Community Detection on Euclidean Random Graphs *

Emmanuel Abbe [†]   François Baccelli, [‡]   Abishek Sankararaman[§]

Abstract

We study the problem of community detection on Euclidean random geometric graphs where each vertex has two latent variables: a binary community label and a $\mathbb{R}^d$ valued location label which forms the support of a Poisson point process of intensity λ. A random graph is then drawn with edge probabilities dependent on both the community and location labels. In contrast to the stochastic block model (SBM) that has no location labels, the resulting random graph contains many more short loops due to the geometric embedding. We consider the recovery of the community labels, partial and exact, using the random graph and the location labels. We establish phase transitions for both sparse and logarithmic degree regimes, and provide bounds on the location of the thresholds, conjectured to be tight in the case of exact recovery. We also show that the threshold of the distinguishability problem, i.e., the testing between our model and the null model without community labels exhibits no phase-transition and in particular, does not match the weak recovery threshold (in contrast to the SBM).

Keywords— Planted Partition, Stochastic Block Model, Random Connection Model, Percolation, Phase Transitions

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 ([17]) such as social sciences ([21],[38]), image segmentation [44], recommendation systems ([25],[42]), web-page sorting [22], and biology ([43], [10]) to name a few. In the present paper, we introduce a new class of spatial random graphs with communities and consider the Community Detection problem on it. Our motivation for a new class of random graph model comes from applications where nodes have geometric attributes, such as in social networks or more generally in graphs of similarities, where the similarity function has metric properties. We study two *regimes* of the random graph - the sparse degree regime and the logarithmic degree regime. The sparse degree regime is one wherein the average node degree does not scale with the total number of nodes of the graph, while the logarithmic degree regime is one

---

where the average node degree is proportional to the logarithm of the total number of nodes.

The random graph will be denoted by G_n, which has a random N_n number of nodes which is Poisson distributed with mean λn. In our formulation, $\lambda > 0$ is a fixed constant that denotes the *intensity* parameter and n is a scaling parameter, and we will consider the asymptotic as $n \to \infty$. 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 λ as the intensity parameter. To draw the edges, we consider two sequences of functions $(f_{in}^{(n)}(\cdot))_{n \in \mathbb{N}}$ and $(f_{out}^{(n)}(\cdot))_{n \in \mathbb{N}}$ such that $1 \geq f_{in}^{(n)}(r) \geq f_{out}^{(n)}(r) \geq 0$ for all $r \geq 0$ and $n \in \mathbb{N}$. Conditional on the node labels, two nodes with location labels $x, y \in B_n$ and community labels $Z_x, Z_y \in \{-1, +1\}$ are connected by an edge in G_n independently of other edges with probability $f_{in}^{(n)}(||x - y||)$ if $Z_x = Z_y$ and with probability $f_{out}^{(n)}(||x - y||)$ if $Z_x \neq Z_y$. In this interpretation, $||\cdot||$ denotes the Euclidean norm on the set B_n if G_n is sparse, or denotes the toroidal metric on B_n in the non-sparse case. We call the random graph G_n *sparse* if the connection functions $f_{in}^{(n)}(\cdot)$ and $f_{out}^{(n)}(\cdot)$ do not depend on n. More precisely, the graph G_n is sparse if for all $r \geq 0$, $f_{in}^{(n)}(r) := f_{in}(r)$ and $f_{out}^{(n)}(r) := f_{out}(r)$, for functions $f_{in}(\cdot)$ and $f_{out}(\cdot)$ satisfying $0 < \int_{x \in \mathbb{R}^d} (f_{in}(||x||) + f_{out}(||x||)) dx < \infty$. 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. In this regime, we draw an edge between two nodes i and j with probability $f_{in}(||X_i - X_j||)$ if the community labels Z_i and Z_j are the same or with probability $f_{out}(||X_i - X_j||)$ if the community labels Z_i and Z_j are different, where $||\cdot||$ denotes the Euclidean norm on the set B_n. Furthermore, the 'boundary effects' due to the edges of the set B_n will not matter asymptotically as $n \to \infty$ as the average degree is uniformly bounded in n (We make this precise in Section 2). We call this the sparse regime as the average degree is a constant independent of n. If the connection functions $f_{in}^{(n)}(\cdot)$ and $f_{out}^{(n)}(\cdot)$ depended on n and further-more satisfy $\int_{x \in \mathbb{R}^d} f_{in}^{(n)}(||x||) dx = C_{in} \log(n)$ and $\int_{x \in \mathbb{R}^d} f_{out}^{(n)}(||x||) dx = C_{out} \log(n)$ for some $C_{in} > C_{out} \geq 0$, for all n, then we call the graph G_n as the *logarithmic degree* regime or simply as the logarithmic regime. This is so since the average degree of any node is proportional to the logarithm of the total number of nodes in G_n. In this case of logarithmic regime, we avoid having to deal with boundary effects by considering the toroidal metric on the set B_n. Precisely, conditional on the location and community labels on nodes, we place an edge between nodes i and j in G_n with probability $f_{in}^{(n)}(||X_i - X_j||)$ if $Z_i = Z_j$ and with probability $f_{out}^{(n)}(||X_i - X_j||)$ if $Z_i \neq Z_j$, where $||\cdot||$ is the toroidal metric on B_n. We provide a more formal description of the random graph process in both regimes in Section 2.

In the sparse regime, we study the problem of 'weak-recovery', which asks how and when one can estimate the community labels of the nodes of G_n better than at random, given observational data of locations labels of all nodes and the graph G_n. We say weak recovery is solvable in the sparse case (made precise in Section 2.2) if there exists an algorithm which takes as input the graph G_n along with the location labels and produces a partition of the nodes 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. In the logarithmic regime, we consider the problem of 'Strong Recovery' or also known as *exact-*

2

recovery, which asks how and when can one recover the partition of nodes into communities *exactly* based on the observation of the random graph G_n and the location labels of all nodes. As this is a stronger requirement on the the estimator, the graph needs to be *sufficiently dense* in order to perform Exact-Recovery. More precisely, we see that the average node degree must scale logarithmically to the number of nodes to capture the phase transition for exact-recovery. In both of these problems, we assume that the estimator has access to the model parameters λ and $f_{in}^{(n)}(\cdot)$ and $f_{out}^{(n)}(\cdot)$. However, we present how one could possibly implement our algorithm in practice, when the connection functions are not known explicitly. From a mathematical perspective, 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.

Main Results and Contributions - Our main results in this paper pertain to fundamental phase-transitions, which dictate how and when we can do Community Detection in the two regimes of interest. In the sparse regime, we show that for every $f_{in}(\cdot)$ and $f_{out}(\cdot)$ that are not identical Lebesgue almost-everywhere, there exists a critical $\lambda_c \in (0, \infty)$ that depends on $f_{in}(\cdot)$ and $f_{out}(\cdot)$, such that if $\lambda < \lambda_c$, then no algorithm can estimate the community labels of G_n better than at random, and if $\lambda > \lambda_c$, then there exists an algorithm that can estimate the community labels better than at random. From a rather straightforward 'monotonicity' argument (in Proposition 4), one can establish the existence of $\lambda_c \in [0, \infty]$ such that the community detection problem shifts from being unsolvable to solvable. Our key technical result in this paper is to establish that this phase-transition is *non-trivial*, i.e., λ_c is neither 0 nor ∞. Furthermore, in certain special cases of $f_{in}(\cdot)$ and $f_{out}(\cdot)$, we are able to characterize exactly the phase-transition point λ_c. However in general, it is hard to characterize explicitly the phase-transition point, as even identifying the phase-transition for percolation of the random connection model, which is conceptually much simpler than CD, has been open for decades (see [32]).

The main contribution in establishing that $\lambda_c > 0$ in Theorem 6 is to identify an *easier* problem than weak-recovery, which we call Information Flow from Infinity. We study this problem by developing coupling and random-cluster like arguments. To show an upper bound, i.e. to establish weak-recovery is solvable for sufficiently high λ, we give a new algorithm called `Good-Bad-Grid` abbreviated as `GBG` 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 by simply considering their neighborhoods. Compared to the SBM, this is 'easier', since the geometric embedding provides common neighbors even in the sparse regime, whereas one needs to go down to neighbors of neighbors of large depth to obtain intersections in the sparse SBM, as done in [4] with the sphere comparison algorithm [1]. However in contrast to the SBM, this is 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 misclassified pairs by partitioning the space B_n into 'good' and 'bad' regions. Hence the name of our algorithm is `Good-Bad-Grid`. To analyze this algorithm, we then couple the partitioning of space with another percolation process to prove that our algorithm will misclassify a fraction strictly smaller than half of 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

---

[1]Counting common neighbors is also exploited in the algorithm of [18]

λ for these cases. Moreover, in Section 4.5 we give a way to implement our algorithm without any knowledge of the model parameters $f_{in}^{(n)}(\cdot), f_{out}^{(n)}(\cdot)$ and λ and is purely 'data-dependent'. Along the way, in Theorem 13, 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 in this regard lies in identifying suitable graph 'triangle profiles' and showing 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 in certain regimes. This is because we can always '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 if λ is small. This phenomenon is new and starkly different from what is observed in the classical Erdős-Rényi based symmetric Stochastic Block Models (SBM) with two communities where the moment one can identify the presence of a partition, one can also recover the partition better than a random guess ([37]). 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 [14],[5].

In the logarithmic degree regime, we give explicit conditions on the model parameters λ, $f_{in}^{(n)}(\cdot)$ and $f_{out}^{(n)}(\cdot)$ under which Exact-Recovery is impossible to solve. We establish this by reducing this problem to a Hypothesis testing problem between Poisson random vectors and them employing the Large-Deviations results of [4] to identify an explicit condition on the model parameters. On the positive side, we show that a direct adaptation of our GBG algorithm to this logarithmic degree regime performs Exact-Recovery if the intensity λ is sufficiently large, thereby establishing the phase-transition for Exact-Recovery. However, we conjecture that the algorithm is sub-optimal and that the lower bound identifies a tight. We describe a procedure that may achieve the lower bound, but leave this as an open problem.

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 ([37],[1]) while real social networks are observed to be *transitive* and sparse. Sparsity in social networks can be understood through 'Dunbar's number' [15], 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 the average 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. Similar phenomena also takes place in graphs of similarities, where vertices are connected based on metric similarity functions. These aspects point out the limitations of the sparse SBM and a large collection of models have been proposed

4

to better fit applications under the realm of Latent Space Models ([19],[20]) and inhomogeneous random graphs ([8]). See also [1] for more references. 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 an 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 ([19]). Such modeling assumptions in the context of multiple communities was also recently verified in parallel independent work [18], where the nodes have both a community label and a location label. The locations labels are sampled uniformly on a sphere and the edges are generated by nodes 'nearby' in this sphere connecting with probabilities that depend on the community labels. Several empirical validations of this model on real data is also conducted in [18] which suggests that such spatial random graph model provides a good fit for several real world networks. However, we note that the sparse SBM enjoys certain advantages over the geometric random graph considered here, namely that of having low diameter. in agreement with the 'small world' phenomena observed in many real world networks (see [46]). Therefore a natural next step is to superimpose an SBM with the type of geometric graphs considered here to obtain both a lot of triangles and small diameter, i.e. a type of small world SBM.

