply it in a non-trivial fashion to

propose estimation algorithms. Our lower bound comes from identifying an *easier* problem than Community Detection called Information Flow from Infinity, which is a new problem. The key idea to show an impossibility for the Information Flow from Infinity problem comes comes from certain 'random-cluster method' and coupling arguments. Such methods are quite popular and have been developed extensively in other contexts for example to study mixing time of Ising Models ([30],[31],[32]). Coupling ideas similar in spirit to ours has also appeared in other estimation contexts such as the reconstruction on trees problem [33] where a lower bound was established using this method. However, in all of the previous literature, this coupling method was applied on fixed graphs (grids, trees, bounded degree graphs), while we need to apply to the continuum (the spatial locations are not discrete) and unbounded degree graphs. Moreover, we need to first identify that the Information Flow from Infinity is the correct mathematical object to study the lower bound, and then develop the coupling method in the continuum for this problem. This makes our application of the random-cluster idea new and non-trivial. We recently learned about [34], which was published independently after the completion of our manuscript, and which studies the Community Detection problem on grids. In the framework of [34], nodes have location labels given by the coordinates of the integer grid and they ask the synchronization question of whether two nodes which are 'far-away' in space belong to the same or different communities. They study this synchronization problem using tools from percolation and random walks on grids which seem related to our methods. Nonetheless, our algorithm is completely different from theirs and in-fact a straight forward adaptation of our algorithm to their setting can give an alternative proof of Theorem 3 in [34].

From a modeling perspective, the work of [35] which considers the Latent Space Models as a part of its model is the closest to our model. They give a spectral algorithm that is proven to work in the logarithmic degree regime. However, the present paper considers the sparse graph case with node degrees a constant and not scaling with population size. This sparse regime requires completely different ideas from the analysis of the logarithmic degree regime.

1.3 Organization of the paper

We give a formal description of the model and the problem statement in Section 2. We then present our main theorem statements in Section 3. The subsequent sections will develop the ideas and the proofs needed for our main results. We describe our `GBG` Algorithm in Section 4 where we first give the idea and then the details of the algorithm. The analysis of our algorithm is performed in Section 5, where the key idea is to construct coupling arguments with site percolations. We establish the lower bound in Section 6, where we first introduce the Information Flow from Infinity problem and then prove that this is easier than Community Detection. Subsequently, we provide a proof of the impossibility result for Information Flow from Infinity. In Section 7, we consider the distinguishability problem and provide a proof of it. Sections 4,6 and 7 which contains the key technical ideas of the algorithm, lower bound and the distinguishability respectively, are independent of each other and can be read in any order.

2 Mathematical Framework and Problem Statement

We describe the mathematical framework based on stationary point processes and state the problem of Community Detection. We just set a common shorthand notation we use throughout the paper.

For an arbitrary $t \in \mathbb{R}_+$, we denote by the term $o_t(1)$ to be a bounded function $t \to a_t$ from $\mathbb{R}_+$ to $\mathbb{R}_+$ such that $\lim_{t\to\infty} a_t = 0$.

2.1 The Planted Partition Random Connection Model

For technical simplicity, we first define an infinite spatial random graph G and consider the Community Detection problem on appropriate finite sub-graphs G_n. We suppose there exists an abstract probability space $(\Omega, \mathcal{F}, \mathbb{P})$ on which we will construct our *infinite* random graph G which is parametrized by $\lambda \in \mathbb{R}_+$, $d \geq 2$ and two functions $f_{in}(\cdot), f_{out}(\cdot) : \mathbb{R}_+ \to [0,1]$ such that for all $r \geq 0$ we have $f_{in}(r) \geq f_{out}(r)$. The nodes of G are indexed using $\mathbb{N}$ and each node has two labels, one denoting its location in $\mathbb{R}^d$, and another being a community label taking values in $\{-1, +1\}$. We first sample the location and community labels of G by a marked point-process, where atoms of the point process represent the location labels on nodes and marks of the point-process represent community labels on nodes. Conditionally on the locations and the community labels, the edges of G will be drawn independently at random between pairs of nodes. Formally, we let $\phi := \{X_1, X_2, \cdots\}$ with $X_i \in \mathbb{R}^d$ for all $i \in \mathbb{N}$, to be a Poisson Point Process (PPP) [2] of intensity λ on $\mathbb{R}^d$. We further assume that $\phi = \{X_1, X_2, \cdots\}$ is enumerated in increasing l_∞ distance, i.e. for all $i \in \mathbb{N}$, $||X_{i+1}||_\infty \geq ||X_i||_\infty$. The interpretation is that node $i \in \mathbb{N}$ of G has for its location label, the random variable $X_i \in \mathbb{R}^d$. We further mark each atom $i \in \mathbb{N}$ of ϕ with random variables $Z_i \in \{-1, +1\}$ and $\{U_{ij}\}_{j\in\mathbb{N}\setminus\{i\}} \in [0,1]^{\mathbb{N}\setminus\{i\}}$ satisfying $U_{ij} = U_{ji}$ for all $i \neq j \in \mathbb{N}$. The sequence $\{Z_i\}_{i\in\mathbb{N}}$ is i.i.d. with each element being uniformly distributed in $\{-1, +1\}$. For every $i \in \mathbb{N}$, the sequence $\{U_{ij}\}_{j\in\mathbb{N}\setminus\{i\}}$ is i.i.d. with each element of $\{U_{ij}\}_{j\in\mathbb{N}\setminus\{i\}}$ being uniformly distributed on $[0,1]$. The interpretation of ϕ and its marks is that every node $i \in \mathbb{N}$ of G has location label given by $X_i \in \mathbb{R}^d$ and a community label given by $Z_i \in \{-1, +1\}$. This community label is moreover i.i.d. and independent of everything else. We will use the marks $\{U_{ij}\}_{j\in\mathbb{N}\setminus\{i\}}$ to sample the graph neighbors of a node i as follows. An edge between two nodes i and j where $i \neq j$ is present in G if and only if $U_{ij} = U_{ji} \leq f_{in}(||X_i - X_j||)\mathbf{1}_{Z_i=Z_j} + f_{out}(||X_i - X_j||)\mathbf{1}_{Z_i\neq Z_j}$. In other words, conditionally on the node labels, an edge is present between node i and j independent of other pair of edges with probability $f_{in}(||X_i - X_j||)$ if $Z_i = Z_j$ or with probability $f_{out}(||X_i - X_j||)$ if $Z_i \neq Z_j$. From Campbell's theorem, the average number of neighbors any node has from its own community is $(\lambda/2)\int_{x\in\mathbb{R}^d} f_{in}(||x||)dx$ and $(\lambda/2)\int_{x\in\mathbb{R}^d} f_{out}(||x||)dx$ in the opposite community. Hence, if $\int_{x\in\mathbb{R}^d} f_{in}(||x||)dx < \infty$, almost-surely, all nodes have finite degree and we call this regime as *sparse*. In the rest of this paper, we will assume that $\int_{x\in\mathbb{R}^d} f_{in}(||x||)dx < \infty$, i.e. the graph is sparse unless we state otherwise.

The random graph G can be viewed as a 'planted-partition' version of the classical random-connection model ([36]). Given $\lambda \in \mathbb{R}_+$, $g(\cdot) : \mathbb{R}_+ \to [0,1]$ and $d \geq 1$, the classical random-connection model $H_{\lambda,g(\cdot),d}$, is a random graph whose vertex set forms a homogeneous PPP of intensity λ on $\mathbb{R}^d$. Conditionally on the locations, edges in $H_{\lambda,g(\cdot),d}$ are placed independently of each other where two points at locations x and y of $\mathbb{R}^d$ are connected by an edge in $H_{\lambda,g(\cdot),d}$ with probability $g(||x-y||)$. This construction can be made precise by letting the edge random variables for each node be marks of the PPP similarly to the construction of G.

[2] This is defined in Appendix B

2.2 The Community Detection Problem

To state the Community Detection problem, we will consider appropriate finite restrictions of the infinite random graph G. Denote by $B_n := \left[-\frac{n^{1/d}}{2}, \frac{n^{1/d}}{2}\right]^d$, the cube of area n in $\mathbb{R}^d$. Let ϕ_n be the restriction of ϕ to B_n and G_n be the restriction of G to the nodes that are located in B_n. Denote by N_n the number of nodes of the graph G_n. From properties of the PPP, the random variable N_n is Poisson distributed with mean λn. Moreover, conditionally on N_n, these nodes are placed uniformly and independently in B_n. Recall that since the atoms of ϕ were numbered in increasing l_∞ distance, $N_n = \inf\{i \in \mathbb{N} : X_i \notin B_n\}$, i.e. the first N_n atoms of ϕ lie in B_n. Notice that the description given here and in the introduction are identical in distribution and hence we decide to use the same notation as the problem definition is only distributional as seen in Definition 2 below. Before describing the problem, we need the definition of 'overlap' between two sequences.

Definition 1. *Given a $t \in \mathbb{N}$, and two sequences $\mathbf{a}, \mathbf{b} \in \{-1, 1\}^t$, the* overlap *between* $\mathbf{a}$ *and* $\mathbf{b}$ *is defined as $\frac{|\sum_{i=1}^t a_i b_i|}{t}$, i.e. the absolute value of the normalized scalar product.*

Definition 2. *Community Detection is said to be* ***solvable*** *for $\lambda, d, f_{in}(\cdot)$ and $f_{out}(\cdot)$ if for every $n \in \mathbb{R}_+$, there exists a sequence of $\{-1, +1\}$ valued random variables $\{\tau_i^{(n)}\}_{i=1}^{N_n}$ which is a* deterministic *function of the observed data G_n* ***and*** *ϕ_n such that there exists a constant $\gamma > 0$ satisfying*

$$\liminf_{n\to\infty} \mathbb{P}\left[\mathcal{O}_n \geq \gamma\right] = 1, \tag{1}$$

where $\mathcal{O}_n$ is the overlap between $\{\tau_i^{(n)}\}_{i=1}^{N_n}$ and $\{Z_i\}_{i=1}^{N_n}$.

Note that we assume that the algorithm has access to the parameters $f_{in}(\cdot)$, $f_{out}(\cdot)$ and λ although this assumption may not always hold in practice. As mentioned, the estimation of model parameters from data itself will form an interesting technical question which we leave for future work. We take an absolute value in the definition of overlap since the distribution of G is symmetric in the community labels. In particular, if we flipped all community labels of G, we would observe a graph which is equal in distribution to G. Thus, any algorithm can produce the clustering only up-to a global sign flip, which we capture by considering the absolute value. We take finite restrictions B_n since the overlap is not well defined if $N_n = \infty$. A natural question then is of 'boundary-effects', i.e. the nodes near the boundary of B_n will have different statistics for neighbors than those far away from the boundary. However since G_n is sparse, except for a $o_n(1)$ fraction of nodes, all nodes in G_n will have the same degree as in the infinite graph G, i.e. the boundary effects are negligible.

A key new feature of our Definition 2 comes from our assumption that the algorithm has knowledge of all location labels on the nodes and it only needs to estimate the missing community labels. Our definition of detection which sometimes is also referred to as weak-recovery in the literature, deems an algorithm successful at community detection if it can correctly classify a fraction larger than half of the nodes. Observe that achieving $\gamma = 0$ (which amounts to correctly classifying exactly half the nodes) in the definition is trivial: this is obtained by associating a $+1$ to all nodes; by the Strong-Law of Large numbers, we will achieve an overlap of 0. There are other notions of recovery of interest such as strong recovery or exact-recovery which asks whether one can achieve $\gamma = 1 - \epsilon$ for all $\epsilon > 0$ or $\gamma = 1$ respectively. However, if the graph G is sparse, we show (in Corollary 3.1), that achieving $\gamma = 1 - \epsilon$ for all $\epsilon > 0$ is not possible. Thus, the interesting question in the sparse

regime is to consider when and how one can achieve any $\gamma > 0$.

The following elementary monotonicity property is evident from the definition of the problem and sets the stage for stating our main results.

Proposition 3. *For every $f_{in}(\cdot), f_{out}(\cdot)$ and d, there exists a $\lambda_c \in [0, \infty]$ such that*

- $\lambda < \lambda_c \implies$ *Community Detection is not solvable.*
- $\lambda > \lambda_c \implies$ *there exists an algorithm (which could possibly take exponential time) to solve Community Detection.*

Proof. Assume, for a given value of λ, there exists a community detection algorithm that achieves an overlap of $\gamma > 0$. Now, given any $\lambda^{'} > \lambda$, we will argue that we can achieve positive overlap. The proof of this follows from the basic thinning properties of the PPP. Given an instance of the problem with intensity $\lambda^{'}$, we will remove every node along with its incident edges independently with probability $1 - \frac{\lambda}{\lambda^{'}}$. We assign a community label estimate of $+1$ to all the removed nodes. For the nodes that remain (which is then an instance of the problem of community detection with intensity λ), we achieve an overlap of γ from the hypothesis that we can achieve positive overlap at intensity λ. Thus, from the independence of the thinning procedure and the community labels and strong law of large numbers, the overlap achieved by this process on an instance of intensity $\lambda^{'}$ will be at-least $\frac{\lambda\gamma}{\lambda^{'}}$ which is strictly positive. Thus, the problem of community detection solvability is monotone in λ. □

This proposition is not that strong since it does not rule out the fact that λ_c is either 0 or infinity. Moreover, this proposition does not tell us anything about polynomial time algorithms, just of the existence or non-existence of any (polynomial or exponential time) algorithms. The first non-trivial result would be to establish that $0 < \lambda_c < \infty$, i.e. the phase transition is strictly non-trivial and then to show that for possibly a larger constant, the problem is solvable efficiently. We establish both of this in Section 3.

Notation - Palm Probability

Before stating the results, we will need an important definition. We define the Palm Probability measure $\mathbb{P}^0$ of a point process in this subsection for future reference. Roughly stated, the Palm measure is the distribution of a marked point process ϕ 'seen from a typical atom'. We refer the reader to [37] for the general theory of Palm measures. However thanks to Slivnyak's theorem [37], we have a simple interpretation of $\mathbb{P}^0$ which is what we will use. The measure $\mathbb{P}^0$ is obtained by first sampling ϕ and G from $\mathbb{P}$ and placing an additional node indexed 0 at the origin of $\mathbb{R}^d$ and equipping it with independent community label and edges. The label of this node at origin will be denoted by $Z_0 \in \{-1, +1\}$ which is uniform and independent of anything else. Conditionally on Z_0, ϕ and the labels $\{Z_i\}_{i \in \mathbb{N}}$, we place an edge between node $i \in \mathbb{N}$ and this extra node at the origin with probability $f_{in}(||X_i||)\mathbf{1}_{Z_i = Z_0} + f_{out}(||X_i||)\mathbf{1}_{Z_i = -Z_0}$ independently of the edges between $j \neq i \in \mathbb{N}$ and the origin.

3 Main Results

3.1 Lower Bound for Community Detection

To state the main lower bound result in Theorem 4 below, we set some notation. For the random connection model graph $H_{\lambda,g(\cdot),d}$, denote by $\mathcal{C}_{H_{\lambda,g(\cdot),d}}(0)$ the set of nodes of $H_{\lambda,g(\cdot),d}$ that are in the same connected component as that of the node at the origin under the measure $\mathbb{P}^0$. Denote by $\theta(H_{\lambda,g(\cdot),d}) := \mathbb{P}^0[|\mathcal{C}_{H_{\lambda,g(\cdot),d}}(0)| = \infty]$ the percolation probability of the random graph $H_{\lambda,g(\cdot),d}$, i.e. the probability (under Palm) that the connected component of the origin has infinite cardinality.

Theorem 4. *If $\theta(H_{\lambda,f_{in}(\cdot)-f_{out}(\cdot),d}) = 0$, then Community Detection is not solvable.*

This theorem states that if the two functions $f_{in}(\cdot)$ and $f_{out}(\cdot)$ are not 'sufficiently far-apart', then no algorithm to detect the partition of nodes can beat a random guess. As a corollary, this says that Community Detection is impossible for $d = 1$.

Corollary 5. *For all $\lambda > 0$, $f_{in}(\cdot), f_{out}(\cdot)$ such that $\int_{x\in\mathbb{R}} f_{in}(||x||)dx < \infty$, Community Detection is not solvable if $d = 1$.*

Proof. This is based on the classical fact that for all $g(\cdot) : \mathbb{R}_+ \to \mathbb{R}_+$ such that $\int_{x\in\mathbb{R}} g(||x||)dx < \infty$, $\theta(H_{\lambda,g(\cdot),1}) = 0$. □

The following corollary gives a quantitative estimate of the percolation probability for higher dimensions in terms of the problem parameters.

Corollary 6. *For all $d \geq 2$, if $\lambda \leq \lambda_{lower} := (\int_{x\in\mathbb{R}^d}(f_{in}(||x||) - f_{out}(||x||))dx)^{-1}$, then Community Detection cannot be solved. Thus, $\lambda_c > (\int_{x\in\mathbb{R}^d}(f_{in}(||x||) - f_{out}(||x||))dx)^{-1}$.*

Proof. From classical results on percolation [36], by comparison with a branching process, we see that $\lambda \int_{x\in\mathbb{R}^d} g(||x||) \leq 1 \implies \theta(H_{\lambda,g(\cdot),d}) = 0$. □

Recall that if the graph G is sparse (finite average degree), then $\int_{x\in\mathbb{R}^d} f_{in}(||x||)dx < \infty$ which implies from Corollary 6 that λ_c is strictly positive in the sparse regime. The following proposition shows that this lower bound is tight for certain specific families of connection functions.

Proposition 7. *For all $d \geq 2$ and $R_1 \geq R_2$, if $f_{in}(r) = \mathbf{1}_{r\leq R_1}$ and $f_{out}(r) = \mathbf{1}_{r\leq R_2}$ and λ is such that $\theta(H_{\lambda,f_{in}(\cdot)-f_{out}(\cdot),d}) > 0$, then community detection can be solved in time proportional to n with the proportionality constant depending on the parameters λ, R_{in} and R_{out}.*

Hence, in view of Theorem 4, for $f_{in}(r) = \mathbf{1}_{r\leq R_1}$ and $f_{out}(r) = \mathbf{1}_{r\leq R_2}$ for some $R_1 \geq R_2$, then if λ is such that $\theta(H_{\lambda,f_{in}(\cdot)-f_{out}(\cdot),d}) = 0$, no algorithm (exponential or polynomial) time can solve Community Detection, while if λ is such that $\theta(H_{\lambda,f_{in}(\cdot)-f_{out}(\cdot),d}) > 0$, then a linear time algorithm exists to solve Community Detection. This gives a sharp phase-transition for this particular set of parameters where the problem shifts from being unsolvable even with unbounded computation to being efficiently solvable.