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,in this paper we assume the location labels on nodes are known exactly to the estimator. 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 missing. The problem with known spatial location labels is itself quite challenging as outlined in the sequel and hence we decided 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 in the sparse regime 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 [37]. 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 in the sparse regime 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 [1] and references therein). This fundamental difficulty renders the weak-recovery prob-

lem in the sparse regime 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 allows one to classify very accurately 'nearby' pairs 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. In the logarithmic degree regime, our main novelty is to notice that the lower bound for community detection can be reduced to reasoning about events in disjoint regions of space. We then exploit this reduction along with the fact that events of a Poisson process in disjoint regions of space are independent, along with certain symmetries of the Euclidean space, to provide an explicit lower bound for exact-recovery. In certain cases, we give a closed form expression of the phase-transition threshold by using the large deviations framework introduced in [4].

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 $f_{out}(\cdot)$ in the sparse case 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. Indeed this is best seen in our lower bound for exact-recovery case, where if $f_{out}^{(n)}(r)$ is identically 0, i.e. there are no cross community edges, then Exact-Recovery might still be possible even before the subgraph on the nodes of each individual communities become fully connected. This takes place since we can use the spatial location information in a non-trivial fashion to estimate the labels of isolated nodes. 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   Related Work

Community Detection on sparse graphs has mostly been studied on the SBM random graph model. The study of SBM has a long history in the different literatures of statistical physics (see [14] and references therein), mathematics (for ex. [7],[37]) and computer science (for ex. [30], [12],[36]). The reader should refer to the survey [1] for further background and references on the SBM. The survey [33] 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 [14], which made a 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 [37],[35], [29] and [9]. 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. The key ideas for all of our results come from different problems, mostly those studied in the theory of percolation [8] and stochastic geometry [40]. Our algorithm is

motivated by certain ideas that appeared in the Interacting Particle Systems literature (for ex. [40] [24],[41],[16]). The papers there developed re-normalization and percolation based ideas to study different particle systems arising in statistical mechanics, and our analysis bears certain similarity to that line of work. 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 this reduction comes from an ergodic argument applied on the spatial random graphs. To study the impossibility of Information Flow from Infinity, we employ certain 'random-cluster method' and coupling arguments. Such methods are quite popular and have proven to be extremely fruitful in other contexts for example to study mixing time of Ising Models ([28],[27],[26]). Similar coupling ideas have also appeared in other estimation contexts, notably the reconstruction on trees problem [34], where a lower bound was established using this method.

In the non-sparse regime, the work of [2] and [4] is most related to our methods for the non-sparse regime. Both papers studies the problem of Exact-Recovery in much greater detail in the generalized SBM, proving both lower bounds and efficient algorithms for exact recovery. In this paper, we use the lower bound framework developed in [4], specifically, the large deviations characterization of hypothesis testing between Poisson random vectors, to provide a lower bound for our spatial random graph model.

From a modeling perspective, the works of [47] and [18] which considers the Latent Space Models as a part of its model is the closest to our model. The paper of [47] gives a spectral algorithm that is proven to work in the logarithmic degree regime. In particular, their methods do not work for the sparse regime which is one of the central topics discussed in the present paper. The algorithm provided in [18] bears certain similarity to ours where nearby nodes' community membership are tested based on the common neighbors. However, it is only guaranteed to work in the logarithmic degree regime. Furthermore, no lower bounds are presented in [18], while the lower bound in [47] is based on the idea that the graph is locally tree-like, which is not the case in our model. Another related problem of group synchronization was introduced in [3], which among other things, also considered the question of how well can one identify the communities of two 'far-away' nodes. This paper establishes certain phase-transitions for weak recovery that look similar to our lower bound, using different ideas from percolation and random walks on grids. 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 [3].

1.2 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 for the sparse case is performed in Section 5, where the key idea is to construct coupling arguments with site percolations. We establish the lower bound for Community Detection in the sparse case 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. In Section 8, we discuss the non-sparse regime and study the Exact-Recovery problem. Thus,

Sections 4,6 and 7 all pertain to the sparse graph problem and can be read in any relative order. Section 8 exclusively only pertains to the non-sparse case and studies the Exact-Recovery problem.

2 Mathematical Framework and Problem Statement

We describe the mathematical framework based on stationary point processes and state the problem of Community Detection. More precisely, we assume the presence of a single infinite marked Poison Point Process (PPP)[2], and consider the random graphs G_n as an appropriate truncations of this infinite object. This representation allows us to couple the graphs G_n for all n on a single probability space. This representation plays a crucial role in certain proofs and arguments. In the sparse case, we can construct a single infinite random graph G and consider G_n as an appropriate finite subgraph of G. In the non-sparse case, there is no direct limiting infinite graph. Nevertheless, we can couple all the graphs G_n on a single probability space by constructing them on a single marked PPP. We set a common shorthand notation we use throughout the paper. For two arbitrary positive sequences $(a_n)_{n\in\mathbb{N}}$ and $(b_n)_{n\in\mathbb{N}}$, we let $b_n = o(a_n)$ to denote the fact that $\lim_{n\to\infty} b_n/a_n = 0$.

2.1 The Planted Partition Random Connection Model

We suppose there exists an abstract probability space $(\Omega, \mathcal{F}, \mathbb{P})$ on which we have an appropriately marked PPP $\bar{\phi}$. We will construct the sequence of random graphs $(G_n)_{n\in\mathbb{N}}$ on this space simultaneously for all n as a measurable function of this marked PPP $\bar{\phi}$. More formally, we assume ϕ to be the support of a homogeneous PPP of intensity λ on $\mathbb{R}^d$ with the enumeration that $\phi := \{X_1, X_2, \cdots\}$. In this representation, $X_i \in \mathbb{R}^d$ for all $i \in \mathbb{N}$. Furthermore, we assume the enumeration to be such that for all $i > j \in \mathbb{N}$, we have $||X_i||_\infty \geq ||X_j||_\infty$, i.e., the points are ordered in accordance to increasing l_∞ distance. 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 $i \neq j \in \mathbb{N}$. We denote by $\bar{\phi}$ to be this marked PPP. 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 this marked point process is that for any node $i \in \mathbb{N}$, its location label is X_i, community label is Z_i and $\{U_{ij}\}_{j\in\mathbb{N}\setminus\{i\}}$ are used to sample the graph neighbors of node i. We describe this construction in both the sparse and logarithmic degree regimes below.

Denote by the set $B_n := \left[-\frac{n^{1/d}}{2}, \frac{n^{1/d}}{2}\right]^d$, the cube of area n in $\mathbb{R}^d$. For all $n \in \mathbb{N}$, we let $N_n := \sup\{i \geq 1 : X_i \in B_n\}$. Since the nodes are enumerated in increasing order of l_∞ distance, it follows that for all $i \in [1, N_n]$, $X_i \in B_n$. Furthermore, from basic properties of PPP, N_n is a Poisson random variable of mean λn. We construct the graph G_n, by assuming its vertex set to be $\{1, \cdots, N_n\}$ and the location label of any node $i \in [1, N_n]$ to be $X_i \in B_n$ and its community label to be $Z_i \in \{-1, +1\}$. However, we use the marks $\{U_{ij}\}_{j\in\mathbb{N}\setminus\{i\}}$ slightly differently depending on whether the graph is sparse or not. Recall that in the sparse regime, we have the connection functions $f_{in}^{(n)}(\cdot)$ and $f_{out}^{(n)}(\cdot)$ to be independent of n. In this case, we first construct an infinite graph G with vertex set $\mathbb{N}$ and place an edge between any two nodes $i, j \in \mathbb{N}$ if and only if $U_{ij} = U_{ji} \leq f_{in}(||X_i - X_j||)\mathbf{1}_{Z_i=Z_j} + f_{out}(||X_i - X_j||)\mathbf{1}_{Z_i\neq Z_j}$. The graph G_n is then the

---

[2]This is defined in Appendix E

induced subgraph of G consisting of the nodes 1 through N_n, i.e., the subgraph of G restricted to the node set with location labels in B_n. In the logarithmic degree regime, we assume that the set B_n is equipped with the toroidal metric rather than the Euclidean metric for simplicity. Formally, for any $x := (x_1, \cdots, x_d), y := (y_1, \cdots, y_d) \in B_n$, the toroidal distance on B_n is given by $||x - y||_{\mathcal{T}_n} = ||(\min(|x_1 - y_1|, n^{1/d} - |x_1 - y_1|), \cdots, \min(|x_d - y_d|, n^{1/d} - |x_d - y_d|))||$, where $||.||$ is the standard Euclidean norm on $\mathbb{R}^d$. For any $i, j \in [1, N_n]$, we draw an edge between nodes i and j in G_n if and only if $U_{ij} = U_{ji} \leq f_{in}^{(n)}(||X_i - X_j||_{\mathcal{T}_n})\mathbf{1}_{Z_i = Z_j} + f_{out}^{(n)}(||X_i - X_j||_{\mathcal{T}_n})\mathbf{1}_{Z_i \neq Z_j}$.

The infinite random graph G in the sparse regime can be viewed as a 'planted-partition' version of the classical random-connection model ([32]). 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.

2.2 The Community Detection Problem

In this paper, we study two different notions of Community Detection - weak recovery and exact recovery, depending on whether the graph G_n is sparse or non-sparse. In the sparse regime, our definition of Community Detection is analogous to the notion of 'weak-recovery' considered in the classicial SBM literature ([14]) and in the logarithmic degree regime, our definition of Community Detection is analogous to the notion of 'Exact-Recovery' ([2]) in the SBM literature. To state the two notions of community recovery, we set more notation. Let ϕ_n be the restriction of the point process ϕ to the set B_n. Notice that the cardinality of ϕ_n is N_n which is distributed as a Poisson random variable of mean λn, and $X_i \in B_n$ for all $i \in [1, N_n]$. Moreover, conditionally on N_n, the location variables $(X_i)_{i \in [1, N_n]}$ are placed uniformly and independently in B_n. 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.*

We define two notions of performance of community detection, weak recovery and exact recovery defined below.

Definition 2. Weak Recovery *is said to be* **solvable in the sparse regime** *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*

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

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

In the above definition, we let the overlap $\mathcal{O}_n := 1$ if $N_n = 0$. In the logarithmic degree regime, we ask for exact-recovery which is formally stated as follows.

Definition 3. Exact-Recovery *is said to be* **solvable in the logarithmic degree regime** *for* $\lambda, d, (f_{in}^{(n)}(\cdot))_{n \in \mathbb{N}}$ *and* $(f_{out}^{(n)}(\cdot))_{n \in \mathbb{N}}$ *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*

$$\lim_{n \to \infty} \mathbb{P}\left[\mathcal{O}_n = 1\right] = 1, \tag{2}$$

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

A key new feature of our Definitions 2 and 3 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. In the sparse regime, we ask when can any estimator assign community labels to the nodes that beats a 'random guess'. Observe that if an estimator guessed every node to be in Community +1, then the achieved overlap $\mathcal{O}_n$ converges almost-surely to 0, thanks to the Strong Law of Large Numbers. Thus, achieving a positive γ asks whether an estimator can asymptotically beat the trivial estimator. In the non-sparse regime, we ask when and how can we recover the commuity label of *all* nodes, upto a global flip.

Observe that we assume the algorithm has access to the parameters $f_{in}^{(n)}(\cdot)$, $f_{out}^{(n)}(\cdot)$ 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_n is symmetric in the community labels. In particular, if we flipped all community labels of G_n, we would observe a graph which is equal in distribution to G_n. 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. In the non-sparse case, since we consider the set B_n as a torus, to precisely avoid the technicalities arising out of considering the edge effects.

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

Proposition 4. *For every* $f_{in}(\cdot), f_{out}(\cdot)$ *and* d, *there exists a* $\lambda_c \in [0, \infty]$ *such that*

- $\lambda < \lambda_c \implies$ *weak-recovery is not solvable.*

- $\lambda > \lambda_c \implies$ *there exists an algorithm (which could possibly take exponential time) to solve weak-recovery.*

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 λ', 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 with probability at-least $1 - o_n(1)$, 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 λ' will be at-least $\frac{\lambda \gamma}{\lambda'}$ with probability at-least $1 - o_n(1)$. Thus, the problem of community detection solvability is monotone in λ in the sparse case. $\qquad \square$

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.

Distinguishability of the Planted Partition in the Sparse Regime

In this paper, we also consider a related problem to weak-recovery, namely the *distinguishability* question in the sparse regime. Roughly speaking, this problem asks whether one can identify if the observed data (G_n, ϕ_n) has communities or not. More precisely, we have the following definition.

Definition 5. *The sparse planted partition model with parameters λ, d and connection functions $f_{in}(\cdot)$ and $f_{out}(\cdot)$ is said to be* distinguishable, *if for every $g(\cdot) : \mathbb{R}_+ \to [0, 1]$, one can identify with probability at-least $\frac{1}{2} + \gamma - o_n(1)$ for some $\gamma > 0$, whether the observed data (G_n, ϕ_n) is a sample of the planted partition with the above parameters, or is a sample of the random connection model $H_{\lambda, g(\cdot), d}$, given an uniform prior over the two models.*

This problem asks if we can even identify the existence of communities, before trying to identify them. Such hypothesis testing questions are of critical importance in practice where one needs to be reasonably sure of the presence of communities in a given graph, before attempting to cluster the graph. Our main result in Theorem 13 is that one can always solve the distinguishability problem, i.e. it exhibits no phase-transition. Thus our results on weak-recovery and distinguishability predicts regimes of the problem (i.e. $d = 1$ and $\lambda < \lambda_c$ for $d \geq 2$), where we can be very sure of the presence of communities, but cannot recover it better than at random.

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 [13] for the general theory of Palm measures. However thanks to Slivnyak's theorem [13], 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. This extra node is also equipped as a mark the sequence $\{U_{0j}\}_{j \geq 1}$, which will be used to sample the edges as before. Informally, we place an edge between this extra point at 0 and any other node $i \in \mathbb{N}$ in the graph G_n with probability $f_{in}^{(n)}(||X_i||)$ if $Z_i = Z_0$, or with probability $f_{out}^{(n)}(||X_i||)$ if $Z_i \neq Z_0$,

independently of everything else. More generally, for any $k \in \mathbb{N}$ and $x_1, \cdots x_k \in \mathbb{R}^d$, we denote by the measure $\mathbb{P}^{x_1, \cdots, x_k}$ to be the union of $\bar{\phi}$ from before with additional points placed at $x_1, \cdots x_k$, with them having independent marks as before. Informally, the community labels of the extra nodes located at $x_1, \cdots, x_k$ is uniform and independent, and will have edges among themselves and the points of ϕ with the same distribution as before. However, for the logarithmic degree regime discussed in Section 8, since the set B_n is equipped with the toroidal metric, we need a definition of Palm measure on the torus rather than Euclidean plane. We give the necessary adaptation of Palm measure needed to handle the torus case in the Appendix F.

3 Main Results

3.1 Lower Bound for Weak Recovery

To state the main lower bound result in Theorem 6 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 6. *If $\theta(H_{\lambda, f_{in}(\cdot) - f_{out}(\cdot), d}) = 0$, then weak-recovery 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 7. *For all $\lambda > 0$, $f_{in}(\cdot), f_{out}(\cdot)$ such that $\int_{x \in \mathbb{R}} f_{in}(||x||)dx < \infty$, weak-recovery 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 $\int_{x \in \mathbb{R}} g(||x||)dx < \infty$, $\theta(H_{\lambda, g(\cdot), 1}) = 0$. $\square$

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

Corollary 8. *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 weak-recovery 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 [32], by comparison with a branching process, we see that $\lambda \int_{x \in \mathbb{R}^d} g(||x||) \leq 1 \implies \theta(H_{\lambda, g(\cdot), d}) = 0$. $\square$

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 8 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 9. *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 weak-recovery 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 6, 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 weak-recovery, 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 weak-recovery. 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.

3.2   Algorithm and an Upper Bound for Weak Recovery

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 10. *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 weak-recovery problem. Moreover, `GBG` when run on data (G_n, ϕ_n), has time complexity order n^2 and storage complexity order n.*

Remark 11. *An upper bound to λ_{upper} is given in Proposition 39.*

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.

We also note that our algorithm is asymptotically optimal in a weak sense made precise in the sequel below. Denote by $\mathcal{O}_\lambda$ as the maximum overlap achieved by our algorithm, with the definition of overlap as given in Definition 1. More precisely, denote by $\mathcal{O}_\lambda$ as

$$\mathcal{O}_\lambda := \sup\{\gamma \geq 0 : \lim_{n \to \infty} \mathbb{P}[\mathcal{O}_n > \gamma] = 1\}, \tag{3}$$

where $\mathcal{O}_n$ is the overlap achieved by the `GBG` algorithm when run on the data G_n and ϕ_n. Thus, for each fixed λ, the overlap $\mathcal{O}_\lambda \in [0, 1]$ is determinstic quantity.

Proposition 12. *We have the limit that $\lim_{\lambda \to \infty} \mathcal{O}_\lambda = 1$.*

In words, this proposition states that as the 'signal' gets stronger, the fraction of nodes correctly classified by our algorithm tends to 1.

On a related note, we also mention in Section 4.5, a practical way of implementing the algorithm when the connection functions $f_{in}(\cdot)$ and $f_{out}(\cdot)$ are not known explicit.

3.3   Distinguishability 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 13. *For every $\lambda > 0$, $d \in \mathbb{N}$ 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 13 implies that this distinguishability problem as stated in Definition 5 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 ([37]) and conjectured not to occur in any arbitrary SBM ([14]). Technically, 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.

3.4   Phase Transition for Exact-Recovery

Our main achievement in the present paper is an explicit necessary condition on the model parameters for Exact Recovery given in the following Theorem. To highlight the ideas, we primarily focus on a simplest non-trivial example of connection functions in the logarithmic regime, where we are able to conjecture a closed form expression for the phase-transition threshold. However, we present the results on the general case in Proposition 53 in the Sequel in Section 8.

Definition 14. *For any $0 \leq b < a \leq 1$, $n \in \mathbb{N}$, $\lambda > 0$ and $d \in \mathbb{N}$, we denote by $\mathcal{G}(\lambda n, a, b, d)$ as the distribution of graph G_n we defined in Section 2 with $f_{in}^{(n)}(r) = a\mathbf{1}_{r \leq \log(n)^{1/d}}$ and $f_{out}^{(n)}(r) = b\mathbf{1}_{r \leq \log(n)^{1/d}}$ for all $r \geq 0$, where the distance on the set B_n is the torridal metric.*

In the results that follow, denote by ν_d to be the volume of the unit Euclidean unit ball in d dimensions.

Theorem 15. *For any $\lambda > 0$, $d \in \mathbb{N}$ and $0 \leq b < a \leq 1$ such that $\lambda \nu_d (1 - \sqrt{ab} - \sqrt{(1-a)(1-b)}) < 1$, Exact-Recovery of $G_n \sim \mathcal{G}_n(\lambda n, a, b, d)$ is not solvable.*

This connection functions are the simplest non-trivial instance of our model that makes the study interesting. We also believe that this necessary condition to be tight. However, at this point, we only present the following conjecture and do not pursue a proof of this.

Conjecture 16. *For any $\lambda > 0$, $d \in \mathbb{N}$ and $0 \leq b < a \leq 1$ such that $\lambda \nu_d (1 - \sqrt{ab} - \sqrt{(1-a)(1-b)}) > 1$, Exact-Recovery of $G_n \sim \mathcal{G}_n(\lambda n, a, b, d)$ is solvable.*

We believe two-round techniques developed in [4] applied to the spatial graph case can be fruitful in establishing this conjecture. To establish the existence of a phase-transition however, we analyze the `GBG` algorithm and adapt it to the non-sparse case to yield the following result.

Theorem 17. *For every $\lambda > 0$, $d \in \mathbb{N}$ and $0 \le b < a \le 1$, there exists $C(a, b, d) > 0$ such that if $\lambda \nu_d (1 - \sqrt{ab} - \sqrt{(1-a)(1-b)}) > C(a, b, d)$, Exact-Recovery of $G_n \sim \mathcal{G}_n(\lambda n, a, b, d)$ is solvable by* `GBG` *algorithm.*

This theorem gives the existence of different phases of the exact-recovery problem depending on λ. In particular, it states that if the intensity λ is sufficiently high, then Exact-Recovery is solvable by our `GBG` algorithm. Using similar ideas as for the sparse regime, we mention in Section 8.2, a way of implementing the algorithm without knowledge of the problem parameters.

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. We make the presentation here assuming that G_n is sparse, although a straightforward adaptation can be made to apply this algorithm in the logarithmic degree regime as well. We will make this adaptation more precise in the sequel in Section 8. The algorithm we present and analyze requires the knowledge of the model parameters $\lambda, f_{in}^{(n)}(\cdot)$ and $f_{out}^{(n)}(\cdot)$, although we show in Section 4.5, that by a simple modification, we can implement the algorithm even if the model parameters are unknown to the algorithm.