Algorithm and an Upper Bound for Community Detection

Our main result in the positive direction is our GBG algorithm described in Section 4. The main theorem statement on the performance of GBG is the following.

Theorem 8. *If $f_{in}(\cdot)$ and $f_{out}(\cdot)$ are such that $\{r \in \mathbb{R}_+ : f_{in}(r) \neq f_{out}(r)\}$ has positive Lebesgue measure and $d \geq 2$, then there exists a $\lambda_{upper} < \infty$ depending on $f_{in}(\cdot), f_{out}(\cdot)$ and d, such that for all $\lambda \geq \lambda_{upper}$, the GBG algorithm solves the Community Detection problem. Moreover, GBG when run on data (G_n, ϕ_n), has time complexity order n^2 and storage complexity order n.*

Remark 9. *An upper bound to λ_{upper} is given in Proposition 32.*

This gives a complete non-trivial phase-transition for the sparse graph case where we have $0 < \lambda_c < \infty$ which implies the existence of different phases. Moreover, we have quantitative bounds $\lambda_{lower} \leq \lambda_c \leq \lambda_{upper}$ for the critical value where the phase-transition occurs.

3.2 Identifiability of the Planted Partition

The key result we show here is that unlike in the traditional Erdős-Rényi setting, the planted partition random connection model is always mutually singular with respect to *any* random connection model without communities. Before precisely stating the result, we set some notation. Denote by $\mathbb{M}_{\mathcal{G}}(\mathbb{R}^d)$ the Polish space of all simple spatial graphs whose vertex set forms a locally finite set of $\mathbb{R}^d$. Thus, our random graph G or the random connection model $H_{\lambda,g(\cdot),d}$ can also be viewed through the induced measure on the space $\mathbb{M}_{\mathcal{G}}(\mathbb{R}^d)$.

Theorem 10. *For every $\lambda > 0$, d and connection functions $f_{in}(\cdot)$ and $f_{out}(\cdot)$ satisfying $1 \geq f_{in}(r) \geq f_{out}(r) \geq 0$ for all $r \geq 0$, and $\{r \geq 0 : f_{in}(r) \neq f_{out}(r)\}$ having positive Lebesgue measure and $g(\cdot) : \mathbb{R}_+ \to [0, 1]$, the probability measures induced on the space of spatial graphs $\mathbb{M}_{\mathcal{G}}(\mathbb{R}^d)$ by G and $H_{\lambda,g(\cdot),d}$ are mutually singular.*

This theorem gives in particular that G and $H_{\lambda,\frac{f_{in}(\cdot)+f_{out}(\cdot)}{2},d}$ are mutually singular. Note that if $g(\cdot) \neq \frac{f_{in}(\cdot)+f_{out}(\cdot)}{2}$, then the average degrees of G and $H_{\lambda,g(\cdot),d}$ are different and hence the empirical average of the degrees in G_n and $H_{\lambda,g(\cdot),d}$ restricted to B_n, will converge almost surely as $n \to \infty$ (thanks to the ergodic property of PPP) to the mean degree, thereby making the two induced measures mutually singular Thus, the only non-trivial random connection model that can possibly be not singular with respect to G is $H_{\lambda,\frac{f_{in}(\cdot)+f_{out}(\cdot)}{2},d}$, i.e. the case of equal average degrees. We show by a slightly different albeit similar ergodic argument in Section 7 that even in the case of equal average degrees, the two induced measures are mutually singular.

This theorem bears on the *distinguishability* problem which is a hypothesis testing question where one needs to predict whether the data (graph along with spatial locations) is drawn from G or $H_{\lambda,\frac{f_{in}(\cdot)+f_{out}(\cdot)}{2},d}$ with a success probability exceeding a half when the prior distribution is uniform over the two models. Theorem 10 implies that this hypothesis testing problem can be solved with probability of success $1 - o_n(1)$ for *all* parameter values. Thus, the distinguishability problem as stated in our spatial case exhibits no phase-transition. A consequence of our results is that in certain regimes ($\lambda < \lambda_c$ for $d \geq 2$ and $\lambda > 0$ for $d = 1$), we can be very sure by observing the data that a partition exists, but cannot identify it better than at random. Such phenomena was proven not to be observed in a symmetric SBM with two communities ([10]) and conjectured not to occur in any arbitrary SBM ([15]).

4 Algorithm for Performing Community Detection

In this section, we outline an algorithm called `GBG` described in Algorithm 3 that has time complexity of order n^2 and storage complexity of order n and solves the Community Detection problem for all sufficiently high λ.

4.1 Key Idea behind the Algorithm - Dense Local Interactions

The main and simple idea in our algorithm is that the graph G_n is 'locally-dense' even though it is globally sparse. This is in contrast to sparse Erdős-Rényi based graphs in which the local neighborhood of a typical vertex 'looks like a tree', our graph will have a lot of triangles due to Euclidean geometry. This simple observation that our graph is locally-dense enables us to propose simple pairwise estimators as described in Algorithm 1 which exploits the fact that two nodes 'nearby' in space have a lot of common neighbors (order λ). For concreteness, consider the case when $f_{in}(r) = a\mathbf{1}_{r\leq R}$ and $f_{out}(r) = b\mathbf{1}_{r\leq R}$ for some $R > 0$ and $0 \leq b < a \leq 1$. This means that points at Euclidean distance of R or lesser are connected by an edge in G with probability either a or b depending on whether the two points have the same community label or not. Moreover from elementary calculations, the number of common graph neighbors for any two nodes of G at a distance αR away for some $\alpha < 2$ is a Poisson random variable with mean either $\lambda c(\alpha)R^d(a^2+b^2)/2$ or $\lambda c(\alpha)R^d ab$ (for some constant $c(\alpha)$ that comes from geometric arguments) depending on whether the two nodes have the same or different community labels. Thus, using a simple strategy consisting of counting the number of common neighbors and thresholding gives a probability of mis-classifying any 'nearby' pair of nodes to be exponentially small in λ. We implement this simple idea in the sub-routine 1 below. Now, to produce the global partition one needs care to aggregate the pairwise estimates into a global partition. Since some pair-wise estimates are bound to be in error, we must identify them and avoid using those erroneous pair-wise estimates (see also Figure 1). We achieve this by classifying regions of space B_n as 'good' or 'bad' and then by considering the pair-wise estimates only in the 'good' regions. We prove that if λ is sufficiently large, then the 'good' regions will have sufficiently large volume and hence will succeed in detecting the communities better than at random.

We summarize our main algorithm below before presenting the formal pseudo-code.

- *Step 1* Partition the region B_n into small constant size cells and based on 'local-geometry' classify each cell as good or bad. This is accomplished in the `Is-A-Good` routine.
- *Step 2* Consider connected components of the Good cells and then in each of them apply the following simple classification rule. We enumerate the nodes in each connected component of Good cells in an arbitrary fashion subject to the fact that subsequent nodes are 'near-by'. Then we sequentially apply the `Pairwise-Classify` Algorithm given in 1.
- *Step 3* Do not classify the nodes in the bad cells and just output an estimate of $+1$ for them.

4.2 Notations and Definitions

In this section, we specify the needed notations for describing our algorithm. We will assume that the connection functions $f_{in}(\cdot)$ and $f_{out}(\cdot)$ satisfy the hypothesis of Theorem 8. Thus, there exists

$0 \leq \tilde{r} < R < \infty$ such that $f_{in}(r) > f_{out}(r)$ for all $r \in [\tilde{r}, R]$. In the rest of this section, we will use the $\tilde{r}$ and R coming from the connection functions.

To describe the algorithm, we need to set some notation. We partition the entire infinite domain $\mathbb{R}^d$ into good and bad regions. However this is just for simplicity and in practice, it suffices to do the partition for the region B_n. We first tessellate the space $\mathbb{R}^d$ into cubes of side-length $\frac{R}{4d^{1/d}}$ where R is as above. We identify the tessellation with the index set $\mathbb{Z}^d$, i.e. the cell indexed z is a cube of side-length $\frac{R}{4d^{1/d}}$ centered at the point $\frac{zR}{4d^{1/d}} \in \mathbb{R}^d$. The subset of $\mathbb{R}^d$ that corresponds to cell z is denoted by Q_z. Hence the cell indexed 0 is the cube of side-length $\frac{R}{4d^{1/d}}$ centered at the origin. We now give several definitions on the terminology used for the $\mathbb{Z}^d$ tessellation and not to be confused with the terminology for describing the graph G_n. We collect all the different notation and terminology in this sub-section for easier access and reference.

Definition 11. *A set $U \subseteq \mathbb{Z}^d$ is said to be $\mathbb{Z}^d$-**connected** if for every $x, y \in U$, there exists a $k \in \mathbb{N}$ and $x_1, \cdots x_k \in U$ such that for all $i \in [0, k+1]$, $||x_i - x_{i-1}||_\infty = 1$, where $x_0 := x$ and $x_{k+1} := y$.*

Definition 12. *For any $z \in \mathbb{Z}^d$, denote by $\mathbb{Z}^d$-**neighbors** of z the set of all $z^{'} \in \mathbb{Z}^d$ such that $||z - z^{'}||_\infty \leq 1$.*

Definition 13. *For any subset $A \subset \mathbb{Z}^d$ and any $k \in \mathbb{N}$, the k **thickening of** A is denoted by $\mathbf{L}_k(A) := \cup_{z \in A} \cup_{z' \in \mathbb{Z}^d : ||z - z'||_\infty \leq k} z^{'}$.*

Definition 14. *For any set $B \subseteq \mathbb{Z}^d$, denote by the set $Q_B := \cup_{z \in B} Q_z$.*

Definition 15. *Let $\mathcal{Z}(\cdot) : \mathbb{R}^d \to \mathbb{Z}^d$ be the projection function, i.e. $\mathcal{Z}(x) := \inf\{z \in \mathbb{Z}^d : ||\frac{Rz}{4d^{1/d}} - x||_\infty \leq 0.5\}$. In case, of more than one z achieving the minimum, we take the lexicographically smallest such z.*

Definition 16. *For any two points $x, y \in \mathbb{R}^d$, denote by $S_R(x, y) := B(x, R) \cap B(y, R)$, i.e. the intersection of two balls of radius R centered at points x and y.*

Definition 17. *For any two points $x, y \in \mathbb{R}^d$ such that $||x - y||_2 < R$, define by the two constants $M_{in}(x, y)$ and $M_{out}(x, y)$ as follows.*

$$M_{in}(x, y) = \int_{z \in S_R(x,y)} (f_{in}(||x - z||) f_{in}(||y - z||) + f_{out}(||x - z||) f_{out}(||y - z||))\, dz$$

$$M_{out}(x, y) = \int_{z \in S_R(x,y)} (f_{in}(||x - z||) f_{out}(||y - z||) + f_{out}(||x - z||) f_{in}(||y - z||)) dz$$

Observe that the definitions of $M_{in}(x, y)$ and $M_{out}(x, y)$ immediately give that

$$M_{in}(x, y) - M_{out}(x, y) = \int_{z \in S_R(x,y)} (f_{in}(||x - z||) - f_{out}(||x - z||))\, (f_{in}(||y - z||) - f_{out}(||y - z||))\, dz.$$

Definition 18. *For any two points $x, y \in \phi$, denote by $E_G^{(R)}(x, y)$ the number of common graph neighbors of x and y in G which are within a distance R from both x and y.*

4.3 Algorithm Description in Pseudo Code

We first present two sub-routines in Algorithms 1 and 2 that classify each cell of $\mathbb{R}^d$ to be either Good or Bad. The algorithm is parametrized by $\epsilon \in \left(0, \frac{1}{2}\right)$ which is arbitrary and fixed.

Algorithm 1 Pairwise Classifier

procedure PAIRWISE-CLASSIFY(i, j, ϕ, G)
  if $E_G^{(R)}(X_i, X_j) > \frac{\lambda}{2}\left(M_{in}(X_i, X_j) + M_{out}(X_i, X_j)\right)$ **then return** 1
  else
return -1
  end if
end procedure

In this algorithm, we classify two nodes as in the same partition if the number of common graph neighbors they have exceeds a threshold. The threshold is the average of the expected number of neighbors if the two nodes in consideration are of the same or opposite communities. Such simple tests suffices for our purpose, although one could imagine a more accurate estimator that also takes into account the number of nodes in $S_R(X_i, X_j)$ that do not have any edges to X_i and X_j; or the location labels of the common neighbors.

Algorithm 2 Is A-Good Testing

procedure IS-A-GOOD(z, G)
  if $|\phi \cap Q_z| < \lambda(R/4)^d(1/d)(1-\epsilon)$ **then return** FALSE
  end if
  $\phi^{(z)} := \phi \cap (\cup_{z':||z-z'||_\infty \le 1} Q_{z'})$
  for all $\forall k \ge 1$, and all $X_1, \cdots X_k \in \phi^{(z)}$ **do**
    if $\prod_{i=1}^{k}$ PAIRWISE-CLASSIFY$(X_i, X_{i+1}, G) = -1$ **then** ▷ Where $X_{k+1} := X_1$
      return FALSE
    end if
  end for
  return TRUE
end procedure

To understand the algorithm, we need some definitions which classify cells of $\mathbb{Z}^d$ into Good or Bad depending on the 'local graph geometry'.

Definition 19. *A cell Q_z is **A-Good** if*

1. $\left|\phi \cap Q_z\right| \ge \lambda\left(\frac{R}{4}\right)^d \frac{1}{d}(1-\epsilon)$*; and*
2. `Is-A-Good`(z, G) *returns TRUE*

*A cell is called **A-Bad** if it is not A-Good.*

The key idea of our simple algorithm lies in the definition of A-Good cells. We classify a cell to be A-Good if there are no 'inconsistencies' in the Pairwise-Estimates. See Figure 1 for an example of

Algorithm 3 GBG

1: **procedure** MAIN-ROUTINE(G_n, ϕ_n)
2: Classify each cell in B_n to be either A-Good or A-Bad using subroutine `Is-A-Good`.
3: Let $\mathcal{D}_1, \cdots \mathcal{D}_k$ be the A-Good $\mathbb{Z}^d$-connected components in B_n.
4: **for** $l = 1, l \leq k$ **do**
5: Let $X_{l_1}, \cdots X_{l_{n_j}} \in \phi_n \cap Q_{\mathcal{D}_j}$ be **maximal** and arbitrary s.t $||\mathcal{Z}(X_{l_o}) - \mathcal{Z}(X_{l_{o+1}})||_\infty \leq 1, \forall 1 \leq o \leq n_j - 1$
6: Set $\hat{\tau}^{(n)}_{l_1} = +1$
7: **for** $w = 2, w \leq n_j$ **do**
8: Set $\hat{\tau}^{(n)}_{l_w} =$ Pairwise-Classify$(l_{w-1}, l_w, \phi_n, G_n)\hat{\tau}^{(n)}_{l_w}$
9: **end for**
10: **end for**
11: **for** $c = 1, c \leq N_n$ **do**
12: **if** $\hat{\tau}^{(n)}_c = 0$ **then**
13: Set $\hat{\tau}^{(n)}_c = +1$
14: **end if**
15: **end for**
16: **return** $\{\hat{\tau}^{(n)}_i\}_{i=1}^{N_n}$
17: **end procedure**

pair-wise inconsistency due to the `Pairwise-Classify` algorithm. In words, a cell is A-Good, if among the nodes of G that either lie in the cell under consideration or in the neighboring cells, there are no inconsistencies in the output returned by the `Pairwise-Classify` algorithm. Moreover, one can test whether a cell is A-Good or not based on the data (ϕ, G) itself as done in Algorithm 2. Thus, we use the nomenclature of *Algorithm*-Good as A-Good.

The main routine in Algorithm 3 proceeds as follows. In Line 3, we extract out all A-Good connected cells in the spatial region B_n. Suppose that there are k A-Good connected components denoted by $\mathcal{D}_1, \cdots, \mathcal{D}_k$. Our algorithm looks at each connected component independently and produces a labeling of the nodes in them. In Line 5, we enumerate all nodes in any A-Good connected component $\mathcal{D}_l$ as $X_{l_1} \cdots X_{l_{n_l}}$ such that for all $1 \leq o < n_l$, we have $||\mathcal{Z}(X_{l_o}) - \mathcal{Z}(X_{l_{o+1}})||_\infty \leq 1$. Such an enumeration of any A-Good connected component is possible since by definition, every A-Good cell is non-empty of nodes. Now, we sequentially estimate the community labels in Line 8 using the `Pairwise-Classify` sub-routine applied on 'nearby' pairs of nodes. In Line 13, we assign an estimate of $+1$, i.e. extract no meaningful clustering for nodes that fall in A-Bad cells. See also Figure 2 for an illustration.

4.4 Complexity and Implementation

We discuss a simple implementation of our algorithm which takes time of order n^2 to run and storage space of order n. The multiplicative constants here depend on λ. We store the locations ϕ_n as a vector whose length is order λn and the graph G_n as an adjacency list. An adjacency list representation is appropriate since G_n is sparse and the average degree of any node is a constant (that depends on λ). Once we sample the locations ϕ_n, the graph G_n takes time of order n^2

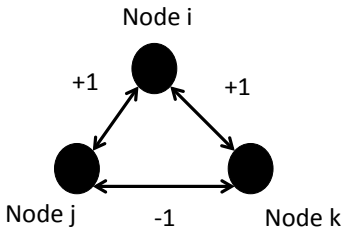

Figure 1: An illustration where Pairwise-Classify leads to inconsistency. The values on the edges represent the output of pairwise classify run on the two end points as inputs. In this example it is clear that for at-least one pair $(i,j),(j,k),(k,i)$, the output of pairwise estimate is different from the ground truth.

to sample. However, if one represented the locations of nodes more cleverly, then the sampling complexity could possibly be reduced from n^2. Moreover, since the average degree is a constant, the storage needed is order n. Given the data ϕ_n and the graph G_n, we pre-process this to store another adjacency list where for every vertex, we store the list of all other vertices within a distance of $2R$ from it. This preprocessing takes order n^2 time and order n space. The space complexity is order n since the graph is sparse. Equipped with this, we create a 'grid-list' where for each coordinate of $\mathbb{Z}^d$, we store the list of vertices whose location is in the considered grid cell. This takes just order n time to build. Moreover, since only a constant number of nodes are in any grid cell and the set B_n contains order n cells, the storage space needed for 'grid-list' is order n. Furthermore, since only a constant number of nodes are in a cell, it takes a constant time to test whether a particular cell is A-Good or A-Bad. Thus, to find $\mathbb{Z}^d$ connected components of Good-cells and produce the clustering takes another order nd time where d is the dimension. This gives our algorithm overall a time complexity of order n^2 and a storage complexity of order n.

5 Analysis and Proof of the Algorithm

The following theorem is the main theoretical guarantee on the performance of the `GBG` algorithm.

Theorem 20. *Let $\epsilon \in \left(0, \frac{1}{2}\right)$ be arbitrarily set in Algorithm 2. Let $\eta \in \left(0, \frac{1}{2}\right)$ be such that $\left(\frac{1}{2}+\eta\right)(1-\epsilon) > \frac{1}{2}$. Then there exists a constant $\lambda_0 < \infty$ depending on $f_{in}(\cdot), f_{out}(\cdot), d, \epsilon$ and η such that for all $\lambda > \lambda_0$, Algorithm 3 will solve Community Detection.*

Remark 21. *We provide an upper bound to the constant λ_0 in the sequel in Proposition 30.*

To prove the main result, we will need an additional classification of the cells of B_n as either T-Good or T-Bad. The nomenclature stands for *Truth*-Good.

Definition 22. *A cell Q_z is **T-Good** if -*

1. $\left|\phi \cap Q_z\right| \geq \lambda \left(\frac{R}{4}\right)^d \frac{1}{d}(1-\epsilon)$*; and*
2. *For all $i,j \in \mathbb{N}$ such that $X_i, X_j \in Q_{\mathbf{L}_1(z)} \cap \phi$, Pairwise-Classify$(i,j,\phi,G)$ returns $\mathbf{1}_{Z_i=Z_j} - \mathbf{1}_{Z_i \neq Z_j}$, i.e. the ground truth.*

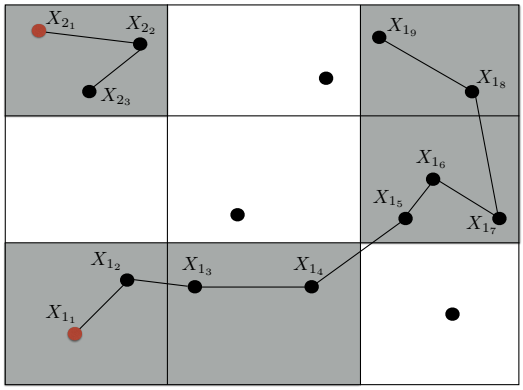

Figure 2: An illustration of algorithm 3. In this example, *we do not draw the graph* G, but only show the locations of the nodes. The shaded cells corresponds to A-Good cells and in this example there are two A-Good connected components. In each component, we outline an arbitrary sequence of points $X_{1_1} \cdots X_{1_9}$ and $X_{2_1}, X_{2_2}, X_{2_3}$ that will be used in line 4 of our main Algorithm 3. The lines then represent how we recursively set the community label estimates of the nodes as in line 8 of Algorithm 3. The estimates for the nodes in A-Bad cell is always set to 1.

If a cell is not T-Good, we call it ***T-Bad****.*

A cell is T-Good, if for any pair nodes which either lie in the cell under consideration or the neighboring cells, the output of the pairwise estimation matches the ground-truth. Of-course since the ground truth is unknown, one cannot test whether a cell is T-Good or not. We introduce the notion of a T-Good cell to aid in the analysis.

The proof of Theorem 20 can be split into three parts. The first part is composed of combinatorial arguments leveraging the definitions of A-Good and T-Good cells. These combinatorial lemmas will conclude that it suffices to ensure that there exists a 'giant' T-Good connected component in the data (ϕ_n, G_n). The next is a local analysis wherein we conclude that the probability a cell is T-Good can be made arbitrarily large by choosing the constant λ sufficiently high. The final step is to couple the process of T-Good cells to that of dependent site percolation on $\mathbb{Z}^d$ to conclude that if a single cell is T-Good with sufficiently high probability, then there exists a giant T-Good component.

5.1 Combinatorial Analysis

The main result in this sub-section we want to establish is the following statement. If we establish this, then the performance of our algorithm will follow from a study of the properties of the random graph G.

Proposition 23. *If there exists a connected component of T-Good cells in the data* (ϕ_n, G_n) *which contains a fraction of nodes of* G_n *strictly larger than a half with probability* $1 - o_n(1)$*, then the output returned by Algorithm 3 solves the Community Detection problem as given in Definition 2.*

The proof of the above proposition is based on the following two elementary combinatorial propositions.

Proposition 24. *If a cell Q_z is* ***T-Good****, then it is also* ***A-Good****. In particular, every connected* ***T-Good*** *component is contained in some connected* ***A-Good*** *component.*

Proof. It suffices to prove that for any $k \in \mathbb{N}$, $\prod_{i=1}^{k}(\mathbf{1}_{Z_i=Z_{i-1}} - \mathbf{1}_{Z_i \neq Z_{i-1}}) = 1$, where $X_k := X_0$ and $Z_k := Z_0$, i.e. a cycle. We can see this by contradiction. Assume $\prod_{i=1}^{k}(\mathbf{1}_{Z_i=Z_{i-1}} - \mathbf{1}_{Z_i \neq Z_{i-1}}) = -1$. This implies that an odd number of $-1's$ exists in the product. This can never be, since this would imply that Z_0 must be both simultaneously $+1$ and -1. Hence, such a product is always $+1$. □

The following proposition is the basis of Line 5 in the `GBG` in Algorithm 3. For every $z \in \mathbb{Z}^d$, denote by $\mathcal{A}(z)$ the *maximal* $\mathbb{Z}^d$ connected set containing z such that for all $u \in \mathcal{A}(z)$, cell u is A-Good.

Proposition 25. *For every $z \in \mathbb{Z}^d$ such that cell z is A-Good, there exists a* unique *partition of $\phi_{\mathcal{A}(z)} := \phi^{(+)}_{\mathcal{A}(z)} \coprod \phi^{(-)}_{\mathcal{A}(z)}$ such that for all $z, z' \in \mathcal{A}(z)$ with $||z - z'||_\infty \leq 1$ and all $X_i \in \phi \cap Q_z$ and $X_j \neq X_i \in \phi \cap Q_{z'}$, we have*

- *If $X_i \in \phi^{(+)}_z$ and $X_j \in \phi^{(-)}_z$ or , if $X_i \in \phi^{(-)}_z$ and $X_j \in \phi^{(+)}_z$, then Pairwise-Classify(i, j, G) will return -1.*
- *If $X_i, X_j \in \phi^{(+)}_z$ or if $X_i, X_j \in \phi^{(-)}_z$, then Pairwise-Classify$(i, j, G)$ returns $+1$.*

Moreover, the partition produced in Line 8 of our Algorithm 3 coincides with this partition.

This Proposition shows that all nodes inside A-Good connected components can be partitioned into two sets uniquely, such that the T-Good sub-component inside the A-Good component will be partitioned according to the underlying ground truth. Moreover, by following any arbitrary enumeration of the nodes of G as done in Line 5 of Algorithm 3, we can now build this unique partition of nodes of the A-Good component. This is what allows our algorithm to be fast.

Proof. We prove the proposition by induction. We will show that, given an unique and consistent partition of all nodes in cells $\mathcal{A}(z)$ upto l_∞ distance of n from z and a certain number of cells at distance $n+1$, we can uniquely extend the consistent partition to one more cell in $\mathcal{A}(z)$ at distance $n+1$ from z. In other words, we construct the unique partition in a 'breadth first manner' from z, by representing the distance on $\mathbb{Z}^d$ using the l_∞ distance. As a corollary, this proof technique establishes that the arbitrary sequence in Line 5 of Algorithm 3 will contain all the nodes of the component and will return the unique consistent partition.

For the base-case, since cell z is A-Good, from Line 6 of Algorithm 2, we can uniquely partition all points in cell z and its 1 step neighbors. The existence of a consistent partition can be argued as follows. Pick an arbitrary $X_i \in \phi \cap Q_z$ which we know by definition to be non-empty and label it $+1$. Now, for all points $X_j \in \cup_{z':||z-z'||_\infty}(\phi \cap Q_{z'})$, label X_j with the value returned by Algorithm 1 run on the input (i, j, G). Thus, we have produced a partition of all the points in the one-step neighbor of z in $\mathcal{A}(z)$ We will first argue that the partition we produced above is consistent, i.e. satisfies the conditions of the present proposition. Indeed, assume to the contrary, that this partition violates the statement of the current proposition. Assume there exists two points X_k and X_l such that they are in opposite sets of the partition with the partitioning procedure stated above whereas Algorithm 1 run on (k, l, G), returns a $+1$. This implies that the product of the outputs of Algorithm 1 run on input (i, k, G) with (k, l, G) and (i, l, G) is -1, thereby violating the fact that cell z is A-Good. We can similarly, argue the absence of two points X_k and X_l such that they are in the same partition,

but Algorithm 1 returns a -1. Now, to argue uniqueness, assume that on the contrary, there exists another consistent partition. This implies that there must exist two points X_i and X_j which are in the same set in one partition and in different sets in the other partition. Thus, clearly one of the partitions must violate the two requirements of this proposition, since Algorithm 1 run on (i, j, G) will produce just one output, thereby invalidating the consistency of at-least one the two partitions.

Now, we show the induction step. Assume there is an unique consistent partition of all nodes in $\phi \cap Q_{\mathcal{A}(z)}$ that are in cells of l_∞ distance of at-most n from z and some cells at a distance of $n+1$ from z. Let $z^{'} \in \mathcal{A}(z)$ be such that $||z - z^{'}||_\infty = n + 1$ and assume that $z^{'}$ has not yet been partitioned. Pick any $z^{''} \in \mathcal{A}(z)$ such that $||z - z^{''}||_\infty \leq n + 1$, $||z^{'} - z^{''}||_\infty = 1$, such that cell $z^{''}$ has already been partitioned. Note from the fact that $\mathcal{A}(z)$ is a connected subset of $\mathbb{Z}^d$, one can always find such a $z^{''}$. Pick $X_i \in \phi \cap Q_{z^{''}}$ arbitrarily. This can be done since we know that $\phi \cap Q_{z^{''}}$ is non-empty by definition of it being in $\mathcal{A}(z)$. To extend the partition, for every point $X_k \in \phi \cap Q_{z^{'}}$, run Algorithm 1 on the input (i, k, G). Place X_k in the same partition as X_i if the algorithm returned $+1$; else place X_k in the opposite partition of X_i. We will now conclude by showing that this extension still respects consistency and is unique. To argue uniqueness and consistency, it suffices to show that for all $X_k \in \phi \cap Q_{z^{'}}$ arbitrary no matter what $X_i \in Q_{z^{''}}$ we pick, the class to which X_k belongs is the same. Let $z^{'''}$ be arbitrary and such that $||z^{'} - z^{'''}||_\infty = 1$ cell and $z^{'''}$ is already partitioned. Let $i^{'} \in \phi \cap Q_{z^{'''}}$ be arbitrary. Now, the points $X_{i^{'}}, X_k, X_i$ are all within a 1 thickening of cell $z^{'}$ which we know is A-Good. Thus, from Line 6 of Algorithm 2, the product of Pairwise-Classify on inputs $(i^{'}, k), (k, i), (i, i^{'})$ is 1. By the induction hypothesis, there is a unique relation between i and $i^{'}$ (they are either in the same or different classes of the unique partition). Therefore, no matter the reference point, the class of a point $X_k \in \phi \cap Q_{z^{'}}$ is unique and consistent.

□

We are now in a position to conclude the proof of Proposition 23.

Proof. Proof of Proposition 23

Proposition 25 justifies Line 5 of Algorithm 3. First note that since every A-Good cell is non-empty of nodes of G, the arbitrary sequence in Line 5 of Algorithm 25 will enumerate all the nodes in each connected component. In other words, the only estimates that will be set in Line 13 of Algorithm 3 are those nodes that fall in the A-Bad cells. Moreover, the partition of the A-Good connected components in Line 8 will coincide with the partition referred to in Proposition 25. From the definition of T-Good components, the unique partition referred to in Proposition 25 will be such that the T-Good component will be partitioned according to the ground truth. Hence, if there exists a T-Good connected component that has a fraction $\alpha > \frac{1}{2}$ of the nodes of G_n, Algorithm 3 will partition this set of nodes in accordance to the ground truth. Thus, the achieved overlap will be at-least $2\alpha - 1 > 0$. This follows since the mis-classification of all nodes apart from this 'giant' connected T-Good component cannot diminish the overlap below $2\alpha - 1$ which is still positive. □

5.2 Local Analysis

The main goal of this subsection is to show that the probability a cell is T-Good can be made arbitrarily high by taking λ sufficiently high. In order to present the arguments, we will need

the definition of a generalized Palm distribution. For any $k \in \mathbb{N}$ and $x_1, \cdots, x_k \in \mathbb{R}^d$, we denote by $\mathbb{P}_{\phi}^{x_1,\cdots x_k}$ to be the Palm distribution of ϕ at $x_1, \cdots x_k$. This measure is the one induced by first sampling ϕ and G and then placing additional points at $x_1, \cdots, x_k$ and equipping them with independent community labels and edges. More precisely, we give these nodes i.i.d. uniform community labels $Z_{-1}, \cdots Z_{-k} \in \{-1, 1\}^k$. Conditionally on all the labels and ϕ, we draw an edge between any $i, j \in \{-k, -(k-1), \cdots\}$ such that at-least one of i or j belong to $\{-k, \cdots, -1\}$ as before, i.e. with probability $f_{in}(||X_i - X_j||)$ if the two nodes have the same community labels or with $f_{out}(||X_i - X_j||)$ if the two nodes have opposite community labels independently of other edges.

Proposition 26. *For any two $x \neq y \in \mathbb{R}^d$ such that $||x - y||_2 < 2R$, then conditionally on the labels of the points at x and y denoted as Z_x and Z_y respectively, we have for all $k \in \mathbb{N}$*

- *If $Z_x = Z_y$, then $\mathbb{P}^{x,y}[E_G^{(R)}(x, y) = k] = \frac{e^{-\lambda M_{in}(x,y)}(\lambda M_{in}(x,y))^k}{k!}$, i.e. is distributed as a Poisson random variable with mean $\lambda M_{in}(x, y)$.*
- *If $Z_x \neq Z_y$, then $\mathbb{P}^{x,y}[E_G^{(R)}(x, y) = k] = \frac{e^{-\lambda M_{out}(x,y)}(\lambda M_{out}(x,y))^k}{k!}$, i.e. is distributed as a Poisson random variable with mean $\lambda M_{out}(x, y)$.*

Proof. Slivnyak's theorem for independently marked PPP gives that conditionally on k points at locations $x_1, \cdots x_k \in \mathbb{R}^d$, the marked point process $\bar{\phi} \setminus \{x_1, \cdots, x_k\}$ has the same distribution as the original marked point process, i.e. is a PPP of intensity λ with independent marks. The independent thinning property of the PPP states that if any point at $x \in \phi$ is retained with probability $p(x)$ and deleted with probability $1 - p(x)$, independently of everything else, then the set of points not deleted forms a (potentially in-homogeneous) PPP.

Notice that the event that any $k \in \phi \setminus \{x, y\}$ such that $k \in B(x, R) \cap B(y, R)$ has an edge to both points x and y in G only depends on the location k and the community labels of points at locations k, x and y and is independent of everything else. Now, since the community labels are i.i.d. and independent of ϕ, the independent thinning property of PPP gives that the distribution of $E_G^{(R)}(x, y)$ is a Poisson random variable.

It remains to notice that the means are precisely $\lambda M_{in}(x, y)$ and $\lambda M_{out}(x, y)$. This follows from the Campbell - Mecke's theorem, that for any $F(\cdot) : \mathbb{R}^d \to \mathbb{R}_+$, we have for independently marked process is

$$\mathbb{E}_{\phi}^{x,y}\left[\sum_{z \in \phi \setminus \{x,y\}} F(z)\right] = \lambda \int_{z \in \mathbb{R}^d} \mathbb{E}_{\phi}^{x,y,z}[F(z)]dz. \tag{2}$$

Now, setting $F(z) := \mathbf{1}_{z \text{ has an edge to } x \text{ and } y} \mathbf{1}_{||z-x||_2 < R} \mathbf{1}_{||z-y||_2 < R}$ will conclude the statement on the means. □

Proposition 27. *For all connection functions $f_{in}(\cdot)$ and $f_{out}(\cdot)$ satisfying the hypothesis of Theorem 8, there exists a constant $c > 0$ such that for all $x \neq y \in \mathbb{R}^d$ satisfying $||x - y||_2 \leq (3/4)R$, we have*

$$\mathbb{P}^{x,y}[(x, y) \text{ is misclassified by Algorithm 1}] \leq e^{-c\lambda}, \tag{3}$$

where the constant c satisfies

$$c \geq \inf_{x,y \in \mathbb{R}^d : ||x-y||_2 \leq 3R/4} (\mathbf{1}_{M_{out}(x,y)>0} M_{out}(x,y) + \mathbf{1}_{M_{out}(x,y)=0} M_{in}(x,y)) h\left(\frac{M_{in}(x,y) - M_{out}(x,y)}{2M_{in}(x,y)}\right), \tag{4}$$

where $h(t) := (1+t)\log(1+t) - t$, for all $t \in \mathbb{R}_+$. In particular, $c > 0$ is strictly positive.

Proof. From Proposition 26, we know that $E_G^{(R)}(x,y)$ is either a Poisson random variable with mean $\lambda M_{in}(x,y)$ if the two nodes have the same community label or is a Poisson random variable of mean $\lambda M_{out}(x,y)$ if the two nodes have opposite community labels. Thus, the probability of mis-classification is then

$$\mathbb{P}^{x,y}[\text{points at } x \text{ and } y \text{ are mis-classified }] = \frac{1}{2}\mathbb{P}\left[X \geq \lambda \frac{M_{in}(x,y) + M_{out}(x,y)}{2}\right] + \frac{1}{2}\mathbb{P}\left[Y \leq \lambda \frac{M_{in}(x,y) + M_{out}(x,y)}{2}\right], \tag{5}$$

where X is a Poisson random variable of mean $\lambda M_{out}(x,y)$ and Y is a Poisson random variable of mean $\lambda M_{in}(x,y)$. The above interpretation is a probabilistic restatement of Algorithm 1. The coefficient $1/2$ denotes the case that the points at x and y could be in the same community or in opposite communities. Thus, by a basic application of Chernoff's bound, we have

$$\mathbb{P}^{x,y}[\text{points at } x \text{ and } y \text{ are mis-classified }] \leq \frac{1}{2} e^{-\lambda M_{out}(x,y) h\left(\frac{M_{in}(x,y) - M_{out}(x,y)}{2M_{in}(x,y)}\right)} + \frac{1}{2} e^{-\lambda M_{in}(x,y) h\left(\frac{M_{in}(x,y) - M_{out}(x,y)}{2M_{in}(x,y)}\right)}, \tag{6}$$

where $h(\cdot)$ is defined in the statement of the proposition.