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 $f_{in}(r) = a\mathbf{1}_{r \le R}$ and $f_{out}(r) = b\mathbf{1}_{r \le R}$ for some $R > 0$ and $0 \le b < a \le 1$. 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 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 Notation 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 10. Thus, there exists $0 \le \tilde{r} < R < \infty$ 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 18. *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 19. *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 $||z - z^{'}||_\infty \le 1$.*

Definition 20. *For any subset $A \subset \mathbb{Z}^d$ and any $k \in \mathbb{N}$, the k **thickening of** A is denoted by $\mathbf{L}_k(A) := \cup_{z \in A} \cup_{z' \in \mathbb{Z}^d : ||z-z'||_\infty \le k} z^{'}$.*

Definition 21. *For any set $B \subseteq \mathbb{Z}^d$, denote by the set $Q_B := \cup_{z \in B} Q_z$.*

Definition 22. *Let $\mathcal{Z}(\cdot) : \mathbb{R}^d \to \mathbb{Z}^d$ be the projection function, i.e. $\mathcal{Z}(x) := \inf\{z \in \mathbb{Z}^d : ||\frac{Rz}{4d^{1/d}} - x||_\infty \le 0.5\}$. In case, of more than one z achieving the minimum, we take the lexicographically smallest such z.*

Definition 23. *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 24. *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 25. *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

1: **procedure** PAIRWISE-CLASSIFY(i,j,ϕ,G)
2:   **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
3:   **else**
4: **return** -1
5:   **end if**
6: **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

1: **procedure** IS-A-GOOD(z,G)
2:   **if** $|\phi \cap Q_z| < \lambda(R/4)^d(1/d)(1-\epsilon)$ **then return** FALSE
3:   **end if**
4:   $\phi^{(z)} := \phi \cap \left(\cup_{z':||z-z'||_\infty \leq 1} Q_{z'}\right)$
5:   **for** all $\forall k \geq 1$, and all $X_1, \cdots X_k \in \phi^{(z)}$ **do**
6:     **if** $\prod_{i=1}^k$ PAIRWISE-CLASSIFY$(X_i, X_{i+1}, G) = -1$ **then**   ▷ Where $X_{k+1} := X_1$
7:       **return** FALSE
8:     **end if**
9:   **end for**
10:   **return** TRUE
11: **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 26. *A cell Q_z is **A-Good** if*

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 $\hat{\tau}_{l_1}^{(n)} = +1$
7:         **for** $w = 2, w \leq n_j$ **do**
8:             Set $\hat{\tau}_{l_w}^{(n)} = \text{Pairwise-Classify}(l_{w-1}, l_w, \phi_n, G_n)\hat{\tau}_{l_{w-1}}^{(n)}$
9:         **end for**
10:     **end for**
11:     **for** $c = 1, c \leq N_n$ **do**
12:         **if** $\hat{\tau}_c^{(n)} = 0$ **then**
13:             Set $\hat{\tau}_c^{(n)} = +1$
14:         **end if**
15:     **end for**
16:     **return** $\{\hat{\tau}_i^{(n)}\}_{i=1}^{N_n}$
17: **end procedure**

---

1. $\left| \phi \cap Q_z \right| \geq \max\left(\lambda \left(\frac{R}{4} \right)^d \frac{1}{d}(1 - \epsilon), 1 \right)$; 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 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.
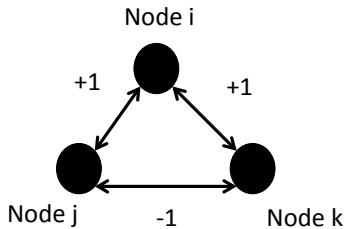
Figure 1: An illustration where Pairwise-Classify leads to inconsistency. The values on the edges represent the output of pairwise classify run on the two end points as inputs. In this example it is clear that for at-least one pair $(i, j), (j, k), (k, i)$, the output of pairwise estimate is different from the ground truth.

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 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.

4.5 Practical Implementation if Model Parameters are Unknown

In this section, we provide a simple alternative that can be used to cluster even when the model parameters $f_{in}^{(n)}, f_{out}^{(n)}(\cdot)$ and λ are unknown to the algorithm. Assume for simplicity, that we know an estimate of R such that $\int_{x \in B(0,R)}(f_{in}(||x||) - f_{out}(||x||))dx > 0$, i.e. the set $\{r \in [0, R] : f_{out}(r) < f_{in}(r)\}$ has non-zero Lebesgue measure. Then, we can change the definition of A-Good as follows. For any grid cell $z, z^{'} \in \mathbb{Z}^d$ such that $||z - z^{'}||_\infty = 1$, consider the subgraph $G_{z,z^{'}}$ of G consisting of nodes whose locations lie either in cell z or $z^{'}$. Consider applying some known partition to the nodes of $G_{z,z^{'}}$, for instance the standard spectral method described in [1]. We can denote the cell z to be A-Good, if the number of points of ϕ in that cell is no smaller than $(1 - \epsilon)$ of the expected
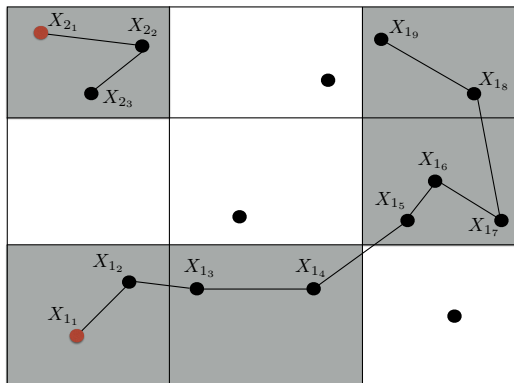
Figure 2: An illustration of algorithm 3. In this example, *we do not draw the graph G*, but only show the locations of the nodes. The shaded cells corresponds to A-Good cells and in this example there are two A-Good connected components. In each component, we outline an arbitrary sequence of points $X_{1_1} \cdots X_{1_9}$ 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.

value, *and* for every z' in the 1-thickening of z, the partition of the nodes of G_z, when the spectral method is applied to the induced sub-graph $G_{z,z'}$ is identical, i.e. for every z', z'' in the 1 thickening of z, the partition of the nodes of G_z is same whether the spectral method is run on the graph $G_{z,z'}$ or $G_{z,z''}$. Since the spectral method of [1] does not need to know the connection functions, one can use this as an alternative definition of A-Good cell in place of Algorithm 2. The only model information in this alternative implementation required is an estimate of R to perform the tessellation of space. Thus in line 3 of Algorithm 3, we can invoke the test described in this paragraph which does not need knowledge of the connection functions, as opposed to involing Algorithm 2 which does need knowledge of the parameters.

From an analysis perspective, we can define an alternate notion of a T-Good cell. We can denote by a cell $z \in \mathbb{Z}^d$ as T-Good, if it contains at-least $(1 - \epsilon)$ of the mean number of points in that cell. Additionally, a grid cell z is T-Good if for every $z' \in \mathbb{Z}^d$ such that $||z - z'||_\infty = 1$, the spectral method on the graph $G_{z,z'}$ matched the ground truth. One can then carry an identical analysis to this spectral method for Is-Good, as done for the version presented in Algorithm 2.

5 Analysis and Proof of the Algorithm in the Sparse Regime

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

Theorem 27. *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 weak-recovery.*

Remark 28. *We provide an upper bound to the constant λ_0 in the sequel in Proposition 37.*

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 29. *A cell Q_z is **T-Good** if -*

1. $\left| \phi \cap Q_z \right| \geq \max \left(\lambda \left(\frac{R}{4} \right)^d \frac{1}{d}(1 - \epsilon), 1 \right);$ *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.*

*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 27 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 comprising of many nodes.

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 30. *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 weak-recovery problem as given in Definition 2.*

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

Proposition 31. *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 $+1$. $\square$

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 32. *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. The proof of this Proposition is quite standard and is defered to the Appendix in B. We are now in a position to conclude the proof of Proposition 30.

Proof. Proof of Proposition 30

Proposition 32 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 32 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 32. From the definition of T-Good components, the unique partition referred to in Proposition 32 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. $\qquad \square$

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 recall 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 $\mathbb{P}^{x_1, \cdots x_k}$ 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 33. *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

$$\mathbb{E}_\phi^{x,y}\left[\sum_{z \in \phi \setminus \{x,y\}} F(z)\right] = \lambda \int_{z \in \mathbb{R}^d} \mathbb{E}_\phi^{x,y,z}[F(z)]dz. \tag{4}$$

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. $\square$

Proposition 34. *For all connection functions $f_{in}(\cdot)$ and $f_{out}(\cdot)$ satisfying the hypothesis of Theorem 10, 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*

$$\mathbb{P}^{x,y}[(x,y) \text{ is misclassified by Algorithm 1}] \leq e^{-c\lambda}, \tag{5}$$

where the constant c satisfies

$$c \geq \inf_{x,y \in \mathbb{R}^d : ||x-y||_2 \leq 3R/4} (\mathbf{1}_{M_{out}(x,y)>0}M_{out}(x,y) + \mathbf{1}_{M_{out}(x,y)=0}M_{in}(x,y))h\left(\frac{M_{in}(x,y) - M_{out}(x,y)}{2M_{in}(x,y)}\right), \tag{6}$$

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 33, 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

$$\mathbb{P}^{x,y}[\text{points at } x \text{ and } y \text{ are mis-classified }] =$$
$$\frac{1}{2}\mathbb{P}\left[X \geq \lambda\frac{M_{in}(x,y) + M_{out}(x,y)}{2}\right] + \frac{1}{2}\mathbb{P}\left[Y \leq \lambda\frac{M_{in}(x,y) + M_{out}(x,y)}{2}\right], \quad (7)$$

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

$$\mathbb{P}^{x,y}[\text{points at } x \text{ and } y \text{ are mis-classified }] \leq$$
$$\frac{1}{2}e^{-\lambda M_{out}(x,y)h\left(\frac{M_{in}(x,y) - M_{out}(x,y)}{2M_{in}(x,y)}\right)} + \frac{1}{2}e^{-\lambda M_{in}(x,y)h\left(\frac{M_{in}(x,y) - M_{out}(x,y)}{2M_{in}(x,y)}\right)}, \quad (8)$$

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

$$\sup_{x,y\in\mathbb{R}^d:||x-y||_2\leq(3/4)R} \mathbb{P}^{x,y}[\text{points at } x \text{ and } y \text{ are mis-classified }] \leq e^{-c\lambda}, \quad (9)$$

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

$\square$

Lemma 35. *For all $z \in \mathbb{Z}^d$,*

$$\mathbb{P}[\text{Cell } z \text{ is T-Good in graph } G] \geq 1 - e^{-\lambda(R/4)^d\frac{1}{d}h(\epsilon)} - \lambda^2(3R/4)^d\frac{1}{d}e^{-c\lambda}, \quad (10)$$

where the constant c and function $h(\cdot)$ are defined in Proposition 34.

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 $e^{-\lambda(R/4d^{1/d})^dh(\epsilon)}$, 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 $\mathbb{P}[X > 0] \leq$

24

$\mathbb{E}[X]$. 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

$$\mathbb{E}[\sum_{i,j \in \mathbb{N}} \mathbf{1}_{X_i, X_j \in \mathbf{L}_1(z)} \mathbf{1} \text{ Algorithm 1 mis-classifies } i \text{ and } j]. \tag{11}$$

From the Moment-Measure expansion and the Campbell-Mecke theorem for an independently marked PPP ([11]), we obtain

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

$$= \lambda^2 \int_{x \in Q_{\mathbf{L}_1(z)}} \int_{y \in Q_{\mathbf{L}_1(z)}} \mathbb{P}^{x,y}[\text{points at } x \text{ and } y \text{ are mis-classified }] dx dy \leq \lambda^2 \left(\frac{3R}{4}\right)^d \frac{1}{d} e^{-c\lambda}. \tag{12}$$

The last inequality follows directly from Proposition 34. Therefore, by a simple union bound, we see that

$$\mathbb{P}[\text{ Cell } z \text{ is T-Bad}] \leq e^{-\lambda(R/4d^{1/d})^d h(\epsilon)} + \lambda^2 \left(\frac{3R}{4}\right)^d \frac{1}{d} e^{-c\lambda}. \tag{13}$$

The proposition is proved by taking complements. $\qquad\square$

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 36. *For every $p \in (0,1)$, and every $f_{in}(\cdot)$ and $f_{out}(\cdot)$ satisfying the hypothesis of Theorem 10, there exists a $\lambda^{'}$ such that for all $\lambda > \lambda^{'}$, and all $z \in \mathbb{Z}^d$, $\mathbb{P}[Cell \ z \ is \ T\text{-}Good] \geq p$.*

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

$$\lim_{\lambda \to \infty} p(\lambda) \geq \lim_{\lambda \to \infty} 1 - e^{-\lambda(R/4)^d \frac{1}{d} h(\epsilon)} - (\lambda^2 + \lambda)(3R/4)^d \frac{1}{d} e^{-c\lambda} = 1, \tag{14}$$

where c is given in Proposition 34.