Now under the assumptions on the connection functions $f_{in}(\cdot)$ and $f_{out}(\cdot)$, for all $r \in [\tilde{r}, R]$, $f_{in}(r) > f_{out}(r)$, we have that, $\inf_{x,y \in \mathbb{R}^d : ||x-y||_2 \leq (3/4)R} M_{in}(x,y) - M_{out}(x,y) > 0$. Moreover, since $M_{in}(x,y)$ and $M_{out}(x,y)$ are non-negative, $M_{in}(x,y) - M_{out}(x,y) > 0$ implies automatically that $M_{in}(x,y) > 0$ for all $x, y \in \mathbb{R}^d$ such that $||x-y||_2 \leq (3R/4)$. Hence, it follows that

$$\sup_{x,y \in \mathbb{R}^d : ||x-y||_2 \leq (3/4)R} \mathbb{P}^{x,y}[\text{points at } x \text{ and } y \text{ are mis-classified }] \leq e^{-c\lambda}, \tag{7}$$

where c is a strictly positive constant as given in the statement of the proposition.

□

Lemma 28. *For all $z \in \mathbb{Z}^d$,*

$$\mathbb{P}[\textit{Cell } z \textit{ is T-Good in graph } G] \geq 1 - e^{-\lambda (R/4)^d \frac{1}{d} h(\epsilon)} - \lambda^2 (3R/4)^d \frac{1}{d} e^{-c\lambda}, \tag{8}$$

where the constant c and function $h(\cdot)$ are defined in Proposition 27.

Proof. This follows from a basic union bound. We will prove an upper bound to a cell being T-Bad. A cell is T-Bad if either the number of points is smaller than $\lambda(R/4d^{1/d})^d(1-\epsilon)$ or there exists two points X_i and X_j in the 1 thickening of the cell $\{z\}$ such that when Algorithm 1 is run on input

(i, j, G), the returned answer is different from the truth.

From a simple Chernoff bound, the probability that a cell has fewer than $\lambda(R/4d^{1/d})^d(1-\epsilon)$ is at-most $e^{-\lambda(R/4d^{1/d})^d h(\epsilon)}$, where $h(\epsilon)$ is strictly positive for all $\epsilon > 0$.

We bound the probability that there exist two nodes that Algorithm 1 mis-classifies by the first moment method. We use the fact that if $X \geq 0$ is a $\mathbb{N}$ valued random variable, then $\mathbb{P}[X > 0] \leq \mathbb{E}[X]$. Hence, the probability that there exists a pair of points of ϕ that are mis-classified is bounded by the average number of pairs of points that are misclassified. Thus, for each cell z, we compute

$$\mathbb{E}[\sum_{i,j\in\mathbb{N}} \mathbf{1}_{X_i,X_j\in\mathbf{L}_1(z)}\mathbf{1}_{\text{ Algorithm 1 mis-classifies } i \text{ and } j}]. \tag{9}$$

From the Moment-Measure expansion and the Campbell-Mecke theorem for an independently marked PPP ([38]), we obtain

$$\begin{aligned}&\mathbb{E}[\sum_{i,j\in\mathbb{N}} \mathbf{1}_{X_i,X_j\in\mathbf{L}_1(z)}\mathbf{1}_{\text{ Algorithm 1 mis-classifies } i \text{ and } j}]\\ &= \lambda^2 \int_{x\in Q_{\mathbf{L}_1(z)}} \int_{y\in Q_{\mathbf{L}_1(z)}} \mathbb{P}^{x,y}[\text{points at } x \text{ and } y \text{ are mis-classified }]dxdy \leq \lambda^2 \left(\frac{3R}{4}\right)^d \frac{1}{d} e^{-c\lambda}. \end{aligned} \tag{10}$$

The last inequality follows directly from Proposition 27. Therefore, by a simple union bound, we see that

$$\mathbb{P}[\text{ Cell } z \text{ is T-Bad}] \leq e^{-\lambda(R/4d^{1/d})^d h(\epsilon)} + \lambda^2 \left(\frac{3R}{4}\right)^d \frac{1}{d} e^{-c\lambda}. \tag{11}$$

The proposition is proved by taking complements. □

Thus, we immediately have the following corollary which is what we will use in the sequel. The key fact to be used here is that the tessellation size R does not depend on λ and only depends on the connection functions $f_{in}(\cdot)$ and $f_{out}(\cdot)$.

Corollary 29. *For every $p \in (0,1)$, and every $f_{in}(\cdot)$ and $f_{out}(\cdot)$ satisfying the hypothesis of Theorem 8, there exists a $\lambda^{'}$ such that for all $\lambda > \lambda^{'}$, and all $z \in \mathbb{Z}^d$, $\mathbb{P}[$Cell z is T-Good$] \geq p$.*

Proof. It suffices to notice that for each fixed $f_{in}(\cdot)$, $f_{out}(\cdot)$ and d, we have

$$\lim_{\lambda\to\infty} p(\lambda) \geq \lim_{\lambda\to\infty} 1 - e^{-\lambda(R/4)^d \frac{1}{d} h(\epsilon)} - (\lambda^2+\lambda)(3R/4)^d \frac{1}{d} e^{-c\lambda} = 1, \tag{12}$$

where c is given in Proposition 27.

□

5.3 Global Analysis

In this section, we present the central tool required to analyze about the 'giant' connected T-Good component in the graph G_n. To do so, we exploit a coupling between the T-Good cells in the graph

G and a certain dependent site percolation process on $\mathbb{Z}^d$.

Denote by $(Y_z)_{z\in\mathbb{Z}^d}$ to be the random $0-1$ field on $\mathbb{Z}^d$ where $Y_z := \mathbf{1}_{\text{Cell } z \text{ is T-Good in } G}$. From the construction of the field, notice that the random field $(Y_z)_{z\in\mathbb{Z}^d}$ is only mildly dependent. Indeed, given any two $z, z^{'} \in \mathbb{Z}^d$, such that $||z - z^{'}||_1 \geq 12d^{1/d}$, we have that Y_z and $Y_{z^{'}}$ are independent random variables. This follows from the fact that we only look upto Euclidean distance of at-most $2R$ from any point inside a cell z to determine whether a cell is T-Good or T-Bad. Since, in an independently marked PPP, events corresponding to disjoint sets of $\mathbb{R}^d$ are independent, the claim follows.

We will now set some notation that will be useful in studying the process $(Y_z)_{z\in\mathbb{Z}^d}$. For any $z \in \mathbb{Z}^d$, cell z is **open in** $\mathbb{Z}^d$ if $Y_z = 1$. Similarly, any edge connecting z and $z^{'}$ is said to be open if both its end points are open. For any $z \in \mathbb{Z}^d$, we denote by $\mathcal{C}(z)$ to be the maximal connected random subset of $\mathbb{Z}^d$ containing z such that all $z^{'} \in \mathcal{C}(z)$ satisfies $Y_{z^{'}} = 1$. The main proposition we want to establish in this section is the following.

Proposition 30. *For every $\eta \in \left(0, \frac{1}{2}\right)$, there exists $\lambda_0(\eta, \epsilon) < \infty$ (where ϵ is set in Algorithm 2) chosen sufficiently high (as a function of $f_{in}(r), f_{out}(r), r \in [0, R]$ and d), such that for all $\lambda > \lambda_0(\eta, \epsilon)$ and all $j \in \mathbb{Z}^d$*

$$\liminf_{n\to\infty} \frac{1}{(2n)^d} \sum_{i\in\mathbb{Z}^d : ||i-j||_\infty \leq n} \mathbf{1}_{i\in\mathcal{C}(j)} \geq \frac{1}{2} + \eta, \tag{13}$$

$\mathbb{P}$ *almost-surely on the event that $\{|\mathcal{C}(j)| = \infty\}$. Moreover, for $\lambda > \lambda_0(\eta, \epsilon)$, and all $j \in \mathbb{Z}^d$, $\mathbb{P}[|\mathcal{C}(j)| = \infty] \geq \frac{1}{2} + \eta$ and $\mathbb{P}[\exists j \in \mathbb{Z}^d : |\mathcal{C}(j)| = \infty] = 1$.*

The key insight out of the proposition we want is to ensure that by taking λ sufficiently high, there exists an infinite open component in the process $(Y_z)_{z\in\mathbb{Z}^d}$, i.e. there exists $z \in \mathbb{Z}^d$ such that $|\mathcal{C}(z)| = \infty$. Moreover, we want to show that this infinite component contains more than half of the sites of $\mathbb{Z}^d$. The reason this does not immediately follow from Corollary 29 is that we have not yet established that the infinite open component in $(Y_z)_{z\in\mathbb{Z}^d}$ if it exists is unique. However [27] provides a clean 'black-box' methodology to establish this and our proposition can be viewed as a direct corollary of Theorem 1 in [27]. We will first dominate the process $(Y_z)_{z\in\mathbb{Z}^d}$ by an independent percolation process which is known to have a unique infinite component and then leverage this domination to conclude the proposition.

Proof. Notice that the process $(Y_z)_{z\in\mathbb{Z}^d}$ is $M := \lceil 12d^{1/d} \rceil$ dependent. Moreover, thanks to Proposition 27, for every $z \in \mathbb{Z}^d$,

$$\mathbb{P}[Y_z = 1 | \sigma(Y_u : u \in \mathbb{Z}^d, ||u - z||_\infty > M)] \geq p(\lambda), \ \mathbb{P} \text{ a.s. }, \tag{14}$$

where $p(\lambda) \to 1$ as $\lambda \to \infty$.

Thus, from Theorem 1 in [27], the law of $(Y_z)_{z\in\mathbb{Z}^d}$ stochastically dominates that of i.i.d. Bernoulli $\tilde{p}(\lambda)$ random variables where $\tilde{p}(\lambda)$ converges to 1 as $p(\lambda)$ converges to 1. More precisely, Theorem 1 from [27] gives the existence of a probability space $(\Omega^{'}, \mathcal{F}^{'}, \mathbb{P}^{'})$ containing two sequences of $\{0, 1\}$ valued random variables $(Y_z^{'})_{z\in\mathbb{Z}^d}$ and $(\tilde{Y}_z^{'})_{z\in\mathbb{Z}^d}$ such that

- The distribution of $(Y_z^{'})_{z\in\mathbb{Z}^d}$ is the same as that of $(Y_z)_{z\in\mathbb{Z}^d}$.
- For all $z \in \mathbb{Z}^d$, $Y_z^{'} \geq \tilde{Y}_z^{'}$, $\mathbb{P}^{'}$ almost-surely.
- $\mathbb{P}^{'}[\tilde{Y}_z^{'} = 1 | \sigma(\tilde{Y}_u^{'} : u \in \mathbb{Z}^d \setminus \{z\})] = \tilde{p}(\lambda)$, $\mathbb{P}^{'}$ almost-surely. In other words, $(\tilde{Y}_z^{'})_{z\in\mathbb{Z}^d}$ is an i.i.d. sequence of Bernoulli random variables with success probability $\tilde{p}(\lambda)$.
- $\tilde{p}(\lambda)$ converges to 1 as $p(\lambda)$ converges to 1.

Denote by $\mathcal{C}^{'}(0)$ and $\tilde{C}^{'}(0)$ the cluster at the origin of the process $(Y_z^{'})_{z\in\mathbb{Z}^d}$ and $(\tilde{Y}_z^{'})_{z\in\mathbb{Z}^d}$ respectively. Denote by $\theta_d(\lambda) := \mathbb{P}^{'}[|\tilde{C}^{'}(0)| = \infty]$. From a direct application of Peirl's argument ([25], Chapter 1), it is also well know that $\theta_d(\lambda) \to 1$ as $\tilde{p}(\lambda) \to 1$. Thanks to Line 4 above, we have $\theta_d(\lambda) \to 1$ as $p(\lambda) \to 1$. From Corollary 29, this can be rephrased as $\lim_{\lambda\to\infty} \theta_d(\lambda) = 1$.

The stochastic domination in Line 2 above yields

$$\frac{1}{(2n)^d} \sum_{i\in\mathbb{Z}^d : ||i-j||_\infty \leq n} \mathbf{1}_{i\in\mathcal{C}^{'}(j)} \geq \frac{1}{(2n)^d} \sum_{i\in\mathbb{Z}^d : ||i-j||_\infty \leq n} \mathbf{1}_{i\in\tilde{\mathcal{C}}^{'}(j)} \ \mathbb{P}^{'} \text{ a.s.} \tag{15}$$

On the event that $|\tilde{\mathcal{C}}^{'}(j)| = \infty$, we have

$$\frac{1}{(2n)^d} \sum_{i\in\mathbb{Z}^d : ||i-j||_\infty \leq n} \mathbf{1}_{i\in\mathcal{C}^{'}(j)} \geq \frac{1}{(2n)^d} \sum_{i\in\mathbb{Z}^d : ||i-j||_\infty \leq n} \mathbf{1}_{|\tilde{\mathcal{C}}^{'}(i)|=\infty} \ \mathbb{P}^{'} \text{ a.s.} \tag{16}$$

This follows from the well known fact that in an independent site percolation process that the infinite component if it exists is unique. In other-words, for all $i, j \in \mathbb{Z}^d$, $|\tilde{\mathcal{C}}^{'}(i)| = \infty$ and $|\tilde{\mathcal{C}}^{'}(j)| = \infty$ implies $\tilde{\mathcal{C}}^{'}(i) = \tilde{\mathcal{C}}^{'}(j)$, $\mathbb{P}^{'}$ almost-surely. Now, taking a limit on both sides, we get that

$$\liminf_{n\to\infty} \frac{1}{(2n)^d} \sum_{i\in\mathbb{Z}^d : ||i-j||_\infty \leq n} \mathbf{1}_{i\in\mathcal{C}^{'}(j)} \geq \liminf_{n\to\infty} \frac{1}{(2n)^d} \sum_{i\in\mathbb{Z}^d : ||i-j||_\infty \leq n} \mathbf{1}_{|\tilde{\mathcal{C}}^{'}(i)|=\infty} \ \mathbb{P}^{'} \text{ a.s.} \tag{17}$$

From Birkhoff's ergodic theorem, it is well known that for all $j \in \mathbb{Z}^d$,

$$\lim_{n\to\infty} \frac{1}{(2n)^d} \sum_{i\in\mathbb{Z}^d : ||i-j||_\infty \leq n} \mathbf{1}_{|\tilde{\mathcal{C}}^{'}(i)|=\infty} = \theta_d(\lambda) \ \mathbb{P}^{'} \text{ a.s.} \tag{18}$$

But since $\lim_{\lambda\to\infty} \theta_d(\lambda) = 1$, for every η and ϵ, we can take $\lambda_0(\eta, \epsilon)$ sufficiently large so that $p(\lambda)$ is sufficiently large which in turn indicates $\tilde{p}(\lambda)$ is sufficiently large so that $\theta_d(\lambda) \geq \frac{1}{2} + \eta$. The proof is concluded by observing that $(Y_z^{'})_{z\in\mathbb{Z}^d} \overset{(d)}{=} (Y_z)_{z\in\mathbb{Z}^d}$.

□

5.4 Concluding that Community Detection is Solvable

The following proposition along with Proposition 23 will conclude the proof of Theorem 20.

Proposition 31. *Let $\epsilon \in \left(0, \frac{1}{2}\right)$ be set in Algorithm 2. For all $\eta \in \left(0, \frac{1}{2}\right)$ such that $(1-\epsilon)\left(\frac{1}{2} + \eta\right) > \frac{1}{2}$, for all $\lambda \geq \lambda_0(\epsilon, \eta)$ where $\lambda_0(\epsilon, \eta)$ is from Proposition 30, the fraction of nodes of G_n that lie in the largest T-Good component, denoted by $\alpha_n \in [0, 1]$ is such that $\liminf_{n\to\infty} \alpha_n > \frac{1}{2}$, $\mathbb{P}$ almost-surely.*

Proof. Observe that the definition of a cell being A-Good or A-Bad is spatially 'local'. More precisely, for all $z \in \mathbb{Z}^d$ such that $z + B(0, 2R) \in B_n$, the event that cell z being A-Good in G_n is the same as cell being A-Good in G. We call cells $z \in \mathbb{Z}^d$ such that $z + B(0, 2R) \in B_n$ *internal* to B_n. Observe that all $z \in \mathbb{Z}^d$ is eventually internal to B_n for all n large enough. Moreover, since each cell is of side $R/(4d^{1/d})$, B_n has at-most $\lceil (4n^{1/d}/Rd^{1/d})^d \rceil$ cells out-of which at-least $\lfloor (4n^{1/d}/Rd^{1/d})^d \rfloor - \lceil 8dn^{1/d} \rceil$ cells are 'internal' to B_n. Thus, the fraction of cells in B_n that are internal to B_n is $1 - o_n(1)$.

From Proposition 30, we know that $\mathbb{P}[|\mathcal{C}(0)| = \infty] \geq \frac{1}{2} + \eta$ and $\mathbb{P}[\exists z \in \mathbb{Z}^d : |\mathcal{C}(z)| = \infty] = 1$. Moreover on the event $\{|\mathcal{C}(z)| = \infty\}$, we know from Proposition 30 that

$$\liminf_{n\to\infty} \frac{1}{(2n)^d} \sum_{i \in \mathbb{Z}^d : ||i-z||_\infty \leq n} \mathbf{1}_{i \in \mathcal{C}(z)} \geq \frac{1}{2} + \eta \ \mathbb{P} \text{ a.s.} \tag{19}$$

However, from an elementary counting argument, we conclude that

$$\liminf_{n\to\infty} \frac{1}{(2n)^d} \sum_{i \in \mathbb{Z}^d : ||i||_\infty \leq n} \mathbf{1}_{i \in \mathcal{C}(z)} \geq \frac{1}{2} + \eta \ \mathbb{P} \text{ a.s.} \tag{20}$$

In other words, the reference point does not matter when considering the limit, which can be seen by the following :

$$\begin{aligned} \frac{1}{(2(n+z))^d} \sum_{i \in \mathbb{Z}^d : ||i-z||_\infty \leq n} \mathbf{1}_{i \in \mathcal{C}(z)} &\leq \frac{1}{(2n)^d} \sum_{i \in \mathbb{Z}^d : ||i-z||_\infty \leq n} \mathbf{1}_{i \in \mathcal{C}(z)} \\ &\leq \frac{1}{(2(n+z))^d} \left(\sum_{i \in \mathbb{Z}^d} \mathbf{1}_{||i-z||_\infty \geq n} \mathbf{1}_{||i||_\infty \leq z+n} + \mathbf{1}_{||i-z||_\infty \leq n} \mathbf{1}_{i \in \mathcal{C}(z)} \right). \end{aligned} \tag{21}$$

But since for every fixed $z \in \mathbb{Z}^d$

$$\begin{aligned} \frac{1}{(2(n+z))^d} \left(\sum_{i \in \mathbb{Z}^d} \mathbf{1}_{||i-z||_\infty \geq n} \mathbf{1}_{||i||_\infty \leq z+n} + \mathbf{1}_{||i-z||_\infty \leq n} \mathbf{1}_{i \in \mathcal{C}(z)} \right) - \frac{1}{(2(n+z))^d} \sum_{i \in \mathbb{Z}^d : ||i-z||_\infty \leq n} \mathbf{1}_{i \in \mathcal{C}(z)} \\ = \mathrm{O}\big(n^{1-d}\big), \end{aligned} \tag{22}$$

it follows that for all $z \in \mathbb{Z}^d$

$$\liminf_{n\to\infty} \frac{1}{(2n)^d} \sum_{i \in \mathbb{Z}^d : ||i-z||_\infty \leq n} \mathbf{1}_{i \in \mathcal{C}(z)} = \liminf_{n\to\infty} \frac{1}{(2n)^d} \sum_{i \in \mathbb{Z}^d : ||i||_\infty \leq n} \mathbf{1}_{i \in \mathcal{C}(z)}. \tag{23}$$

Let $z \in \mathbb{Z}^d$ be arbitrary and condition on the event $\{|\mathcal{C}(z)| = \infty\}$. On this event, Equation (23) and Proposition 30 along with the fact that the fraction of cells in B_n that are internal is $1 - o_n(1)$ give that the fraction of internal cells in B_n in the connected T-Good component of cell z (i.e. in $\mathcal{C}(z)$) is $\frac{1}{2} + \eta - o_n(1)$. Since there are at-least $\lambda(R/4)^d(1/d)(1-\epsilon)$ nodes of G in each T-Good cell, the number of nodes of G_n in this T-Good connected component is at-least $(\lfloor (4n^{1/d}/R)^d \rfloor - \lceil 8dn^{1/d} \rceil) \left(\frac{1}{2} + \eta\right) \lambda(R/4)^d(1-\epsilon) > \frac{1}{2}(\lfloor (4n^{1/d}/R)^d \rfloor - \lceil 8dn^{1/d} \rceil)\lambda(R/4)^d$ since

we assumed that $(1-\epsilon)\left(\frac{1}{2}+\eta\right) > \frac{1}{2}$. Moreover, from elementary Chernoff and Borell Cantelli arguments, we get that for every fixed $\epsilon^{'} > 0$, there exists a random $n_{\epsilon'}$ such that for all $n \geq n_{\epsilon'}$, the number of nodes in G_n is less than or equal to $\lceil (4n^{1/d}/R)^d \rceil \lambda (R/4)^d (1+\epsilon^{'})$ almost-surely. Now, fix an $\epsilon^{'} > 0$, such that there exists a $\gamma > 0$ satisfying $\frac{(1/2+\eta)(1-\epsilon)}{(1+\epsilon^{'})} = \frac{1}{2} + \gamma$. Thus, for n larger than $n_{\epsilon'}$, the fraction of nodes in G_n lying the T-Good component of cell z is α_n, where

$$\alpha_n \geq \frac{\left(\lfloor (4n^{1/d}/R)^d \rfloor - \lceil 8dn^{1/d} \rceil\right)\left(\frac{1}{2}+\eta\right)\lambda (R/4)^d (1-\epsilon)}{\lceil (4n^{1/d}/R)^d \rceil \lambda (R/4)^d (1+\epsilon^{'})} \geq \frac{1}{2} + \gamma - o_n(1) \tag{24}$$

almost-surely, i.e., $\lim_{n\to\infty} \mathbb{P}\left[\alpha_n > \frac{1}{2} \middle| |\mathcal{C}(z)| = \infty\right] = 1$. But since $\mathbb{P}[\exists z \in \mathbb{Z}^d : |\mathcal{C}(z)| = \infty] = 1$, we can drop the conditioning on the event $\{|\mathcal{C}(z)| = \infty\}$ and conclude that with probability 1, a fraction of nodes of G_n strictly larger than half lie in a connected T-Good component.

□

5.5 An Upper Bound to the Constant $\lambda(\epsilon, \eta)$

In this section, we can use Peirl's argument directly to provide an upper bound λ_{upper} as a function of $f_{in}(\cdot).f_{out}(\cdot)$ and d such that `GBG` will perform community detection for all $\lambda \geq \lambda_{upper}$. This is just a sufficient condition and we present it here for completeness although the expression by itself provides no further insight into the problem.

Proposition 32. *If λ satisfies*

$$\frac{1}{3}\sum_{n\geq M} n \left(3\left(e^{-\lambda (R/4d^{1/d})^d h(\epsilon)} + \lambda^2 \left(\frac{3R}{4}\right)^d \frac{1}{d} e^{-c\lambda}\right)^{1/M}\right)^n \leq \frac{1}{2} - \eta, \tag{25}$$

where the constant c is given in Proposition 27, $M := \lceil 12d^{1/d} \rceil$, ϵ is set in Algorithm 2 and $h(x) := (1+x)\log(1+x) - x$, then $\lambda \geq \lambda_0(\epsilon, \eta)$.

This proof is a standard application of Peirl's estimate similar to Chapter 2 [25]. We outline it in the Appendix C for the sake of completeness. Observe that we did not use Peirl's argument directly in Proposition 30 since we did not prove that the infinite component is unique, which we needed there.

6 Lower Bound for Community Detection

The goal of this section is to prove Theorem 4. The central idea is to consider the problem of how well can one estimate whether two uniformly randomly chosen nodes of G_n belong to the same or opposite communities better than at random. This problem is indeed easier than Community Detection which requires one to produce an entire partition of the nodes of G_n. We will show that the natural way to understand the pairwise classification problem is through another problem which we call 'Information Flow through Infinity' which we define in the sequel in Section 6.2. Informally, this problem asks whether one can estimate with success probability larger than a half, the community label of any node chosen uniformly at random from G_n, given the graph,

the spatial locations **and** the true community labels of all nodes whose spatial locations are *far* away (at infinity) from this chosen node. Subsequently, the core technical argument of this section is to establish an impossibility result for Information Flow from Infinity which we state below in Theorem 38. To aid us in developing the technical arguments, it is instructive to first consider the proof of Proposition 7 (which was stated in Section 3), which identifies a special case of connection functions $f_{in}(\cdot)$ and $f_{out}(\cdot)$ when the phase-transition is sharp.

6.1 Proof of Proposition 7

Let $R_{in} > R_{out} \geq 0$ be arbitrary and consider the two functions to be $f_{in}(r) = \mathbf{1}_{r \leq R_{in}}$ and $f_{out}(r) = \mathbf{1}_{r \leq R_{out}}$. In words, two points of opposite communities are connected if and only if their distance is lesser than R_{out} and two points of the same community are connected if and only if their distance is smaller than R_{in}. In this example, it is clear that for any two points $X_i, X_j \in \phi$, no matter their community labels Z_i and Z_j, if $||X_i - X_j||_2 \leq R_{out}$, then i and j are always connected in G. Similarly, any two points X_i and X_j such that $||X_i - X_j||_2 > R_{in}$ are never connected by an edge in G no matter their community labels Z_i and Z_j. Hence, the *informative* pairs of points in this example are those X_i, X_j such that $||X_i - X_j||_2 \in (R_{out}, R_{in}]$. Moreover, it is immediate that, if $||X_i - X_j||_2 \in (R_{out}, R_{in}]$ and $i \sim_G j$, then $Z_i = Z_j$. On the other hand if $||X_i - X_j||_2 \in (R_{out}, R_{in}]$ and $i \nsim_G j$, then it must be the case that $Z_i \neq Z_j$. For any two points X_i and X_j such that $||X_i - X_j||_2 \in [0, R_{out}] \cup (R_{in}, \infty)$, the presence or absence of an edge is not informative as it is a certain event.

This example motivates the following simple algorithm for Community Detection. Partition the nodes of G_n into $\mathcal{D}_1, \cdots \mathcal{D}_k$ where each component $\mathcal{D}_i$ is a maximal set of nodes $\{X_{i_1}, \cdots X_{i_{l_i}}\}$ of G_n such that for all $j \in [1, l_i]$, we have $||X_{i_{j-1}} - X_{i_j}|| \in (R_{out}, R_{in}]$. In words, we form another graph T_n from the points ϕ_n such that any two nodes i and j of G_n are connected in T_n if and only if $||X_i - X_j|| \in (R_{out}, R_{in}]$. Then $\mathcal{D}_1, \cdots \mathcal{D}_k$ are the connected components of the graph T_n. The algorithm works by considering and labeling each connected component $\mathcal{D}_i$ independently of other components. For each cluster $i \in [1, k]$, estimate the node label of X_{i_1} to be $+1$. Then for every $j \in [2, l_i]$, recursively estimate the node label by the following procedure-

- If $i_{j-1} \sim_{G_n} i_j$ then set $Z_{i_j} = Z_{i_{j-1}}$.
- If $i_{j-1} \not\sim_{G_n} i_j$ then set $Z_{i_j} = -Z_{i_{j-1}}$.

This algorithm considers each of the connected component of T_n enumerated in an arbitrary manner and then labels the nodes in these components. The following very elementary proposition explains when this algorithm will perform well.

Proposition 33. *Let $R_{in} > R_{out} \geq 0$ be arbitrary such that $f_{in}(r) = \mathbf{1}_{r \leq R_{in}}$ and $f_{out}(r) = \mathbf{1}_{r \leq R_{out}}$. If $\theta(H_{\lambda, f_{in}(\cdot) - f_{out}(\cdot), d}) > 0$, then the procedure described above solves Community Detection for this set of parameters.*

Note that in view of Theorem 38, Proposition 33 will imply Proposition 7.

Proof. Notice that if $f_{in}(r) = \mathbf{1}_{r \leq R_{in}}$ and $f_{out}(r) = \mathbf{1}_{r \leq R_{out}}$, then $f_{in}(r) - f_{out}(r) = \mathbf{1}_{R_{out} \leq r < R_{in}}$. From the properties of the construction of the graph, any two $i \neq j \in \mathbb{N}$ such that $||X_i - X_J|| \in (R_{out}, R_{in}]$ satisfies -

- $Z_i = Z_j$ if $i \sim_G j$
- $Z_i \neq Z_j$ if $i \nsim_G j$.

Hence, it is clear that the algorithm described in the preceding paragraph partitions each cluster $\mathcal{D}_i$, $i \in [1, k]$ exactly in accordance to the ground truth. However, it could be that the estimated signs in each of the connected components $\mathcal{D}_i$ could be flipped from the underlying ground truth and hence the achieved overlap can still be small even though we partition each cluster $\mathcal{D}_i$ accurately. To argue that the overlap achieved by the algorithm is not too small, a sufficient condition is that there exists a unique giant (of size $cn - o(n)$ for some $c > 0$) component of T_n and all other connected components are $o(n)$. Then, we will have by the strong-law of large numbers that the overlap achieved will be c, i.e. the mislabeling in all small components will 'cancel' each other out and in particular cannot drive the overlap of c achieved in the giant component to 0. From the definition of percolation, a unique giant component in T_n exists if and only if $\theta(H_{\lambda, f_{in}(\cdot) - f_{out}(\cdot), d}) > 0$ since $T_n \overset{(d)}{=} H_{\lambda, f_{in}(\cdot) - f_{out}(\cdot), d}$.

□

In the sequel, we will generalize the above example to come up with the general lower bound for Community Detection problem. To do so, we introduce another problem of Information Flow from Infinity, which is new to the best of our knowledge and can be of independent interest. We show this is an *easier* problem than Community Detection and hence an impossibility result for this translates to an impossibility for Community Detection.

6.2 The Information Flow from Infinity Problem

This problem refers to how well can one estimate the community label of a tagged node of a graph better than at random, given some extra 'information at infinity'. We make this problem precise by posing this question under the Palm Probability measure $\mathbb{P}^0$. Recall that the Palm measure is the distribution of the graph G 'seen from a typical node'. Slivnyak's theorem [37] gives that the measure $\mathbb{P}^0$ is obtained by first sampling ϕ and G from $\mathbb{P}$ and placing an additional node indexed 0 at the origin of $\mathbb{R}^d$ and equipping it with independent community label and edges. The label of this node at origin is denoted by $Z_0 \in \{-1, +1\}$ which is uniform and independent of anything else. Conditionally on Z_0, ϕ and the labels $\{Z_i\}_{i \in \mathbb{N}}$, we place an edge between node $i \in \mathbb{N}$ and this extra node at the origin with probability $f_{in}(||X_i||)\mathbf{1}_{Z_i = Z_0} + f_{out}(r)\mathbf{1}_{Z_i = -Z_0}$ independent of the edges between $j \neq i \in \mathbb{N}$ and the origin.

We now set some notation to define 'information given at infinity'. For every $r \in \mathbb{R}_+$, denote by $\phi^{(r)}$ and $G^{(r)}$ the point-process and graph, in which every vertex $i \in \mathbb{N}$ (which is at location $X_i \in \mathbb{R}^d$) is equipped with the random variable $Z_i \mathbf{1}_{||X_i||_2 \geq r}$. Note that this is not a mark since it is not translation invariant, but is a random variable associated with vertex i. In words, we retain the community label marks on nodes of G at a Euclidean distance of r or more from the origin and delete (i.e. set to 0) the community label of those nodes which are located at distances less than r from the origin.

Definition 34. *We say* Information Flows from Infinity *if for every $r \in \mathbb{R}_+$ there exists a random variable $\tau_r \in \{-1, +1\}$, measurable (deterministic function) with respect to the observed data*

$(\phi^{(r)}, G^{(r)})$ *and a constant* $\gamma > 0$ *such that*

$$\liminf_{r\to\infty} \mathbb{P}^0[\tau_r = Z_0] \geq \frac{1}{2} + \gamma. \tag{26}$$

We say information 'flows' from infinity if we are able to non-trivially estimate the community label at origin, given 'information at infinity'. Note that for each r, there exists algorithms (i.e. τ_r) such that $\mathbb{P}^0[\tau_r = Z_0] > \frac{1}{2}$. However, the non-trivial question is to understand if the limit as $r \to \infty$ is still strictly larger than a half. This definition is similar in spirit to those considered in Ising models to detect phase-transition for multiplicity of Gibbs states (as in [39] and [10]). Nonetheless, the problem as stated in the continuum space context is new to the best of our knowledge and its study could be of independent interest. We first establish a monotonicity property of this problem and connect it with the Community Detection Problem.

Proposition 35. *For every* $d \in \mathbb{N}$ *and* $f_{in}(\cdot), f_{out}(\cdot) : \mathbb{R}_+ \to [0,1]$*, the limit* $\lim_{r\to\infty} \sup_{\tau_r \in \sigma((\bar{G}^{(r)}, \bar{\phi}^{(r)}))} \mathbb{P}^0[\tau_r = Z_0]$ *exists. Moreover,* $\lambda \to \lim_{r\to\infty} \sup_{\tau_r \in \sigma((\bar{G}^{(r)}, \bar{\phi}^{(r)}))} \mathbb{P}^0[\tau_r = Z_0]$ *is non-decreasing.*

Note the supremum is over all possible estimators of the community label at origin.

Proof. Denote by $\tilde{\xi}(\lambda, r) := \sup_{\tau_r \in \sigma((\bar{G}^{(r)}, \bar{\phi}^{(r)}))} \mathbb{P}^0_\phi[\tau_r = Z_0]$. Notice that, for each fixed λ and $r^{'} \geq r$, we have $\sigma((\bar{G}^{(r^{'})}, \bar{\phi}^{(r^{'})})) \subseteq \sigma((\bar{G}^{(r)}, \bar{\phi}^{(r)}))$. This follows from the fact that sample path-wise,$(\bar{G}^{(r^{'})}, \bar{\phi}^{(r^{'})})$ is a measurable function of $(\bar{G}^{(r)}, \bar{\phi}^{(r)})$ which is obtained by zeroing all revealed labels in the set $B_r^{\complement} \cap B_{r'}$. Hence, the limit in proposition 35 exists.

It remains to prove that $\xi(\lambda) := \lim_{r\to\infty} \tilde{\xi}(\lambda, r)$ is non-decreasing in λ. It suffices to prove that $\tilde{\xi}(\lambda, r)$ is non-decreasing in λ for every r. We show this by using a standard coupling argument used to prove monotonicity of percolation probabilities (for example in Chapter 2, [36]). The basis of the coupling argument is the independent thinning property and Slivnyak's theorem of the PPP and the associated random connection model. These two theorems gives the following two facts. Let (ϕ, G) be a Poisson Point Process of intensity λ and G is the block model graph for some connection functions $f_{in}(\cdot)$ and $f_{out}(\cdot)$ under measure $\mathbb{P}$. Then if each node of G along with its incident edges are removed independently with probability p, the resulting point process $\phi^{'}$ is an instance of a PPP with intensity λp and the resulting graph $G^{'}$ is the associated block model graph with the same connection functions $f_{in}(\cdot)$ and $f_{out}(\cdot)$. Slivnyak's theorem for (ϕ, G) gives that if we place an extra node at origin and equip it with independent community label and edges, the resulting point-process and graph is equal in distribution to (ϕ, G) under the Palm measure $\mathbb{P}^0$.

Thus given a problem instance at intensity λ under measure $\mathbb{P}^0$, we can independently remove nodes of G other than the one at origin with probability p. The resulting graph and the Information Flow from Infinity problem will be that at intensity λp. Thus, the best performance at intensity λ cannot be smaller than that at intensity λp. Since p was arbitrary, we have that the best performance at intensity λ cannot be smaller than that at any intensity $\lambda^{'} \leq \lambda$. In other words, for all $r \geq 0$, $\tilde{\xi}(\lambda^{'}, r) \leq \tilde{\xi}(\lambda, r)$. □

We will need the following classical result on the ergodic property of marks of a stationary point process.

Proposition 36. *([37]) Let $\phi := \{X_1, X_2, \cdots\}$ be a homogeneous PPP with its atoms enumerated in an arbitrary measurable way. Let each atom $i \in \mathbb{N}$ be assigned a translation invariant mark random variable $J_i \in \Xi$ taking values in an arbitrary Borel measurable space $(\Xi, \mathfrak{g})$. Let $B_n := \left[-\frac{n^{1/d}}{2}, \frac{n^{1/d}}{2}\right]^d$ be the box of volume n and let $X_{j^{(n)}} \in \phi$ be chosen uniformly at random among the atoms of ϕ that lie in B_n, if any. Then for all $A \in \mathfrak{g}$, the limit $\lim_{n\to\infty} \mathbb{P}[J_{j^{(n)}} \in A]$ exists and satisfies $\lim_{n\to\infty} \mathbb{P}[J_{j^{(n)}} \in A] = \mathbb{P}^0[J_0 \in A]$, where J_0 is the mark of the atom of ϕ at origin under $\mathbb{P}^0$.*

The following proposition establishes that Community Detection is harder than Information Flow from Infinity.