$\qquad\square$

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 37. *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 $j \in \mathbb{Z}^d$*

$$\liminf_{n \to \infty} \frac{1}{(2n)^d} \sum_{i \in \mathbb{Z}^d : ||i-j||_\infty \leq n} \mathbf{1}_{i \in \mathcal{C}(j)} \geq \frac{1}{2} + \eta, \tag{15}$$

$\mathbb{P}$ *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 36 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 [24] provides a clean 'black-box' methodology to establish this and our proposition can be viewed as a direct corollary of Theorem 1 in [24]. 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 $M := \lceil 12 d^{1/d} \rceil$ dependent. Moreover, thanks to Proposition 34, for every $z \in \mathbb{Z}^d$,

$$\mathbb{P}[Y_z = 1 | \sigma(Y_u : u \in \mathbb{Z}^d, ||u - z||_\infty > M)] \geq p(\lambda), \ \mathbb{P} \text{ a.s. }, \tag{16}$$

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

Thus, from Theorem 1 in [24], 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 [24] 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 $\tilde{\mathcal{C}}'(0)$ 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 $\theta_d(\lambda) := \mathbb{P}'[|\tilde{\mathcal{C}}'(0)| = \infty]$. From a direct application of Peirl's argument ([8], 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 36, this can be rephrased as $\lim_{\lambda\to\infty} \theta_d(\lambda) = 1$.

The stochastic domination in Line 2 above yields

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

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

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

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

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

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

$$\lim_{n\to\infty} \frac{1}{(2n)^d} \sum_{i\in\mathbb{Z}^d: ||i-j||_\infty \leq n} \mathbf{1}_{|\tilde{\mathcal{C}}'(i)|=\infty} = \theta_d(\lambda) \quad \mathbb{P}' \text{ a.s.} \tag{20}$$

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 $(Y'_z)_{z\in\mathbb{Z}^d} \stackrel{(d)}{=} (Y_z)_{z\in\mathbb{Z}^d}$.

$\square$

5.4   Concluding that Weak-Recovery is Solvable

The following proposition along with Proposition 30 will conclude the proof of Theorem 27.

Proposition 38. *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 37, 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 37, 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 37 that

$$\liminf_{n \to \infty} \frac{1}{(2n)^d} \sum_{i \in \mathbb{Z}^d : ||i-z||_\infty \leq n} \mathbf{1}_{i \in \mathcal{C}(z)} \geq \frac{1}{2} + \eta \ \mathbb{P} \text{ a.s.} \tag{21}$$

However, from an elementary counting argument, we conclude that

$$\liminf_{n \to \infty} \frac{1}{(2n)^d} \sum_{i \in \mathbb{Z}^d : ||i||_\infty \leq n} \mathbf{1}_{i \in \mathcal{C}(z)} \geq \frac{1}{2} + \eta \ \mathbb{P} \text{ a.s.} \tag{22}$$

In other words, the reference point does not matter when considering the limit, which can be seen easily by the following in Equation (23) .

$$\frac{1}{(2(n+z))^d} \sum_{i \in \mathbb{Z}^d : ||i-z||_\infty \leq n} \mathbf{1}_{i \in \mathcal{C}(z)} \leq \frac{1}{(2n)^d} \sum_{i \in \mathbb{Z}^d : ||i-z||_\infty \leq n} \mathbf{1}_{i \in \mathcal{C}(z)}$$

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

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

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

$$= \mathrm{O}(n^{1-d}),$$

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 37 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

28

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

almost-surely, i.e., $\lim_{n\to\infty} \mathbb{P}\left[\alpha_n > \frac{1}{2} \Big| |\mathcal{C}(z)| = \infty\right] = 1$. 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.

$\square$

5.5 Proof of Proposition 12

From Proposition 38, we know that for every $\epsilon \in (0,1)$ and $\eta \in (0, \frac{1}{2}])$, there exists $\lambda_0(\epsilon, \eta) < \infty$, such that for all $\lambda > \lambda_0(\epsilon, \eta)$, the GBG algorithm achieves an overlap of $(\frac{1}{2} + \eta)(1 - \epsilon)$. The proof is concluded by noticing that for any $\delta \in (\frac{1}{2}, 1)$, we can choose $\epsilon \in (0, 1)$ and $\eta \in (\frac{1}{2}, 1)$ such that $(\frac{1}{2} + \eta)(1 - \epsilon) > \delta$.

5.6 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 39. *If λ satisfies*

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

where the constant c is given in Proposition 34, $M := \lceil 12d^{1/d} \rceil$, 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 [8]. We outline it in the Appendix C for the sake of completeness. Observe that we did not use Peirl's argument directly in Proposition 37 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 6. 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 45. To aid us in developing the technical arguments, it is instructive to first consider the proof of Proposition 9 (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 9

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 $i_{j-1} \nsim_{G_n} i_j$ 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 40. *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 45, Proposition 40 will imply Proposition 9.

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 -

30

- $Z_i = Z_j$ 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 $T_n \overset{(d)}{=} H_{\lambda, f_{in}(\cdot) - f_{out}(\cdot), d}$.

$\square$

In the sequel, we will generalize the above example to come up with the general lower bound for Community Detection problem.

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 obtained by placing an additional node at the origin and equipping it with an independent community label and edges to other existing nodes. 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 41. *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 [6] and [37]). We first establish a monotonicity property of this problem and connect it with the Community Detection Problem.

31

Proposition 42. *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 42 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, [32]). 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 ϕ' 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 $r \geq 0$, $\tilde{\xi}(\lambda', r) \leq \tilde{\xi}(\lambda, r)$. $\qquad\qquad\square$

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

Proposition 43. *([13]) 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 44. *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 44.

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 43. 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

33

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 45. *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 44, 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 45, then we will conclude the proof of Theorem 6.

Before presenting the proof of Theorem 45, we illustrate a few example setting where the bound in Equation 28 is tight and loose respectively. In view of Lemma 44 and Proposition 9, the following corollary where Equation (28) is tight holds.

Corollary 46. *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 45 is achieved in certain examples. However, Theorem 45 is not an accurate characterization of the Information Flow from Infinity problem as evidenced in the following example.

Proposition 47. *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 47 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 [45], 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 [45], and hence perfect recovery is not possible. We present the complete proof in the Appendix A.

6.4 The Information Graph and Proof of Theorem 45

In this section, we generalize the example of the previous section and give a proof of Theorem 45. 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 $I \stackrel{(d)}{=} H_{\lambda, f_{in}(\cdot)-f_{out}(\cdot),d}$, 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 48. *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 49. *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 48 with the observed data G, ϕ and I.*

The proof of this follows from Lemma 48 and an induction argument. The proof can be found in Appendix D. We now present the main probabilistic observation in the sequel in Lemma 50 which essentially states that the community labels on disconnected components are independent.

Lemma 50. *For all $\lambda > 0$, on the event $\{|V_I(0)| < \infty\}$,*

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

Proof. From Lemma 49, 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 48. 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.

$$\mathbb{P}^0_\phi[(Z_k)_{k \in V_I(0)} = \mathbf{s} | \phi, \{\{U_{kl}\}_{l>k}\}_{k \in \mathbb{N} \cup \{0\}}, (Z_k)_{k \in V_I^c(0)}, G = g]$$

$$= \frac{\mathbb{P}^0_\phi[G = g | \phi, \{\{U_{kl}\}_{l>k}\}_{k \in \mathbb{N} \cup \{0\}}, (Z_k)_{k \in V_I^c(0)}, (Z_k)_{k \in V_I(j)} = \mathbf{s}]}{\mathbb{P}^0_\phi[G = g | \phi, \{\{U_{kl}\}_{l>k}\}_{k \in \mathbb{N} \cup \{0\}}, (Z_k)_{k \in V_I^c(0)}]} \tag{29}$$

$$\mathbb{P}^0_\phi[(Z_k)_{k \in V_I(0)} = \mathbf{s} | \phi, \{\{U_{kl}\}_{l>k}\}_{k \in \mathbb{N} \cup \{0\}}, (Z_k)_{k \in V_I^c(0)}]$$

$$\overset{(a)}{=} \frac{1}{\sum_g \mathbb{P}^0_\phi[G = g | \phi, \{\{U_{kl}\}_{l>k}\}_{k \in \mathbb{N} \cup \{0\}}, (Z_k)_{k \in V_I^c(0)}]} \left(\frac{1}{2}\right)^{|V_I(0)|} \quad \text{a.s. on the event } \{|V_I(0)| < \infty\} \tag{30}$$

$$\overset{(b)}{=} \frac{1}{2} \text{ a.s. on the event } \{|V_I(0)| < \infty\} \tag{31}$$

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 49 which states that g is consistent with the data $(\phi, \{\{U_{kl}\}_{l>k}\}_{k \in \mathbb{N} \cup \{0\}}, (Z_k)_{k \in V_I^c(0)})$ 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 $\{|V_I(0)| < \infty\}$,

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

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 $\{|V_I(0)| < \infty\}$,

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

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 49 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 49, 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).

$\square$

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

Corollary 51. *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 45

We are now ready to conclude the proof of Theorem 45. 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

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

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,

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

$$\sup_{A\in\sigma(((G^{(r)},\phi^{(r)}))} \mathbb{E}^0[\mathbf{1}_{E_r}\left(\mathbf{1}_A\mathbf{1}_{Z_0=+1} + \mathbf{1}_{A^\complement}\mathbf{1}_{Z_0=-1}\right)] + \mathbb{E}^0[\mathbf{1}_{E_r^\complement}\left(\mathbf{1}_A\mathbf{1}_{Z_0=+1} + \mathbf{1}_{A^\complement}\mathbf{1}_{Z_0=-1}\right)]$$

$$\leq \sup_{A\in\sigma(((G^{(r)},\phi^{(r)})))} \mathbb{E}^0[\mathbf{1}_{E_r}\left(\mathbf{1}_A\mathbf{1}_{Z_0=+1} + \mathbf{1}_{A^\complement}\mathbf{1}_{Z_0=-1}\right)] + \mathbb{P}^0[E_r^\complement]$$

$$\overset{(a)}{\leq} \sup_{A\in\sigma((G^{(r)},\phi^{(r)},\{\{U_{kl}\}_{l>k}\}_{k\in\mathbb{N}\cup\{0\}})} \mathbb{E}^0[\mathbf{1}_{E_r}\left(\mathbf{1}_A\mathbf{1}_{Z_0=+1} + \mathbf{1}_{A^\complement}\mathbf{1}_{Z_0=-1}\right)] + \mathbb{P}^0[E_r^\complement]$$

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

$$\overset{(c)}{=} \sup_{A\in\sigma(G,\phi,\{\{U_{kl}\}_{l>k}\}_{k\in\mathbb{N}\cup\{0\}},(Z_k)_{k\in V_I^\complement(0)})} \frac{1}{2}\mathbb{E}^0[\mathbf{1}_A\mathbf{1}_{E_r}] + \frac{1}{2}\mathbb{E}^0[\mathbf{1}_{A^\complement}\mathbf{1}_{E_r}] + \mathbb{P}^0[E_r^\complement]$$

$$= \frac{1}{2}\mathbb{P}^0[E_r] + \mathbb{P}^0[E_r^\complement].$$

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 51. 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.

$$\lim_{r\to\infty}\sup_{A\in\sigma((G^{(r)},\phi^{(r)}))} \mathbb{E}^0[\mathbf{1}_A\mathbf{1}_{Z_0=+1} + \mathbf{1}_{A^\complement}\mathbf{1}_{Z_0=-1}] \leq \lim_{r\to\infty}\frac{1}{2}\mathbb{P}^0[E_r] + \mathbb{P}^0[E_r^\complement]$$

$$= \frac{1}{2}\mathbb{P}^0[E] + \mathbb{P}^0[E^\complement]$$

$$= \frac{1}{2}(1 + \theta(H_{\lambda,f_{in}(\cdot)-f_{out}(\cdot),d})). \qquad (35)$$

The limit on the LHS exists from Proposition 4 and the limit on the RHS exists since E_r are non-decreasing events.