Lemma 37. *If there exists a Community Detection algorithm (polynomial or exponential time) that achieved an overlap of $\gamma > 0$, then $\lim_{r\to\infty} \sup_{\tau_r \in \sigma((\bar{G}^{(r)}, \bar{\phi}^{(r)}))} \mathbb{P}^0[\tau_r = Z_0] \geq \frac{1+\gamma}{2}$.*

Proof. We will assume that we cannot solve Information Flow from Infinity problem and then conclude that Community Detection is not solvable. More precisely, we will assume that $\lim_{r\to\infty} \sup_{\tau_r \in \sigma((\bar{G}^{(r)}, \bar{\phi}^{(r)}))} \mathbb{P}^0[\tau_r = Z_0] \leq \frac{1}{2}$ and then argue that no Community Detection algorithm can achieve a positive overlap. A Community Detection algorithm achieves an overlap $\gamma > 0$ if when run on the data (G_n, ϕ_n), it produces an output $\{\tau_i^{(n)}\}_{i=1}^{N_n}$ satisfying

$$\frac{|\sum_{i=1}^{N_n} \tau_i^{(n)} Z_i|}{N_n} \geq \gamma, \tag{27}$$

with probability $1 - o_n(1)$. Now, an *easier* question corresponds to asking if any two uniformly randomly chosen nodes (with replacement) of G_n belong to the same or opposite community. This question is easier than Community Detection since one way to answer this pairwise question is to first produce a partition of all nodes of G_n and then answer the question for the two randomly chosen nodes. Note that an overlap of γ can be achieved if and only if a fraction $(1 + \gamma)/2$ of the nodes have been correctly classified. Hence the chance that any two uniformly chosen nodes are classified correctly is at-least $(1 + \gamma)/2$. Since we can achieve an overlap of γ with probability $1 - o_n(1)$, the chance that two uniformly randomly chosen nodes of G_n to be correctly classified is at-least $(1 + \gamma)/2 - o_n(1)$. Hence, if we show that the best estimator for answering whether any two randomly chosen nodes from G_n belong to the same or opposite community has a success probability of at-most $\frac{1}{2} + o_n(1)$, then no algorithm exists for solving Community Detection. In the rest of the proof, we will show that if the Information Flow from Infinity cannot be solved, then for every $\epsilon > 0$, the best estimator to estimate whether any two randomly chosen nodes of G_n belong to the same or opposite communities will succeed with probability at-most $\frac{1}{2} + \epsilon + o_n(1)$. This will conclude the proof that no algorithm exists for solving Community Detection in view of the preceding discussion and hence the proof of Lemma 37.

Let $\epsilon > 0$ be arbitrary. Under the assumption that Information Flow from Infinity cannot be solved, there exists a $r > 0$ such that $\sup_{\tau_r \in \sigma((G^{(r)}, \phi^{(r)}))} \mathbb{P}^0[\tau_r = Z_0] \leq \frac{1}{2} + \frac{\epsilon}{2}$. In words, choose a r such that the Information Flow from Infinity cannot succeed with probability larger than $\frac{1}{2} + \frac{\epsilon}{2}$. Now, let n be large enough such that for two uniformly randomly chosen nodes of G_n denoted by i and j to be $||X_i - X_j|| > r$ with probability at-least $1 - \frac{\epsilon}{2}$. Now, assume that we are on the event that $||X_i - X_j|| > r$. Conditionally on this event, the probability that any pairwise estimator correctly tells whether the two nodes i and j are in the same or opposite community will succeed

with probability at-most $\frac{1}{2}+\frac{\epsilon}{2}$. This follows since conditionally on X_i and X_j, we can make the pairwise problem *easier* by revealing all community labels of nodes at a distance of larger than r from X_i and asking whether we can now guess the community label at X_i. This will enable us to answer the pairwise question of whether X_i and X_j lie in the same community or not since we will know the true label of X_j when the labels of nodes at distances r or more from X_i are revealed. This now is a problem of finding a mark τ_i of the atom i of ϕ which denotes the best community label estimate of X_i given ϕ, G and all community labels of nodes at a distance of r or more from X_i. Since X_i was an uniformly randomly chosen point from $\phi \cap B_n$, the chance that $\tau_i = Z_i$ is equal to the Palm probability that the best community label estimate of the node at origin is correct given ϕ, G and the true community labels of all nodes at a distance r or more from the origin. This follows from a direct application of Proposition 36. Thus the probability $\tau_i = Z_i$ is bounded from above by $\frac{1}{2}+\frac{\epsilon}{2}+o_n(1)$. On the complementary event that $||X_i - X_j|| < r$, we use the trivial bound that the pairwise estimation is always successful. Hence by the law of total probability, the success probability of the pairwise estimator cannot be larger than $\frac{1}{2}+\epsilon+o_n(1)$. In other words, for every $\epsilon > 0$, there exists a $n_\epsilon < \infty$, such that for all $n \geq n_\epsilon$, the probability that we correctly identify the community membership of two uniformly randomly chosen nodes of G_n is at-most $\frac{1}{2}+\epsilon$.

□

The following is the main technical result on the Information Flow from Infinity problem.

6.3 Main Result on Information Flow from Infinity

Theorem 38. *For every $\lambda, f_{in}(\cdot), f_{out}(\cdot)$ and d, the following limit exists and satisfies*

$$\lim_{r\to\infty} \sup_{\tau_r \in \sigma(G^{(r)}, \phi^{(r)})} \mathbb{P}^0[\tau_r = Z_0] \leq \frac{1}{2}\left(1 + \theta(H_{\lambda, f_{in}(\cdot)-f_{out}(\cdot), d})\right). \tag{28}$$

Recall that $\theta(H_{\lambda,f_{in}(\cdot)-f_{out}(\cdot),d})$ is the percolation probability of the classical random connection model where any two nodes of ϕ located at $x, y \in \mathbb{R}^d$ are connected by an edge with probability $f_{in}(||x-y||) - f_{out}(||x-y||)$. The supremeum is over all valid estimators of the community label at origin and hence if $\theta(H_{\lambda,f_{in}(\cdot)-f_{out}(\cdot),d}) = 0$, then there is no estimator that will solve the Information Flow from Infinity problem. In view of Lemma 37, we also get that if $\theta(H_{\lambda,f_{in}(\cdot)-f_{out}(\cdot),d}) = 0$, then there is no algorithm (polynomial or exponential time) to solve Community Detection. Thus, if we prove Theorem 38, then we will conclude the proof of Theorem 4.

Before presenting the proof of Theorem 38, we illustrate a few example setting where the bound in Equation 28 is tight and loose respectively. In view of Lemma 37 and Proposition 7, the following corollary where Equation (28) is tight holds.

Corollary 39. *For all $\lambda > 0, R_1 \geq R_2$, if $f_{in}(r) = \mathbf{1}_{r\leq R_1}$ and $f_{out}(r) = \mathbf{1}_{r\leq R_2}$*

$$\lim_{r\to\infty} \sup_{\tau_r \in \sigma(G^{(r)}, \phi^{(r)})} \mathbb{P}^0[\tau_r = Z_0] = \frac{1}{2}\left(1 + \theta(H_{\lambda, f_{in}(\cdot)-f_{out}(\cdot), d})\right).$$

In other words, we see that the inequality in Theorem 38 is achieved in certain examples. However, Theorem 38 is not an accurate characterization of the Information Flow from Infinity problem as evidenced in the following example.

Proposition 40. *For all $d \geq 2$, if $f_{in}(r) = \min\left(1, \frac{1}{\sqrt{r}} + \frac{1}{r^{d-1/4}}\right)$ and $f_{out}(r) = \min\left(1, \frac{1}{\sqrt{r}}\right)$, the inequality in Equation (28) is strict for all values of $\lambda > 0$.*

The example in Proposition 40 corresponds to the case when the degree of each node is almost-surely infinite. Thus, $\theta(H_{\lambda, f_{in}(\cdot) - f_{out}(\cdot), d}) = 1$ in this case. However, using results from [40], one can argue that perfect recovery is impossible in this example, i.e. $\lim_{r\to\infty} \sup_{\tau_r \in \sigma(G^{(r)}, \phi^{(r)})} \mathbb{P}^0[\tau_r = Z_0] < 1$. The key tool, is to see that if perfect recovery were to be possible, then it would be the case that either of the following two pairs of point-process will be mutually singular.

1. The point process formed by the location of those nodes of G that have an edge to the origin and have a community label Z_0 and the point process formed by the location of those nodes of G having an edge to the origin and having a community label of $-Z_0$ are mutually singular.

2. Or, the point process corresponding to the locations of those nodes of G that have a community label Z_0 and do not have an edge to the origin and the point process corresponding to the locations of those nodes of G that have a community label $-Z_0$ and do not have an edge to the origin are mutually singular.

We will argue that in our example, neither is possible by alluding to a theorem from [40], and hence perfect recovery is not possible. We present the complete proof in the Appendix A A.

6.4 The Information Graph and Proof of Theorem 38

In this section, we generalize the example of the previous section and give a proof of Theorem 38. To do so, we define a general information graph and conclude that if this constructed information graph does not percolate, then one cannot solve the Information Flow from Infinity problem.

We denote by I the information graph whose vertex set is ϕ. The random graph I is constructed just based on the positions of the points and the random elements $\{\{U_{ij}\}_{j>i}\}_{i\in\mathbb{N}}$. Recall that the graph G was built by connecting any two points $i < j \in \mathbb{N}$ if $U_{ij} \leq \mathbf{1}_{Z_i = Z_j} f_{in}(||X_i - X_j||) + \mathbf{1}_{Z_i \neq Z_j} f_{out}(||X_i - X_j||)$. Using the same random elements, we connect any $i < j \in \mathbb{N}$ by an edge in graph I if $U_{ij} \in [f_{out}(||X_i - X_j||), f_{in}(||X_i - X_j||)]$. We denote by $i \sim_I j$ the event that points i and j are connected by an edge in I. Hence the graphs I and G are coupled and built on the same probability space using the same set of random elements. For each $i \in \mathbb{N}$, we denote by $V_I(i) \subseteq \mathbb{N}$ the random subset of the nodes contained in the connected component of node i in graph I. Note that the information graph $I \overset{(d)}{=} H_{\lambda, f_{in}(\cdot) - f_{out}(\cdot), d}$, i.e. the I graph we constructed is equal in distribution to the graph of the Poisson Random Connection model with vertex set forming a PPP of intensity λ and connecting any two vertices at distance r away with probability $f_{in}(r) - f_{out}(r)$ independently of everything else. This equality in distribution follows from the fact that $\{\{U_{kl}\}_{l>k}\}_{k\in\mathbb{N}}$ is an i.i.d. uniform $[0, 1]$ sequence. The following structural lemma justifies the term information graph.

Lemma 41. *From the way we have coupled the construction of G and I, we have*

- *If $i \sim_I j$ and $i \sim_G j$, then $Z_i = Z_j$.*
- *If $i \sim_I j$ and $i \nsim_G j$, then $Z_i \neq Z_j$.*

Proof. This follows from the following construction of G and I as follows.

- $i \sim_G j$ if and only if $U_{ij} \leq f_{in}(||X_i - X_j||)\mathbf{1}_{Z_i=Z_j} + f_{out}(||X_i - X_j||)\mathbf{1}_{Z_i \neq Z_j}$.
- $i \sim_I j$ if and only if $U_{ij} \in [f_{out}(||X_i - X_j||), f_{in}(||X_i - X_j||)]$.
- $\forall r \geq 0$, $f_{in}(r) \geq f_{out}(r)$.

The lemma follows since the $\{U_{ij}\}_{0 \leq i < j}$ are the same with which we build both the random graphs G and I. □

We can iterate the above lemma from edges to connected components of I which forms a crucial structural lemma.

Lemma 42. *For all $j \in \mathbb{N}$, conditional on G, ϕ, I, there are exactly two possible sequences $(Z_k)_{k \in V_I(j)}$ which are complements of each other that are consistent in the sense of Lemma 41 with the observed data G, ϕ and I.*

Proof. Denote by $T_I(j)$ be the breadth first spanning tree of I constructed with j as the root. Thus in the tree $T_I(j)$, and for all $k \in V_I(j)$, there is exactly one path from j to k in the tree $T_I(j)$.

The existence of two labelings is not so difficult since we know there exists one underlying true labeling which generated the data G, I. But since, the model is symmetric, the complement of the true labels will also be consistent in the sense of Lemma 41. Thus, there are at least two labeling consistent with the observed data G, I. These two labels of $V_I(j)$ can be constructed explicitly which we do in the next paragraph. We then show, that there are no other that can be consistent in the sense of Lemma 41, which will conclude the proof.

To construct the two possible labelings, first assume that $Z_j = +1$. Now conditionally on this and G, each neighbor of j in $T_I(j)$ will have exactly one possible community label estimate that is consistent in the sense of Lemma 41. Now, by induction, we can construct the labels of $V_I(j)$. Assume, that conditionally on $Z_j = +1$ and G, we have a unique set of labels for all vertices in $T_I(j)$ at graph distance of less than or equal to k. Let $u \in T_I(j)$ be an arbitrary vertex such that it is at graph distance $k+1$ from j in $T_I(j)$. Since $T_I(j)$ is a tree, there is a unique vertex v in $V_I(j)$ such that $v \sim_{T_I(j)} u$ and v is at a distance of k from j. Thus, conditionally on $Z_j = +1$, Z_v is a fixed community label due to the induction hypothesis. Since Z_v is fixed, then there is a unique label for Z_u that will be consistent in the sense of Lemma 41. Since u was arbitrary, we can uniquely assign a community label to all vertices at graph distance of $k+1$ from j in $T_I(j)$. Hence, by induction, conditionally on $Z_j = +1$, there is a unique community estimate for all vertices in $V_I(j)$. Similarly, if we assumed $Z_j = -1$, we will find another unique labeling for the vertices in $V_i(j)$ which will be the complement of the unique labeling obtained by assuming $Z_j = +1$. This, gives us that there exist at-least two labelings of $V_i(j)$ that are complements of each other and consistent with the observed data G and I in the sense of Lemma 41.

To see that there can be no other possibilities, we argue by contradiction. Assume there are two labelings and a vertex k such that in one of the labelings $Z_j = +1, Z_k = +1$ and in the other $Z_j = +1, Z_k = -1$. It is clear that at-most one of the above labelings will be consistent in the tree $T_I(j)$ in the sense of Lemma 41. This establishes that the two sequences we constructed in the

previous paragraph which are complements of each other are the only two possible sequences that are consistent in the sense of Lemma 41.

□

The following lemma, which follows from simple observations, essentially says the community labels on disconnected components of I are independent.

Lemma 43. *For all $\lambda > 0$, on the event $\{|V_I(0)| < \infty\}$,*

$$\mathbb{P}^0\left[Z_0 = +1 \middle| G, \{\{U_{kl}\}_{l>k}\}_{k\in\mathbb{N}\cup\{0\}}, \phi, \{Z_i\}_{i\in V(I_{(0)})^c}\right] = \frac{1}{2} \text{ a.s.}$$

Proof. From Lemma 42, we know that conditionally on ϕ, G, I, there are exactly two possible sequences $\{Z_k\}_{k\in V_I(0)}$ that are consistent with the observed data in the sense of Lemma 41. Denote these two sequences by $\mathbf{s}$ and $\mathbf{s}^{\complement}$. It suffices to show that conditionally on $\phi, \{\{U_{kl}\}_{l>k}\}_{k\in\mathbb{N}\cup\{0\}}, G$ and $\{Z_k\}_{k\in V_I^c(0)}$, the two sequences $\mathbf{s}$ and $\mathbf{s}^{\complement}$ are equally likely. We will denote by g the realization of the random graph G. To conclude the lemma, we use Bayes' conditional rule as follows.

$$\begin{aligned}
&\mathbb{P}^0_\phi[(Z_k)_{k\in V_I(0)} = \mathbf{s}|\phi, \{\{U_{kl}\}_{l>k}\}_{k\in\mathbb{N}\cup\{0\}}, (Z_k)_{k\in V_I^c(0)}, G = g] \\
&= \frac{\mathbb{P}^0_\phi[G = g|\phi, \{\{U_{kl}\}_{l>k}\}_{k\in\mathbb{N}\cup\{0\}}, (Z_k)_{k\in V_I^c(0)}, (Z_k)_{k\in V_I(j)} = \mathbf{s}]}{\mathbb{P}^0_\phi[G = g|\phi, \{\{U_{kl}\}_{l>k}\}_{k\in\mathbb{N}\cup\{0\}}, (Z_k)_{k\in V_I^c(0)}]} \qquad (29) \\
&\mathbb{P}^0_\phi[(Z_k)_{k\in V_I(0)} = \mathbf{s}|\phi, \{\{U_{kl}\}_{l>k}\}_{k\in\mathbb{N}\cup\{0\}}, (Z_k)_{k\in V_I^c(0)}] \\
&\overset{(a)}{=} \frac{1}{\sum_g \mathbb{P}^0_\phi[G = g|\phi, \{\{U_{kl}\}_{l>k}\}_{k\in\mathbb{N}\cup\{0\}}, (Z_k)_{k\in V_I^c(0)}]} \left(\frac{1}{2}\right)^{|V_I(0)|} \text{ a.s. on the event } \{|V_I(0)| < \infty\} \qquad (30) \\
&\overset{(b)}{=} \frac{1}{2} \text{ a.s. on the event } \{|V_I(0)| < \infty\} \qquad (31)
\end{aligned}$$

The first equality follows from rewriting the events using Baye's conditional rule. In the rest of the proof, we justify steps (a) and (b). We prove the equalities and also justify that one can apply conditional Baye's rule without worrying about the 0 by 0 situation almost-surely.