7 Identifiability of the Partition and Proof of Theorem 13

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 [13]). 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 52. *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 13.

Proof. First if $g(\cdot) \neq \frac{f_{in}(\cdot)+f_{out}(\cdot)}{2}$ Lebesgue almost everywhere, 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 $g(\cdot) = (f_{in}(\cdot) + f_{out}(\cdot)/2$ Lebesgue almost-everywhere. 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 [11], we get that

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

$$f_{out}(||x-y||)\left(\frac{f_{in}(||x||)f_{out}(||y||) + f_{out}(||x||)f_{in}(||y||)}{4} \right))\lambda^2 dx dy,$$

and

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

Now observe that

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

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.

$\square$

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 ([14],[5]).

8 The Exact-Recovery Problem

In this section, we provide a lower bound for the exact-recovery problem as stated in Definition 3. Recall that for the exact-recovery case, we equipped the set B_n with the torridal metric rather than the usual Euclidean metric. This is done mainly to simplify the presentation of the results. Nevertheless, one could establish identical results to the case when the set B_n equipped with the Euclidean metric albeit with significantly more heavier notation to handle the 'edge effects'. We note that for any $x := (x_1, \cdots, x_d), y := (y_1, \cdots, y_d) \in B_n$, the torroidal distance is given by $||x - y||_{\mathcal{T}_n} = ||(\min(|x_1 - y_1|, n^{1/d} - |x_1 - y_1|), \cdots, \min(|x_d - y_d|, n^{1/d} - |x_d - y_d|))||$, where $||.||$ is the Euclidean norm on $\mathbb{R}^d$. The key result we establish about this model is a lower bound or a necessary condition to perform Exact-Recovery. We then observe that the `GBG` algorithm presented earlier achieves Exact-Recovery, if the intensity λ is sufficiently high, thereby establishing a phase-transition. We establish the lower bound by exploiting recent advances in the understanding of error exponents for the probability of error in distinguishing between two Poisson random vectors developed in [4]. Furthermore, we also conjecture this necessary condition to also be sufficient. The exact proof of this conjecture is left open in this paper. Another important note is that since we view the set B_n as a torus in this section, the notion of Palm probability needs more care in stating. In the most general sense, we must employ the moment measure expansions of a Poisson Point Process on compact topological groups to discuss the Palm measure on the torus. We provide a justification of the moment measure equations we use in this section in the Appendix F.

8.1 Lower Bound for Exact-Recovery

In this section, we prove Theorem 15. To do so, we first give a general result in Proposition 53 using the Genie aided argument introduced in [2], and then subsequently prove an explicit formula for the lower bound stated in Theorem 15 using the large deviation results of [4] for hypothesis testing between Poisson random vectors. To state the result, we define a notion of *Flip-Bad* for nodes. Roughly speaking, we say a node $i \in [1, N_n]$ is Flip-Bad in G_n, if on flipping the community label of node i from Z_i to $-Z_i$, the likelihood of the observed data (G_n, ϕ_n) increases. More formally, given data $(\phi_n, G_n, (Z_i)_{i=1}^{N_n})$, the likelihood is defined as

$$\mathcal{L}(\phi_n, G_n, (Z_i)_{i=1}^{N_n}) := \frac{e^{-\lambda n}(\lambda n)^{N_n}}{N_n!} \left(\frac{1}{2n}\right)^{N_n} \prod_{1 \leq i < j \leq N_n : Z_i = Z_j, i \sim j} f_{in}^{(n)}(||X_i - X_j||)$$

$$\prod_{1 \leq i < j \leq N_n : Z_i \neq Z_j, i \sim j} f_{out}^{(n)}(||X_i - X_j||) \prod_{1 \leq i < j \leq N_n, Z_i = Z_j, i \nsim j} (1 - f_{in}^{(n)}(||X_i - X_j||))$$

$$\prod_{1 \leq i < j \leq N_n, Z_i \neq Z_j, i \nsim j} (1 - f_{out}^{(n)}(||X_i - X_j||)),$$

with the empty product being equal to 1. Thus, it is easy to see that for every n, $0 < \mathcal{L}((\phi_n, G_n, (Z_i)_{i=1}^{N_n}))$ ≤ 1 almost-surely, since for every n, N_n is finite almost surely. We say a node j is *Flip-Bad in G_n* if

$$\mathcal{L}((\phi_n, G_n, (Z_i)_{i=1}^{N_n})) \leq \mathcal{L}((\phi_n, G_n, (\tilde{Z}_i^{(j)})_{i=1}^{N_n})),$$

where $\tilde{Z}_i^{(j)} = Z_i$ if $i \neq j$ and $\tilde{Z}_j^{(j)} = -Z_j$. Thus, we use the term Flip-Bad, where a node is 'bad' if on flipping its community label, the observed data becomes more likely. We use this definition

to reason about the maximum-likelihood estimator. More formally, for each node $i \in [1, N_n]$, the Maximum-Likelihood (ML) estimate of the community label is denoted as $\hat{Z}_i$ and satisfies

$$
\begin{aligned}
(\hat{Z}_i)_{i \in [1, N_n]} &= \arg \max_{z \in \{-1, +1\}^{N_n}} \mathbb{P}[(Z_i)_{i \in [1, N_n]} = z | \phi_n, G_n] \\
&= \arg \max_{z \in \{-1, +1\}^{N_n}} \mathcal{L}((\phi_n, G_n, z))
\end{aligned}
\tag{37}
$$

In words, it is the optimal estimate for the community labels given the observed data ϕ_n and G_n, where optimality refers to the fact that it minimizes the probability that the vector is $(\hat{Z}_i)_{i \in [1, N_n]}$ is different from the ground truth $(Z_i)_{i \in [1, N_n]}$ among all possible estimators. More formally, the ML estimator $(\hat{Z}_i)_{i \in [1, N_n]}$ in Equation (37) satisfies

$$
\mathbb{P}[(\hat{Z}_i)_{i \in [1, N_n]} \neq (Z_i)_{i \in [1, N_n]}] = \inf_{(\tilde{Z}_i)_{i \in [1, N_n]} = \mathcal{A}(\phi_n, G_n)} \mathbb{P}[(\tilde{Z}_i)_{i \in [1, N_n]} \neq (Z_i)_{i \in [1, N_n]}],
$$

where the infimum is over all measurable functions $\mathcal{A}$ of the data ϕ_n and G_n. This asserts that the ML estimator for the community labels is the optimal estimator that minimizes probability of error. The following proposition gives a structural condition on the model parameters of G_n to identify a sufficient condition when the ML estimator will fail.

Proposition 53. *If the model parameters of G_n satisfies*

$$
\limsup_{n \to \infty} \frac{\int_{y \in B_n} \mathbb{E}^{0, y}[\mathbf{1}_{0 \text{ is Flip-Bad in } G_n \cup \{0, y\}} \mathbf{1}_{y \text{ is Flip-Bad in } G_n \in \cup \{0, y\}}] m_{n, d}(dy)}{n \mathbb{E}^0[\mathbf{1}_{0 \text{ is flip-bad in } G_n \cup \{0\}}]^2} \leq 1,
\tag{38}
$$

$$
\lim_{n \to \infty} n \mathbb{E}^0[\mathbf{1}_{0 \text{ is flip-bad in } G_n \cup \{0\}}] = \infty
\tag{39}
$$

where $m_{n, d}$ is the Haar measure on the torus B_n, then Exact-Recovery is not solvable.

The condition in Equation (38) states that the event of two 'far-away' nodes being Flip-Bad are asymptotically uncorrelated. Such statements are true for instance if the functions $f_{in}^{(n)}(\cdot)$ and $f_{out}^{(n)}(\cdot)$ have support that is $o(n)$. For example, if the connection functions satisfy $f_{in}^{(n)}(r) = a_n \mathbf{1}_{r \leq R_n}$ and $f_{out}^{(n)}(r) = b_n \mathbf{1}_{r \leq R_n}$, for some $0 \leq b_n \leq a_n \leq 1$ and $R_n = o(n^{1/d})$, then they satisfy the condition. See for example also Lemma 55. On the other hand, Equation (38) is satisfied even if the model is reduced to the classical SBM without geometry, i.e., any two nodes in G_n are connected only based on their community labels and not the location labels if the average degree is proportional to the logarithm of the population size.

Proof. If there exists a node in G_n that is Flip-Bad, then the Maximum Likelihood estimator for the community labels will not match the ground truth. We will show that if Equations (38) and (39) hold, then $\mathbb{P}[\exists i \in [1, N_n] : i \text{ is Flip-Bad in } G_n] = 1 - o_n(1)$, which will conclude the proof of the result.

To do this, define by $Y_n := \sum_{i \in [1, N_n]} \mathbf{1}_{i \text{ is Flip-Bad in } G_n}$. From the classical method of moments, it suffices to establish that $\limsup_{n \to \infty} \frac{\mathbb{E}[Y_n^2]}{\mathbb{E}[Y_n]^2} \leq 1$ and $\mathbb{E}[Y_n] \to \infty$ as $n \to \infty$. Indeed, if this were the case, then by Chebychev's inequality, we would have

$$
\mathbb{P}[Y_n = 0] \leq \frac{\mathbb{E}[Y_n^2]}{\mathbb{E}[Y_n]^2} - 1,
$$

which will converge to 0. It remains to compute the first and second moment of Y_n.

From Campbell's theorem (see Appendix F for Palm probability on the torus), we have

$$\mathbb{E}[Y_n] = \lambda \int_{y \in B_n} \mathbb{E}^y[\mathbf{1}_{y \text{ is Flip-Bad in } G_n \cup \{y\}}] m_{d,n}(dy)$$
$$= \lambda n \mathbb{E}^0[\mathbf{1}_{0 \text{ is Flip-bad in } G_n \cup \{0\}}], \tag{40}$$

where the second inequality follows from the symmetry in the torus. Thus from the hypothesis of the theorem in Equation (39), $\mathbb{E}[Y_n]$ converges to ∞. To compute the second moment, we use the factorial moment expansion of the Poisson Process as follows.

$$\mathbb{E}[Y_n^2] = \mathbb{E}[(\sum_{i \in [1,N_n]} \mathbf{1}_{i \text{ is Flip-Bad in } G_n})^2]$$

$$= \mathbb{E}[Y_n] + \mathbb{E}[\sum_{i \neq j} \mathbf{1}_{i \text{ is Flip-Bad in } G_n} \mathbf{1}_{j \text{ is Flip-Bad in } G_n}]$$

$$\overset{(a)}{=} \mathbb{E}[Y_n] + \lambda^2 \int_{x \in B_n} \int_{y \in B_n} \mathbb{E}^{x,y}[\mathbf{1}_{x \text{ is Flip-Bad in } G_n \cup \{x,y\}} \mathbf{1}_{y \text{ is Flip-Bad in } G_n \in \cup \{x,y\}}] m_{n,d}(dx) m_{n,d}(dy)$$

$$\overset{(b)}{=} \mathbb{E}[Y_n] + \lambda^2 n \int_{y \in B_n} \mathbb{E}^{0,y}[\mathbf{1}_{0 \text{ is Flip-Bad in } G_n \cup \{0,y\}} \mathbf{1}_{y \text{ is Flip-Bad in } G_n \in \cup \{0,y\}}] m_{n,d}(dy), \tag{41}$$

where equality (a) follows from the 2nd order Moment Measure expansion of a Poisson process and (b) follows from the symmetry in the torus. Thus, from Equations (40) and (41), we have,

$$\frac{\mathbb{E}[Y_n^2]}{\mathbb{E}[Y_n]^2} = \frac{1}{\mathbb{E}[Y_n]} + \frac{\lambda^2 \int_{y \in B_n} \mathbb{E}^{0,y}[\mathbf{1}_{0 \text{ is Flip-Bad in } G_n \cup \{0,y\}} \mathbf{1}_{y \text{ is Flip-Bad in } G_n \in \cup \{0,y\}}] m_{n,d}(dy)}{(\lambda n \mathbb{E}^0[\mathbf{1}_{0 \text{ is Flip-bad in } G_n \cup \{0\}}])^2}. \tag{42}$$

Thanks to the assumption on the connection functions in Equation (38) and the first moment in Equation (42), we have that $\limsup_{n \to \infty} \frac{\mathbb{E}[Y_n^2]}{\mathbb{E}[Y_n]} \leq 1$. $\qquad\square$

In the rest of the section, we consider the model given in Definition 14 where the connection functions $f_{in}^{(n)}(\cdot)$ and $f_{out}^{(n)}(\cdot)$ take the form $f_{in}(r) = a\mathbf{1}_{r \leq \log(n)^{1/d}}$ and $f_{out}^{(n)}(r) = b\mathbf{1}_{r \leq \log(n)^{1/d}}$ for some $0 \leq b < a \leq 1$, to illustrate how one can use the previous proposition to obtain a closed form expression for the phase-transition threshold. We first prove Theorem 15 to provide an exact necessary condition in terms of the model parameters λ, a and b for achieving exact-recovery. The proof of Theorem 15 follows from the next two lemmas. Recall that, we denote by ν_d for all $d \in \mathbb{N}$ as the volume of the unit Euclidean ball in d dimensions.

Lemma 54. *For all $\lambda > 0$, $d \in \mathbb{N}$ and $0 \leq b < a \leq 1$, if $G_n \sim \mathcal{G}(\lambda n, a, b, d)$, then*
$\mathbb{E}^0[\mathbf{1}_{0 \text{ is Flip-Bad in } G_n \cup \{0\}}] = e^{-\lambda \nu_d \log(n)(1 - \sqrt{ab} - \sqrt{(1-a)(1-b)} - o(1))}$.

Lemma 55. *For all $\lambda > 0$, $d \in \mathbb{N}$ and $0 \leq b < a \leq 1$ such that $\lambda \nu_d \log(n)(1 - \sqrt{ab} - \sqrt{(1-a)(1-b)}) < 1$, the graph $G_n \sim \mathcal{G}(\lambda n, a, b, d)$ satisfies Equation (38).*

Proof. of Theorem 15.
From Lemma 55, we know that the connection functions satisfy the model assumption in Equation (38). From Lemma 54, we know that if $\lambda \nu_d(1 - \sqrt{ab} - \sqrt{(1-a)(1-b)}) < 1$, then $\mathbb{E}^0[\mathbf{1}_{0 \text{ is flip-bad in } G_n \cup \{0\}}] = n^{-1+\delta-o(1)}$ for some $\delta > 0$. Thus, $n\mathbb{E}^0[\mathbf{1}_{0 \text{ is flip-bad in } G_n \cup \{0\}}] = n^{\delta-o(1)}$ which converges to ∞ as n goes to ∞. The proof is now complete thanks to Proposition 53. $\qquad\square$

We now provide the proofs of the associated lemmas.

Proof. of Lemma 54.
This lemma is a corollary of Lemma 11 proven in [4], where the error exponent for hypothesis testing between Poisson random vectors was established. To assess whether the node at 0 is Flip-Bad, we need to decide given the locations and true community labels of the neighbors and non-neighbors of node at 0, whether it belongs to community $+1$ or -1. However, since the connection functions have support of $\log(n)^{1/d}$ and events in disjoint regions of space are independent, it suffices to consider the neighbors and non-neighbors within the ball of radius $\log(n)^{1/d}$ around 0. (Note from the model that there are no neighbors of the node at 0 at a distance larger than $\log(n)^{1/d}$). The number of neighbors of nodes in the same commuity as 0 is a Poisson random variable of mean $\lambda/2\nu_d a \log(n)$ and the number of neighbors in the opposite community is another independent Poisson random variable of mean $\lambda/2\nu_d b \log(n)$. The independence follows from elementary independent thinning property of the PPP. Similarly, the number of non-neighbors in the same and opposite community as 0 and within a distance of $\log(n)^{1/d}$ of 0 are independent Poisson random variables of mean $\lambda/2\nu_d(1-a)\log(n)$ and $\lambda/2\nu_d(1-b)\log(n)$ respectively. This argument again follows from the independent thinning property of an independently marked Poisson Process.

Thus the probability of a node at 0 being flip-bad is equal to the error made by an optimal hypothesis tester between two random vectors $(\lambda/2)\nu_d \log(n)(a, b, 1-a, 1-b)$ and $(\lambda/2)\nu_d \log(n)(b, a, 1-b, 1-a)$, given an uniform prior over the two models. Furthermore, the components of the observed random vector are independent. Thus, we are in a setting to apply the CH-divergence theorem of [4] to characterize the error probability of the optimal hypothesis testing error. More precisely, applying Lemma 11 from [4] the error probability in identifying between the two Poisson random vectors with uniform prior $(\lambda/2)\nu_d \log(n)\mu$ and $(\lambda/2)\nu_d \log(n)\nu$ where the vectors $\mu := (a, b, 1-a, 1-b)$ and $\nu := (b, a, 1-b, 1-a)$ is equal to $\mathbb{E}^0[\mathbf{1}_{0 \text{ is flip-bad in } G_n \cup \{0\}}]$ is given by

$$\mathbb{E}^0[\mathbf{1}_{0 \text{ is flip-bad in } G_n \cup \{0\}}] := n^{-(\lambda/2)\nu_d D_+(\mu,\nu)+o(1)}, \tag{43}$$

where the CH-Divergence $D_+(\mu, \nu)$ ([4]) is given by

$$D_+(\mu, \nu) := \max_{t \in [0,1]} \sum_{x \in \mathcal{X}} (t\mu(x) + (1-t)\nu(x) - \mu(x)^t \nu(x)^{1-t}). \tag{44}$$

Here $\mathcal{X} := [1, 2, 3, 4]$, and $\mu(x)$ ($\nu(x)$) for $x \in \mathcal{X}$ refers to the xth component of the vector μ (ν). Evaluating Equation (44) yields that $D_+(\mu, \nu) = 1 - \sqrt{ab} - \sqrt{(1-a)(1-b)}$ with the maximum being achieved at $t = 1/2$ due to symmetry in the vectors μ and ν. Substituting Equation (44) into (43) yields the result. $\square$

Proof. of Lemma 55.

This lemma follows from some straightforward calculations exploiting the spatial independence

across the Poisson process. To verify Equation (38), consider the following chain of equations.

$$\int_{y\in B_n} \mathbb{E}^{0,y}[\mathbf{1}_{0 \text{ is Flip-Bad in } G_n\cup\{0,y\}}\mathbf{1}_{y \text{ is Flip-Bad in } G_n\in\cup\{0,y\}}]m_{n,d}(dy) =$$

$$\int_{y\in B(0,2\log(n)^{1/d})} \mathbb{E}^{0,y}[\mathbf{1}_{0 \text{ is Flip-Bad in } G_n\cup\{0,y\}}\mathbf{1}_{y \text{ is Flip-Bad in } G_n\in\cup\{0,y\}}]m_{n,d}(dy)$$

$$+ \int_{y\in B_n\cap B(0,2\log(n)^{1/d})^\complement} \mathbb{E}^{0,y}[\mathbf{1}_{0 \text{ is Flip-Bad in } G_n\cup\{0,y\}}\mathbf{1}_{y \text{ is Flip-Bad in } G_n\in\cup\{0,y\}}]m_{n,d}(dy)$$

$$\leq \int_{y\in B(0,2\log(n)^{1/d})} \mathbb{E}^{0,y}[\mathbf{1}_{0 \text{ is Flip-Bad in } G_n\cup\{0,y\}}]m_{n,d}(dy)$$

$$+ \int_{y\in B_n\cap B(0,2\log(n)^{1/d})^\complement} \mathbb{E}^{0,y}[\mathbf{1}_{0 \text{ is Flip-Bad in } G_n\cup\{0,y\}}\mathbf{1}_{y \text{ is Flip-Bad in } G_n\in\cup\{0,y\}}]m_{n,d}(dy) \qquad (45)$$

The key observation to make is that for x,y such that $||x-y|| > 2\log(n)^{1/d}$, we have

$$\mathbb{E}^{x,y}[\mathbf{1}_{x \text{ is Flip-Bad in } G_n\cup\{x,y\}}\mathbf{1}_{y \text{ is Flip-Bad in } G_n\in\cup\{x,y\}}] = \mathbb{E}^x[\mathbf{1}_{x \text{ is Flip-Bad in } G_n\cup\{x\}}]\mathbb{E}^y[\mathbf{1}_{y \text{ is Flip-Bad in } G_n\cup\{y\}}]$$
$$(46)$$

This follows since in an independently marked Poisson Process, events on disjoint sets are independent. Further more, from the symmetry in the torus, for all $x \in B_n$, we also have

$$\mathbb{E}^x[\mathbf{1}_{x \text{ is Flip-Bad in } G_n\cup\{x\}}] = \mathbb{E}^0[\mathbf{1}_{0 \text{ is Flip-Bad in } G_n\cup\{0\}}]. \qquad (47)$$

Thus, we get from Equations (45),(46) and (47) that

$$\int_{y\in B_n} \mathbb{E}^{0,y}[\mathbf{1}_{0 \text{ is Flip-Bad in } G_n\cup\{0,y\}}\mathbf{1}_{y \text{ is Flip-Bad in } G_n\in\cup\{0,y\}}]m_{n,d}(dy) \leq$$

$$(n - 2^d\nu_d\log(n))\mathbb{E}^0[\mathbf{1}_{0 \text{ is Flip-Bad in } G_n\cup\{0\}}]^2 + 2^d\nu_d\log(n)\mathbb{E}^0[\mathbf{1}_{0 \text{ is Flip-Bad in } G_n\cup\{0\}}]. \qquad (48)$$

Thus, since $\lambda\nu_d\log(n)(1 - \sqrt{ab} - \sqrt{(1-a)(1-b)}) > 1$, we have from Lemma 54 that, $n\mathbb{E}^0[\mathbf{1}_{0 \text{ is Flip-Bad in } G_n\cup\{0\}}] = n^\gamma$ for some $\gamma > 0$. Hence Equation (48) implies Equation (38) if $\lambda\nu_d\log(n)(1 - \sqrt{ab} - \sqrt{(1-a)(1-b)}) > 1$.