Note that conditionally on $\phi, \{\{U_{kl}\}_{l>k}\}_{k\in\mathbb{N}\cup\{0\}}, \{Z_k\}_{k\in\mathbb{N}}$, the graph G is fixed and deterministic. Thus, the numerator in step (a) is 1 almost-surely. This follows from Lemma 42 which states that g is consistent with the data $(\phi, \{\{U_{kl}\}_{l>k}\}_{k\in\mathbb{N}\cup\{0\}}, (Z_k)_{k\in V_I^{\complement}(0)})$ if $(Z_k)_{k\in V_I(0)} = \mathbf{s}$ or $\mathbf{s}^{\complement}$. Furthermore, the process $(Z_k)_{k\in\mathbb{N}}$ is an i.i.d. sequence independent of everything else. Hence, given any random finite subset $A \in \mathbb{N}$ independent of $(Z_k)_{k\in\mathbb{N}}$, the labels $(Z_k)_{k\in A}$ are uniform over $\{-1, 1\}^{|A|}$. Now, since $|V_I(0)| < \infty$, and $V_I(0)$ is a function of $(\phi, \{\{U_{kl}\}_{l>k}\}_{k\in\mathbb{N}\cup\{0\}})$ which is independent of $(Z_k)_{k\in\mathbb{N}}$, it follows that, on the event $\{|V_I(0)| < \infty\}$,

$$\mathbb{P}^0_\phi[(Z_k)_{k\in V_I(0)} = \mathbf{s}|\phi, \{\{U_{kl}\}_{l>k}\}_{k\in\mathbb{N}\cup\{0\}}, (Z_k)_{k\in V_I^c(0)}] = \left(\frac{1}{2}\right)^{|V_I(0)|} \text{ a.s.} \qquad (32)$$

Moreover, the above expression is non-zero almost surely since $|V_I(0)| < \infty$. This justifies step (a). To conclude the proof, it suffices to show that on the event $\{|V_I(0)| < \infty\}$,

$$\sum_g \mathbb{P}^0_\phi[G = g | \phi, \{\{U_{kl}\}_{l>k}\}_{k\in\mathbb{N}\cup\{0\}}, (Z_k)_{k\in V_I^c(0)}] = 2\left(\frac{1}{2}\right)^{|V_I(0)|} \text{ a.s .} \tag{33}$$

This will conclude the proof by noticing that the above expression is non-zero almost-surely.

Observe that the summation in Equation (33) is over the various community labels $(Z_k)_{k\in V_I(0)}$. Thus, the summation is over the $2^{|V_I(0)|}$ different choices for $(Z_k)_{k\in V_I(0)}$. However, given ϕ and $\{\{U_{kl}\}_{l>k}\}_{k\in\mathbb{N}\cup\{0\}}$, one can construct the I graph. Then, Lemma 42 states that the total number of possible choices for the labels $(Z_k)_{k\in V_I(0)}$ is now only two, which we denoted by $\mathbf{s}$ and $\mathbf{s}^{\complement}$ in this proof. However, again from Lemma 42, conditionally on those two sequences, the graph constructed from the data $(\phi, \{\{U_{kl}\}_{l>k}\}_{k\in\mathbb{N}\cup\{0\}}, (Z_k)_{k\in V_I^c(0)}, (Z_k)_{k\in V_I(0)} = \mathbf{s})$ and from the data $(\phi, \{\{U_{kl}\}_{l>k}\}_{k\in\mathbb{N}\cup\{0\}}, (Z_k)_{k\in V_I^c(0)}, (Z_k)_{k\in V_I(0)} = \mathbf{s}^{\complement})$ is g, the observed graph. Hence, the proof of the claim follows from Equation (32).

□

The following is an immediate corollary of the definition of conditional expectation.

Corollary 44. *For all events $A \in \sigma(G, \phi, \{\{U_{kl}\}_{l>k}\}_{k\in\mathbb{N}\cup\{0\}}, (Z_k)_{k\in V_I^{\complement}(0)})$, we have $\mathbb{E}^0[\mathbf{1}_E\mathbf{1}_A\mathbf{1}_{Z_0=+1}] = \mathbb{E}^0[\mathbf{1}_E\mathbf{1}_A\mathbf{1}_{Z_0=-1}] = \frac{1}{2}\mathbb{E}^0[\mathbf{1}_E\mathbf{1}_A]$, where E is the event that $V_I(0)$ is finite.*

6.5 Proof of Theorem 38

We are now ready to conclude the proof of Theorem 38. Notice that since $\tau_r \in \{-1, +1\}$, we can represent it as $\tau_r = \mathbf{1}_A - \mathbf{1}_{A^{\complement}}$, for some $A \in \sigma((G^{(r)}, \phi^{(r)}))$. Hence, we have

$$\sup_{\tau_r\in\sigma((G^{(r)},\phi^{(r)}))} \mathbb{P}^0_\phi[\tau_n^{(\delta)} = Z_0] = \sup_{A\in\sigma((G^{(r)},\phi^{(r)}))} \mathbb{E}^0_\phi[\mathbf{1}_A\mathbf{1}_{Z_0=+1} + \mathbf{1}_{A^{\complement}}\mathbf{1}_{Z_0=-1}]. \tag{34}$$

For every $m \in \mathbb{N}$, denote by E_m the event that $C_I(0) \subseteq B_m$, i.e. the event that the connected component of the point at the origin in I is contained in the set B_m. The sets E_m are non-decreasing. Moreover, from the definition of percolation, $\mathbb{P}^0[\cup_{m\in\mathbb{N}} E_m] = \lim_{m\to\infty} \mathbb{P}^0[E_m] = 1 - \theta(H_{\lambda, f_{in}(\cdot)-f_{out}(\cdot),d})$. Let $r \in \mathbb{R}$ be arbitrary, and condition on the event E_r. We have,

$$
\begin{aligned}
&\sup_{A\in\sigma((G^{(r)},\phi^{(r)}))} \mathbb{E}^0[\mathbf{1}_A\mathbf{1}_{Z_0=+1}+\mathbf{1}_{A^{\complement}}\mathbf{1}_{Z_0=-1}] = \\
&\sup_{A\in\sigma(((G^{(r)},\phi^{(r)}))} \mathbb{E}^0[\mathbf{1}_{E_r}\left(\mathbf{1}_A\mathbf{1}_{Z_0=+1}+\mathbf{1}_{A^{\complement}}\mathbf{1}_{Z_0=-1}\right)] + \mathbb{E}^0[\mathbf{1}_{E_r^{\complement}}\left(\mathbf{1}_A\mathbf{1}_{Z_0=+1}+\mathbf{1}_{A^{\complement}}\mathbf{1}_{Z_0=-1}\right)] \\
&\qquad \leq \sup_{A\in\sigma(((G^{(r)},\phi^{(r)})))} \mathbb{E}^0[\mathbf{1}_{E_r}\left(\mathbf{1}_A\mathbf{1}_{Z_0=+1}+\mathbf{1}_{A^{\complement}}\mathbf{1}_{Z_0=-1}\right)] + \mathbb{P}^0[E_r^{\complement}] \\
&\qquad \overset{(a)}{\leq} \sup_{A\in\sigma((G^{(r)},\phi^{(r)},\{\{U_{kl}\}_{l>k}\}_{k\in\mathbb{N}\cup\{0\}})} \mathbb{E}^0[\mathbf{1}_{E_r}\left(\mathbf{1}_A\mathbf{1}_{Z_0=+1}+\mathbf{1}_{A^{\complement}}\mathbf{1}_{Z_0=-1}\right)] + \mathbb{P}^0[E_r^{\complement}] \\
&\qquad \overset{(b)}{\leq} \sup_{A\in\sigma(G,\phi,\{\{U_{kl}\}_{l>k}\}_{k\in\mathbb{N}\cup\{0\}},(Z_k)_{k\in V_I^{\complement}(0)})} \mathbb{E}^0[\mathbf{1}_{E_r}\left(\mathbf{1}_A\mathbf{1}_{Z_0=+1}+\mathbf{1}_{A^{\complement}}\mathbf{1}_{Z_0=-1}\right)] + \mathbb{P}^0[E_r^{\complement}] \\
&\qquad \overset{(c)}{=} \sup_{A\in\sigma(G,\phi,\{\{U_{kl}\}_{l>k}\}_{k\in\mathbb{N}\cup\{0\}},(Z_k)_{k\in V_I^{\complement}(0)})} \frac{1}{2}\mathbb{E}^0[\mathbf{1}_A\mathbf{1}_{E_r}] + \frac{1}{2}\mathbb{E}^0[\mathbf{1}_{A^{\complement}}\mathbf{1}_{E_r}] + \mathbb{P}^0[E_r^{\complement}] \\
&\qquad = \frac{1}{2}\mathbb{P}^0[E_r] + \mathbb{P}^0[E_r^{\complement}].
\end{aligned}
$$

Step (a) follows from enlarging the sigma algebra over which we are searching for a solution. Step (b) follows from the fact that on the event E_r, $V_I(0) \subseteq B(0,r)$. Thus, revealing more labels will only preserve the inequality. Step (c) follows from Corollary 44. Now, since the sets E_r are non-decreasing , we get the theorem by taking a limit as r goes to infinity on both sides, i.e.

$$
\begin{aligned}
\lim_{r\to\infty} \sup_{A\in\sigma((G^{(r)},\phi^{(r)}))} \mathbb{E}^0[\mathbf{1}_A\mathbf{1}_{Z_0=+1}+\mathbf{1}_{A^{\complement}}\mathbf{1}_{Z_0=-1}] &\leq \lim_{r\to\infty} \frac{1}{2}\mathbb{P}^0[E_r] + \mathbb{P}^0[E_r^{\complement}] \\
&= \frac{1}{2}\mathbb{P}^0[E] + \mathbb{P}^0[E^{\complement}] \\
&= \frac{1}{2}(1+\theta(H_{\lambda,f_{in}(\cdot)-f_{out}(\cdot),d})). \qquad (35)
\end{aligned}
$$

The limit on the LHS exists from Proposition 3 and the limit on the RHS exists since E_r are non-decreasing events.

7 Identifiability of the Partition and Proof of Theorem 10

The key technical tool is the ergodicity of the PPP which is summarized in the following lemma. We need to set some notation that are needed to state the lemma. Denote by $\mathbb{M}_{\Xi}(\mathbb{R}^d)$ the set of all 'marked' point processes on $\mathbb{R}^d$ where each point is assigned a 'mark' from the measure space Ξ, with its associated sigma-algebra. The set $\mathbb{M}_{\Xi}(\mathbb{R}^d)$ is a Polish space, has a natural topology and hence an associated sigma-algebra (see [37]). Denote by $\theta : \mathbb{R}^d \times \mathbb{M}_{\Xi}(\mathbb{R}^d) \to \mathbb{M}_{\Xi}(\mathbb{R}^d)$ the 'shift' operator which is a measurable function where $\theta(x,\psi)$ retains the same marks but translates all points of ψ by a vector x.

Lemma 45. *Let $C_n \subset \mathbb{R}^d$ be a sequence of L_p, $p \in [1,\infty]$ balls centered at the origin with radius going to infinity as $n \to \infty$. Let $f : \mathbb{M}_{\Xi}(\mathbb{R}^d) \to \mathbb{R}_+$ be a measurable function such that $\mathbb{E}^0_{\phi}[f] < \infty$.*

Then, the following limit exists :

$$\lim_{n\to\infty} \frac{\sum_{i\in\mathbb{N}} \mathbf{1}_{X_i\in C_n} f\circ\theta(X_i,G)}{\sum_{i\in\mathbb{N}} \mathbf{1}_{X_i\in C_n}} = \mathbb{E}^0[f] \ \mathbb{P}\ a.s.\ . \tag{36}$$

We now prove Theorem 10.

Proof. First if $g(\cdot) \neq \frac{f_{in}(\cdot)+f_{out}(\cdot)}{2}$, then G and $H_{\lambda,g(\cdot),d}$ are mutually singular and this can be seen through the following elementary argument. Fix some $L<\infty$ such that $\int_{x\in\mathbb{R}^d:||x||\leq L} g(||x||)dx \neq \int_{x\in\mathbb{R}^d:||x||\leq L}((f_{in}(x)+f_{out}(||x||))/2)dx$ are both finite. Such a L exists since $g(\cdot)\neq\frac{f_{in}(\cdot)+f_{out}(\cdot)}{2}$. Now, we apply the ergodic theorem, where every node $i\in\mathbb{N}$ of ϕ is equipped with a mark $\Xi_i\in\mathbb{N}$, which denotes the number of graph neighbors of node i at a distance of at-most L from X_i, i.e. $\Xi_i = |\{j\in\mathbb{N}\setminus\{i\} i\sim_G j, ||X_i-X_j||\leq L\}|$. Thus, the Ergodic theorem implies that the measure induced by G will be concentrated on the set

$$\left\{ g\in\mathbb{M}_{\mathcal{G}}(\mathbb{R}^d) : \lim_{n\to\infty} \frac{\sum_{i,j\in\mathbb{N}} \mathbf{1}_{||X_i||\leq n,||X_i-X_j||\leq L}\mathbf{1}_{i\sim_g j}}{\sum_{i,j\in\mathbb{N}} \mathbf{1}_{||X_i||\leq n,||X_i-X_j||\leq L}} = \int_{x\in\mathbb{R}^d:||x||\leq L} (f_{in}(x)+f_{out}(||x||))/2dx \right\},$$

while the measure induced by $H_{\lambda,g(\cdot),d}$ will be concentrated on the set

$$\left\{ g\in\mathbb{M}_{\mathcal{G}}(\mathbb{R}^d) : \lim_{n\to\infty} \frac{\sum_{i,j\in\mathbb{N}} \mathbf{1}_{||X_i||\leq n,||X_i-X_j||\leq L}\mathbf{1}_{i\sim_g j}}{\sum_{i,j\in\mathbb{N}} \mathbf{1}_{||X_i||\leq n,||X_i-X_j||\leq L}} = \int_{x\in\mathbb{R}^d:||x||\leq L} g(||x||)dx \right\}.$$

Thus, the only case to consider is the one where $g(\cdot)=(f_{in}(\cdot)+f_{out}(\cdot)/2$. From linearity of expectation, the average degree of any node $i\in\mathbb{N}$ in both graphs G and $H_{(\lambda,\frac{f_{in}(\cdot)+f_{out}(\cdot)}{2},d)}$ is the same and equal to $(\lambda/2)\int_{x\in\mathbb{R}^d}(f_{in}(||x||)+f_{out}(||x||))dx$ and thus empirical average of the degree does not help. However, we see that the triangle profiles differ in the two models which we leverage to prove the Theorem.

For ease of notation, we denote by $H := H_{\lambda,(f_{in}(\cdot)+f_{out}(\cdot))/2,d}$. Define

$$\Delta = \mathbb{E}^0\left[\sum_{x\neq y\neq 0} h(x,y)\mathbf{1}((0,x)\in E,(0,y)\in E,(x,y)\in E)\right].$$

Denote by Δ_G and Δ_H the value of the above expression if the underlying graphs were G and H respectively. From the moment measure expansion of PPPs [38], we get that

$$\Delta_G = \int_{x\in\mathbb{R}^d}\int_{y\in\mathbb{R}^d} h(x,y)(f_{in}(||x-y||)\left(\frac{f_{in}(||x||)f_{in}(||y||)+f_{out}(||x||)f_{out}(||y||)}{4}\right)+ f_{out}(||x-y||)\left(\frac{f_{in}(||x||)f_{out}(||y||)+f_{out}(||x||)f_{in}(||y||)}{4}\right))\lambda^2 dxdy, \tag{37}$$

and

$$\Delta_H = \int_{x\in\mathbb{R}^d}\int_{y\in\mathbb{R}^d} h(x,y)\left(\frac{f_{in}(||x||)+f_{out}(||x||)}{2}\right)\left(\frac{f_{in}(||y||)+f_{out}(||y||)}{2}\right) \left(\frac{f_{in}(||x-y||)+f_{out}(||x-y||)}{2}\right)\lambda^2 dxdy. \quad (38)$$

Now observe that

$$\Delta_G - \Delta_H = \int_{x\in\mathbb{R}^d}\int_{y\in\mathbb{R}^d} h(x,y)\frac{1}{2}\left(f_{in}(||x||)-f_{out}(||x||)\right)\left(f_{in}(||y||)-f_{out}(||y||)\right) \left(f_{in}(||x-y||)-f_{out}(||x-y||)\right)\lambda^2 dxdy. \quad (39)$$

From the fact that $f_{in}(r) \geq f_{out}(r)$ and $f_{in}(r)$ is not equal to $f_{out}(r)$ Lebesgue almost everywhere, there exists a positive bounded function $h(x,y)$ such that $0 \leq \Delta_H < \Delta_G < \infty$. Choose one such test function $h(\cdot,\cdot)$, for ex. $h(x,y) = \mathbf{1}_{||x||\leq R}\mathbf{1}_{||y||\leq R}\mathbf{1}_{f_{in}(||x||)>f_{out}(||x||)}\mathbf{1}_{f_{in}(||y||)>f_{out}(||y||)}\mathbf{1}_{f_{in}(||x-y||)>f_{out}(||x-y||)}$ and consider the following estimator:

Algorithm 4 Detect-Partitions

Given the data, i.e. locations of nodes and the graph, pick a L large enough and compute

$$\Delta^{(L)} := \frac{\sum_{i\in\mathbb{N}}\mathbf{1}(|X_i|\leq L)\tilde{h}(X_i)}{\sum_{i\in\mathbb{N}}\mathbf{1}(|X_i|\leq L)},$$

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

From ergodicity, we know that $\lim_{L\to\infty}\Delta^{(L)} = \Delta_G$, $\mathbb{P}$ almost-surely if the data is the block model graph or $\lim_{L\to\infty}\Delta^{(L)} = \Delta_H$ $\mathbb{P}$ almost surely if the graph is drawn according to the null model. We can apply the ergodic theorem since a spatial random graph is a marked point process (as described in Section 2.1). Thus, the measures induced by (ϕ, G) and $(\phi, H_{\lambda,\frac{f_{in}(\cdot)+f_{out}(\cdot)}{2},d})$ are mutually singular. Moreover, the above algorithm when tested on the finite data (G_n, ϕ_n) runs in time proportional to λn with the multiplicative constants depending on the connection functions $f_{in}(\cdot)$ and $f_{out}(\cdot)$ with success probability $1 - o_n(1)$ with the $o_n(1)$ term depending on the function $h(\cdot,\cdot)$ chosen.

□

We investigated the singularity of measures in order to understand the question of distinguishability of the planted partition model. This is a hypothesis testing problem of whether the data (graph and spatial locations) is drawn from the distribution of G or from the distribution of $H_{\lambda,g(\cdot),d}$ with a probability of success exceeding a half given a uniform prior over the models. This problem is in some sense easier than Community Detection, since this asks for asserting whether a partition exists or not, which intuitively should be simpler than finding the partition. Indeed, we show this in our model by proving that the distinguishability problem is trivially solvable while community detection undergoes a phase transition and is solvable only under certain regimes. In the general sparse SBM however, the equivalence between distinguishability and community detection is only conjectured and not yet proven ([15],[16]).

8 Conclusions and Open Problems

In this paper, we introduced the problem of community detection in a spatial random graph where there are two equal sized communities. Our main technical contributions are in identifying the problem of Information Flow from Infinity and connecting that with the Community Detection problem and giving a simple lower bound criterion. For developing the algorithm, we noticed that a spatial graph is sparse due to the fact that all interactions are dense, but localized which is starkly different from the reason why an Erdős-Rényi graph is sparse. We leveraged this difference to propose an algorithm for community detection by borrowing further ideas from dependent site percolation processes. However, this is just a first step and there are a plenty of open questions just concerning the model we introduced.