$\square$

8.2 Upper Bound for Exact-Recovery - Proof of Theorem 17

In the present paper, we are only able to establish the presence of the phase transition by proving Theorem 17. We believe a 'two-round' information theoretic argument can be employed to prove this result. A possible strategy is to first show that for any $0 \leq b < a \leq 1$, a 'large' (i.e. all but $o(n)$) nodes will be correctly classified by the ML estimator with high probability. Further, if the parameters satisfied $\lambda\nu_d(1 - \sqrt{ab} - \sqrt{(1-a)(1-b)}) > 1$, then all nodes will be correctly classified with high probability. One can possibly make this efficient by means of 'sample splitting' arguments of [4]. One can sub-sample the edges so that the graph is almost sparse so that a large fraction of the nodes can be correctly labelled by the ML estimator. Then, we can 'clean-up', i.e. estimate a corrected community label estimate using the edges not used in the first round. This conjecture is also reminiscent of the 'local to global' phenomena that occurs in many random graph

models ([2],[36],[8],[31]), where an obvious local necessary condition also turns out to be sufficient. However, as a corollary to the GBG algorithm introduced above, we Theorem 17 which establsihes that Exact-Recovery can be solved if the intensity λ is sufficiently high.

Proof. of Theorem 17
Notice that with $R = \log(n)^{1/d}/2d$, there are at-most n grid-cells in B_n. If we show that the chance that a grid cell is T-BAD is $n^{-1-\delta}$ for some $\delta > 0$, then by an union bound argument, we can assert that with probability at-least $1 - n^{-\delta}$, all grid-cells will be T-GOOD. This follows since there are at-most $\lceil 4^d dn/\log(n) \rceil$ number of grid cels. If all grid cells are T-Good, then by Proposition 32, all nodes in G_n will be correctly partitioned by the algorithm.

Set $R = \log(n)^{1/d}/2d$ and $\epsilon \in (0,1)$ arbitrary as parameters of the GBG algorithm. From Proposition 34, we know that for any $z, z^{'} \in \mathbb{Z}^d$ such that $||z - z^{'}||_{\infty} = 1$, the probability that the pairwise classifier makes an error is at-most $e^{-c(d,a,b)\lambda \log(n)}$, where $c(d,a,b) > 0$ is a positive constant. From Lemma 35, it is clear, that the chance that a cell is T-BAD is at-most $n^{-c(d,a,b)\lambda} \log(n)^d C + n^{-c^{'}(\epsilon)\lambda n}$, where c and $c^{'}$ are two constants. Thus, by choosing λ sufficiently high, we can ensure that the probability of a grid-cell being T-BAD is at-most $n^{-1-\delta}$ for some $\delta > 0$. From an union bound, with probability at-least $1 - n^{-\delta}$, all nodes will be T-GOOD, and hence Exact-Recovery is achieved thanks to Proposition 32. □

Here we also want to remark that using ideas from Section 4.5, the GBG algorithm can be implemented without knowledge of the parameters. In particular, one can use standard spectral or SDP algorithms to decide whether a cell is T-GOOD or not, as explained in Section 4.5. This can be done so since the sub-graph of G_n restricted to nodes within the 1 thickening of any grid cell z is dsitributed as the SBM with random $\log(n)$ number of nodes and having a constant connection probability of either a or b among nodes of the same or opposite communities. Thus, one can employ standard spectral or SDP algorithms 'off the shelf' to implement the algorithm without knowledge of a, b or λ.

9 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. We studied this problem in both the sparse and non-sparse regime. Our main technical contributions in the sparse graph case 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. In the Exact-Recovery setting, we give a lower bound that we conjecture to be tight. However, this is just a first step and there are a plenty of open questions just concerning the model we introduced.

1) Is Weak-Recovery and Information Flow from Infinity equivalent ? - In this paper, we proved that weak-recovery 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 weak-recovery monotone non-decreasing in λ ? - We saw in the proof of Proposition 42 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) Can one resolve Conjecture 16 to identify the critical phase-transition point for Exact-Recovery - This conjecture is reminiscent of the local to global phenomenon consistently observed in various random graph models ([1], [39],[36]). In these settings, an obvious local condition, i.e., there being no flip-bad node, also turns out to be sufficient for exact-recovery. Establishing such a result in our model will help us obtain a better understanding of our random graph model and also may aid in improved algorithms for practical situations.

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 $f_{in}^{(n)}(\cdot)$ within communities and $f_{out}^{(n)}(\cdot)$ across communities, a simple adaptation of our algorithm can give a sufficient condition in both the sparse and logarithmic degree regime, although will be sub-optimal. Our lower bound technique in the sparse regime can also be applied in the setting of many communities (see Theorem 2 in [3] 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 9 where the phase transition for weak-recovery 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. As far as the logarithmic degree regime, our lower bound framework can easily extend using the same large-deviations result for Poisson hypothesis testing developed in [4]. However, an algorithm achieving this threshold in the logarithmic regime with multiple communities is not yet known.

4) Characterization of the Phase-Transition for Weak-Recovery - 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 47. 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 aspect concerns the possible gaps between information versus computation thresholds. 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

References

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

[2] Emmanuel Abbe, Afonso S Bandeira, and Georgina Hall. Exact recovery in the stochastic block model. *arXiv preprint arXiv:1405.3267*, 2014.

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

[4] Emmanuel Abbe and Colin Sandon. Community detection in general stochastic block models: Fundamental limits and efficient algorithms for recovery. In *Foundations of Computer Science (FOCS), 2015 IEEE 56th Annual Symposium on*, pages 670–688. IEEE, 2015.

[5] 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.

[6] 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.

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

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

[9] 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.

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

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

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

[13] 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.

[14] 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.

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

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

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

[18] Sainyam Galhotra, Arya Mazumdar, Soumyabrata Pal, and Barna Saha. The geometric block model. *arXiv preprint arXiv:1709.05510*, 2017.

[19] 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.

[20] 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.

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

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

[23] Günter Last. *Modern Random Measures: Palm Theory and Related Modelsl*. Univ. Karlsruhe, Fak. für Mathematik, 2008.

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

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

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

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

[28] 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.

[29] 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.

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

[31] Ronald Meester, Mathew D Penrose, and Anish Sarkar. The random connection model in high dimensions. *Statistics & probability letters*, 35(2):145–153, 1997.

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

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

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

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

[36] Elchanan Mossel, Joe Neeman, and Allan Sly. Consistency thresholds for the planted bisection model. In *Proceedings of the forty-seventh annual ACM symposium on Theory of computing*, pages 69–75. ACM, 2015.

[37] 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.

[38] 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.

[39] Mathew Penrose. *Random geometric graphs*. Number 5. Oxford University Press, 2003.

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

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

[42] 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.

[43] 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.

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

[45] 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.

[46] Jeffrey Travers and Stanley Milgram. The small world problem. *Phychology Today*, 1(1):61–67, 1967.

[47] 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.

A   Proof of Proposition 47

The goal of this section is to establish Proposition 47 in which the equality in Theorem 45 cannot be achieved for any $\lambda > 0$. The functions $f_{in}(\cdot)$ and $f_{out}(\cdot)$ of Proposition 47 satisfy $\int_{r \geq 0}(f_{in}(r) - f_{out}(r))r\,dr = \infty$. Thus, the bound from Theorem 45 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 [45] 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 [45] after being adapted to our setting, states the following.

Lemma 56. *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 57. *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 56 for our example.

Notice that the chosen example satisfies the following.

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

$$\int_{x \in \mathbb{R}^2} \left(1 - \sqrt{\frac{1 - f_{in}(||x||)}{1 - f_{out}(||x||)}}\right)^2 dx < \infty \tag{49}$$

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 (49) and Lemma 56, 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   Proof of Proposition 32

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.

$\square$

C Proof of Proposition 39

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 35, 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.

D   Proof of Lemma 49

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 48. 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 48, 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 48. 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 48. 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 48.

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 48. 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 48.

$\square$

E   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.

F   Palm Measure on the Torus

In this section, we recap the basic properties of the Poisson Point Process on the torus. As the torus is a locally compact unimodular topological group, the definition of Palm measures and the moment measure equations of PPP on a torus follows directly from the theory described in [23]. We reproduce the key results needed for our paper here. Fix a $n \in \mathbb{N}$ and consider the d dimensional torus on the set $B_n := \left[-\frac{n^{1/d}}{2}, \frac{n^{1/d}}{2}\right]^d$ equipped with the Haar measure denoted as $m_{n,d}(\cdot)$ which is invariant under translations on the torus. Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space on which we have a stationary independently marked PPP on the torus B_n of intensity $\lambda > 0$ with marks in an arbitrary Polish space $\mathcal{K}$. Denote by the atoms of this point process by $\phi_n := \{X_1, \cdots, X_{N_n}\}$ enumerated in an arbitrary manner and the corresponding marks as $\{K_1, \cdots, K_{N_n}\}$. From the definition of the PPP, N_n is a Poisson random variable of mean λn independent of everything else and conditional on N_n, $(X_i)_{i=1}^{N_n}$ are i.i.d. random variables that are uniformly distributed in the set B_n. Conditionally on N_n and $(X_i)_{i=1}^{N_n}$, the sequence $(K_i)_{i=1}^{N_n}$ are i.i.d. In this framework, for any $k \in \mathbb{N}$ and $x_1, \cdots, x_k \in B_n$, the Palm measure $\mathbb{P}^{x_1, \cdots, x_k}$ corresponds to adding fictitious atoms at locations $\{x_1, \cdots, x_k\}$ and equipping them with independent marks having the same law as K_1. In other words, thanks to Slivnyak's theorem ([23]), the atoms of the PPP under the Palm measure $\mathbb{P}^{x_1, \cdots, x_k}$ is $\phi_n \cup \{x_1, \cdots x_k\}$ where ϕ_n is the law of the point process under $\mathbb{P}$, with the marks of these additional points having the same distribution as that of K_1 and independent of everything else. For this framework, for any function $f(\cdot) : B_n \to \mathbb{R}_+$ such that $\int_{x \in B_n} f(x) m_{n,d}(dx) < \infty$ we have the following version of Campbell's theorem ([23]) -

$$\mathbb{E}[\sum_{x \in \phi_n} f(x)] = \lambda \int_{x \in B_n} \mathbb{E}^x[f(x)] m_{d,n}(dx). \tag{50}$$

More generally, we have the following k-th order moment measure expansions of the Poisson process. Let $f(\cdot) : B_n^k \to \mathbb{R}_+$ be such that $\int_{x_1, \cdots, x_k} f(x_1, \cdots, x_k) \prod_{i=1}^k m_{n,d}(dx_i) < \infty$. Then, we have

$$\mathbb{E}[\sum_{\substack{x_1, \cdots, x_k \in \phi_n \\ \neq}} f(x_1, \cdots, x_k)] = \lambda^k \int_{x_1, \cdots, x_k \in B_n} \mathbb{E}^{x_1, \cdots, x_k}[f(x_1, \cdots, x_k)] \prod_{i=1}^k m_{n,d}(dx_i). \tag{51}$$

We use Equations (50) and (51) in Section 8 to compute the first and second moments of the number of Flip-Bad nodes in G_n.