1) Are Community Detection and Information Flow from Infinity equivalent ? - In this paper, we proved that Community Detection was harder than Information Flow from Infinity. However, our algorithm and its analysis showed that it can solve Community Detection whenever it can solve Information Flow from Infinity. Thus a natural question is whether these two problems undergo a phase-transition at the same point ? Moreover is there a relation between the optimal overlap achievable in Community Detection and the optimal success probability of estimating the community label of the origin in the Information Flow from Infinity problem ?

2) Is the optimal overlap in Community Detection Monotone ? - We saw in the proof of Proposition 35 that the optimal success probability of correctly labeling the origin in the Information Flow from Infinity problem is monotone in λ. However, for Community Detection, we only established that solvability is monotone and not the optimal overlap achievable.

3) More than 2 *Communities* - In this paper, we focused exclusively on the case of two communities in the network, and an immediate question is that of 3 or more communities. In the symmetric case where the connection function is $f_{in}(\cdot)$ within communities and $f_{out}(\cdot)$ across communities, a simple adaptation of our algorithm can give a sufficient condition, although will be sub-optimal. Our lower bound technique can also be applied in the setting of many communities (see Theorem 2 in [34] for example) thereby establishing the *existence* of a non-trivial phase transition for any number of communities in the symmetric setting. But the open question is to identify examples similar to Proposition 7 where the phase transition can be tight. A quest for such examples can possibly lead to better understanding of even the 2 community case considered in this paper. Moreover, unified algorithmic techniques capable of handling non-symmetric case also is of interest since our algorithm does not generalize in a straight forward way to the non-symmetric setting.

4) Characterization of the Phase-Transition - An obvious but harder question is whether one can characterize if not compute the exact phase-transition for either Community Detection or Information Flow from Infinity. We show that our lower bound is capturing the phase-transition only in very specific cases and may not be tight in general due to corner cases similar to Proposition 40. We also have no reason to believe that our algorithm is optimal in any sense. Thus, a structural characterization of the phase-transition is still far from being understood.

5) Computational Phase-Transition - Another deep issue is computational complexity. Is there a regime where Community Detection is solvable, but no polynomial (in n) time and space algorithms

that operate on (G_n, ϕ_n) are known to exist ?

6) Estimating the Model Parameters - How does one efficiently estimate the connection functions $f_{in}(\cdot)$ and $f_{out}(\cdot)$ from the data of just the graph G and the spatial locations ϕ.

Acknowledgements

This work was supported by a grant of the Simons Foundations (#197982 to The University of Texas at Austin). The authors thank numerous discussions with Emmanuel Abbe for suggesting the problem and pointing out the literature. The authors also thank Gustavo de Veciana, Sanjay Shakkottai, Joe Neeman and Alex Dimakis for providing comments on an earlier version of this work [41] presented during the first author's Ph.D. candidacy talk.

References

[1] Santo Fortunato. Community detection in graphs. *Physics reports*, 486(3):75–174, 2010.

[2] Paul W Holland, Kathryn Blackmond Laskey, and Samuel Leinhardt. Stochastic blockmodels: First steps. *Social networks*, 5(2):109–137, 1983.

[3] Mark EJ Newman, Duncan J Watts, and Steven H Strogatz. Random graph models of social networks. *Proceedings of the National Academy of Sciences*, 99(suppl 1):2566–2572, 2002.

[4] Jianbo Shi and Jitendra Malik. Normalized cuts and image segmentation. *IEEE Transactions on pattern analysis and machine intelligence*, 22(8):888–905, 2000.

[5] Greg Linden, Brent Smith, and Jeremy York. Amazon. com recommendations: Item-to-item collaborative filtering. *IEEE Internet computing*, 7(1):76–80, 2003.

[6] Shaghayegh Sahebi and William W Cohen. Community-based recommendations: a solution to the cold start problem. In *Workshop on recommender systems and the social web, RSWEB*, 2011.

[7] Ravi Kumar, Prabhakar Raghavan, Sridhar Rajagopalan, and Andrew Tomkins. Trawling the web for emerging cyber-communities. *Computer networks*, 31(11):1481–1493, 1999.

[8] Andrey A Shabalin, Victor J Weigman, Charles M Perou, and Andrew B Nobel. Finding large average submatrices in high dimensional data. *The Annals of Applied Statistics*, pages 985–1012, 2009.

[9] Jingchun Chen and Bo Yuan. Detecting functional modules in the yeast protein–protein interaction network. *Bioinformatics*, 22(18):2283–2290, 2006.

[10] Elchanan Mossel, Joe Neeman, and Allan Sly. Reconstruction and estimation in the planted partition model. *Probability Theory and Related Fields*, 162(3):431–461, 2015.

[11] Emmanuel Abbe. Community detection and stochastic block models: recent developments. *arXiv preprint arXiv:1703.10146*, 2017.

[12] Robin IM Dunbar. Neocortex size as a constraint on group size in primates. *Journal of human evolution*, 22(6):469–493, 1992.

[13] Mark S Handcock, Adrian E Raftery, and Jeremy M Tantrum. Model-based clustering for social networks. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 170(2):301–354, 2007.

[14] Peter D Hoff, Adrian E Raftery, and Mark S Handcock. Latent space approaches to social network analysis. *Journal of the american Statistical association*, 97(460):1090–1098, 2002.

[15] Aurelien Decelle, Florent Krzakala, Cristopher Moore, and Lenka Zdeborová. Asymptotic analysis of the stochastic block model for modular networks and its algorithmic applications. *Physical Review E*, 84(6):066106, 2011.

[16] Jess Banks, Cristopher Moore, Joe Neeman, and Praneeth Netrapalli. Information-theoretic thresholds for community detection in sparse networks. In *Conference on Learning Theory*, pages 383–416, 2016.

[17] Béla Bollobás, Svante Janson, and Oliver Riordan. The phase transition in inhomogeneous random graphs. *Random Structures & Algorithms*, 31(1):3–122, 2007.

[18] Frank McSherry. Spectral partitioning of random graphs. In *Foundations of Computer Science, 2001. Proceedings. 42nd IEEE Symposium on*, pages 529–537. IEEE, 2001.

[19] Amin Coja-Oghlan. Graph partitioning via adaptive spectral techniques. *Combinatorics, Probability and Computing*, 19(2):227–284, 2010.

[20] Elchanan Mossel, Joe Neeman, and Allan Sly. Consistency thresholds for the planted bisection model. In *Proceedings of the Forty-Seventh Annual ACM on Symposium on Theory of Computing*, pages 69–75. ACM, 2015.

[21] Cristopher Moore. The computer science and physics of community detection: landscapes, phase transitions, and hardness. *arXiv preprint arXiv:1702.00467*, 2017.

[22] Elchanan Mossel, Joe Neeman, and Allan Sly. A proof of the block model threshold conjecture. *arXiv preprint arXiv:1311.4115*, 2013.

[23] Laurent Massoulié. Community detection thresholds and the weak ramanujan property. In *Proceedings of the 46th Annual ACM Symposium on Theory of Computing*, pages 694–703. ACM, 2014.

[24] Charles Bordenave, Marc Lelarge, and Laurent Massoulié. Non-backtracking spectrum of random graphs: community detection and non-regular ramanujan graphs. In *Foundations of Computer Science (FOCS), 2015 IEEE 56th Annual Symposium on*, pages 1347–1357. IEEE, 2015.

[25] Béla Bollobás and Oliver Riordan. *Percolation*. Cambridge University Press, 2006.

[26] Mathew D Penrose. On a continuum percolation model. *Advances in applied probability*, 23(03):536–556, 1991.

[27] Thomas M Liggett, Roberto H Schonmann, and Alan M Stacey. Domination by product measures. *The Annals of Probability*, 25(1):71–95, 1997.

[28] Mathew D Penrose. Existence and spatial limit theorems for lattice and continuum particle systems. *Prob. Surveys*, 5:1–36, 2008.

[29] Richard Durrett. *Lecture notes on particle systems and percolation*. Brooks/Cole Pub Co, 1988.

[30] Eyal Lubetzky and Allan Sly. Information percolation and cutoff for the stochastic ising model. *Journal of the American Mathematical Society*, 29(3):729–774, 2016.

[31] Eyal Lubetzky and Allan Sly. Universality of cutoff for the ising model. *arXiv preprint arXiv:1407.1761*, 2014.

[32] Eyal Lubetzky and Allan Sly. An exposition to information percolation for the ising model. *arXiv preprint arXiv:1501.00128*, 2014.

[33] Elchanan Mossel. Reconstruction on trees: beating the second eigenvalue. *Annals of Applied Probability*, pages 285–300, 2001.

[34] Emmanuel Abbe, Laurent Massoulie, Andrea Montanari, Allan Sly, and Nikhil Srivastava. Group synchronization on grids. *arXiv preprint arXiv:1706.08561*, 2017.

[35] Jiaming Xu, Laurent Massoulié, and Marc Lelarge. Edge label inference in generalized stochastic block models: from spectral theory to impossibility results. In *COLT*, pages 903–920, 2014.

[36] Ronald Meester and Rahul Roy. *Continuum percolation*, volume 119. Cambridge University Press, 1996.

[37] Daryl J Daley and David Vere-Jones. *An introduction to the theory of point processes: volume II: general theory and structure*. Springer Science & Business Media, 2007.

[38] Sung Nok Chiu, Dietrich Stoyan, Wilfrid S Kendall, and Joseph Mecke. *Stochastic geometry and its applications*. John Wiley & Sons, 2013.

[39] P. M. Bleher, J. Ruiz, and V. A. Zagrebnov. On the purity of the limiting gibbs state for the ising model on the bethe lattice. *Journal of Statistical Physics*, 79(1):473–482, 1995.

[40] AN Shiryaev. Absolute continuity and singularity of probability measures in functional spaces. In *Proceedings of the International Congress of Mathematicians, Helsinki*, pages 209–225, 1978.

[41] Abishek Sankararaman. Spatial stochastic models in wireless and data networks. `http://abishek90.github.io/QualReport.pdf`, May 2017.

A Proof of Proposition 40

The goal of this section is to establish Proposition 40 in which the equality in Theorem 38 cannot be achieved for any $\lambda > 0$. The functions $f_{in}(\cdot)$ and $f_{out}(\cdot)$ of Proposition 40 satisfy $\int_{r\geq 0}(f_{in}(r) - f_{out}(r))rdr = \infty$. Thus, the bound from Theorem 38 predicts that $\lim_{r\to\infty} \sup_{\tau_r \in \sigma(G^{(r)},\phi^{(r)})} \mathbb{P}^0_\phi[\tau_r =$

$Z_0] \leq 1$. However, we shall show that for every $\lambda > 0$, in our particular example, the probability of correctly estimating Z_0 is strictly smaller than 1. In-fact, we will show something slightly stronger. We will show that given the labels of *every node* other than the node at origin, the probability of correctly estimating Z_0 is strictly less than 1. The key tool to conclude about this example is the following classical result from [40] which states that the measures induced by two Poisson Point Processes on $\mathbb{R}^d$ are either absolutely continuous with respect to each other or are mutually singular. More precisely, the result from [40] after being adapted to our setting, states the following.

Lemma 46. *Let μ_1 and μ_2 be two measures on $\mathbb{R}^d$ such that $\mu_1 \sim \mu_2$ and $\mu_1(\mathbb{R}^d) = \mu_2(\mathbb{R}^d) = \infty$. Let P_{μ_1} and P_{μ_2} be the probability measures on the space of locally finite counting measures on $\mathbb{R}^d$ induced by two Poisson Point Processes having intensity measures μ_1 and μ_2 respectively. Then the following dichotomy holds.*

- *If $\int_{x\in\mathbb{R}^d} \left(1 - \sqrt{\frac{d\mu_1}{d\mu_2}}\right)^2 d\mu_2 < \infty$, then $P_{\mu_1} \sim P_{\mu_2}$*
- *If $\int_{x\in\mathbb{R}^d} \left(1 - \sqrt{\frac{d\mu_1}{d\mu_2}}\right)^2 d\mu_2 = \infty$, then $P_{\mu_1} \perp P_{\mu_2}$*

The following lemma will conclude that the lower bound is strictly sub-optimal in this example.

Lemma 47. *For every $\lambda > 0$ and $d \geq 2$, if $f_{in}(r) = \min\left(1, \frac{1}{r} + \frac{1}{r^{d-1/4}}\right)$ and $f_{out}(r) = \min\left(1, \frac{1}{r}\right)$, $\sup_{\tau\in\sigma(G,\phi,\{Z_i\}_{i\geq 1})} \mathbb{P}^0[\tau = Z_0] < 1$.*

Proof. From the independent thinning property of the Poisson Point Process, the origin partitions the process $\phi \setminus \{0\}$ into 4 independent Poisson Processes,

1. The point process $\phi_{+,e}$ which are the locations of those nodes of G that have an edge to the origin and have for community label Z_0. From properties of G, the intensity measure of $\phi_{+,e}$ which we denote by $\mu_{in}(\cdot)$, has a density with respect to Lebesgue measure given by $f_{in}(||\cdot||)$.
2. The point process $\phi_{-,e}$ which are the locations of those nodes of G that have an edge to the origin and have for community label $-Z_0$. The intensity measure of this point process which we denote as $\mu_{out}(\cdot)$ admits $f_{out}(||\cdot||)$ as its density with respect to Lebesgue measure.
3. The point process $\phi_{+,n}$ which are the locations of those nodes of G that do not have an edge to the origin and have for community label Z_0. The intensity measure of this point process which we denote as $\tilde{\mu}_{in}(\cdot)$ admits $1 - f_{in}(||\cdot||)$ as its density with respect to Lebesgue measure.
4. The point process $\phi_{-,n}$ which are the locations of those nodes of G that do not have an edge to the origin and have for community label $-Z_0$. The intensity measure of this point process which we denote as $\tilde{\mu}_{out}(\cdot)$ admits $1 - f_{out}(||\cdot||)$ as its density with respect to Lebesgue measure.

Since, the process of graph neighbors of the origin and graph non-neighbors of the origin are independent, it suffices to conclude that both the optimal estimators for Z_0 based on the data of just the neighbors and based on the data of just non-neighbors have a strictly positive chance of being wrong. In other words, it suffices to conclude that the measures induced on the set of locally finite counting measures on $\mathbb{R}^d$ by the process $\phi_{+,e}$ and $\phi_{-,e}$ are not mutually singular and the measures

induced by $\phi_{+,n}$ and $\phi_{-,n}$ are not mutually singular either. To do so, we will directly use Lemma 46 for our example.

Notice that the chosen example satisfies the following.

$$\begin{aligned}\int_{x\in\mathbb{R}^2}\left(1-\sqrt{\frac{f_{out}(||x||)}{f_{in}(||x||)}}\right)^2 dx<\infty\\ \int_{x\in\mathbb{R}^2}\left(1-\sqrt{\frac{1-f_{in}(||x||)}{1-f_{out}(||x||)}}\right)^2 dx<\infty\end{aligned}\tag{40}$$

Furthermore, notice that $\frac{d\mu_{out}}{d\mu_{in}}(\cdot)=\frac{f_{out}(\cdot)}{f_{in}(\cdot)}$ and $\frac{d\tilde{\mu}_{in}}{d\tilde{\mu}_{out}}(\cdot)=\frac{1-f_{in}(\cdot)}{1-f_{out}(\cdot)}$. Hence, from Equations (40) and Lemma 46, we see that for every $\lambda>0$, we have $P_{\mu_{in}}\sim P_{\mu_{out}}$ and $P_{\tilde{\mu}_{in}}\sim P_{\tilde{\mu}_{out}}$. Thus, Z_0 cannot be estimated perfectly without errors from the data and thus $\sup_{\tau\in\sigma(G,\phi,\{Z_i\}_{i\geq 1})}\mathbb{P}^0[\tau=Z_0]<1$. □

B Definition of PPP

A homogeneous PPP of intensity λ on $\mathbb{R}^d$ is a random process $\phi:=\{X_1,X_2,\cdots\}$ with each $X_i\in\mathbb{R}^d$ such that the following two holds

- For every bounded Borel set B, the cardinality of the set $\phi(B):=|\{i\in\mathbb{N}:X_i\in B\}|$ is a Poisson random variable with mean $\lambda|B|$ where $|B|$ is the volume (Lebesgue measure) of the set B.
- For any $k\in\mathbb{N}$ and any *disjoint* bounded Borel measurable sets $B_1,\cdots,B_k$, the random variables $\phi(B_1),\cdots\phi(B_k)$ are mutually independent.

C Proof of Proposition 32

To that the existence of a giant T-Good component is equivalent to asserting that a certain dependent site percolation process on $\mathbb{Z}^d$ percolates with probability $\frac{1}{2}+\eta$. We will denote by $q:=\max(1-e^{-\lambda(R/4d^{1/d})^d h(\epsilon)}+\lambda^2\left(\frac{3R}{4}\right)^d\frac{1}{d}e^{-c\lambda},0)$. From Proposition 28, the process by which cells of $\mathbb{Z}^d$ are T-Good is stochastically dominated by a dependent percolation process on $\mathbb{Z}^d$ with l_∞ dependence radius M and marginal probability at any vertex being equal to q. Thus to prove the proposition, we need an upper bound on q such that the probability of percolation is at-least $\frac{1}{2}+\eta$. Now, since the dependence is in l_∞, for any fixed q, one can construct a natural coupling of the site-percolation across dimensions. In particular, the restriction of the percolation process to a smaller dimension yields the same distribution as that of the original dependent percolation with the same parameters in a smaller dimension. Therefore, the percolation probability is non-decreasing with dimensions.

In two dimensions, the proposition is proven by using the 'self-dual' property of the planar grid $\mathbb{Z}^d$. Denote by the dual grid as $\mathbb{Z}^d+\frac{1}{2}$, the grid shifted by a half. Thus, every edge in the original grid intersects an unique edge in the dual grid and vice-versa. Let a circuit of length n be a sequence

of non repeating sites $x_0, e_1, \cdots x_n$ such that for all $i \in [0, n]$, $x_i \sim x_{i+1}$, where $x_{n+1} := x_0$. Call an edge in the dual grid closed if it intersects a closed edge in the original grid. It is easy to check that the original grid percolates if and only if there are no circuit containing the origin in the dual grid consisting of all closed edges. From a straightforward combinatorial argument, it is easy to see that there are at-most 4.3^n different closed circuits of length n and the probability that each one of them is closed is at-most $\left(\left(e^{-\lambda(R/4d^{1/d})^d h(\epsilon)} + \lambda^2 \left(\frac{3R}{4}\right)^d \frac{1}{d} e^{-c\lambda}\right)^{1/M}\right)^n$. A simple union bound now yields the proposition.