Connectivity of Random Annulus Graphs and the Geometric Block Model

Sainyam Galhotra
University of Massachusetts Amherst, USA
sainyam@cs.umass.edu

Arya Mazumdar
University of Massachusetts Amherst, USA
arya@cs.umass.edu

Soumyabrata Pal
University of Massachusetts Amherst, USA
spal@cs.umass.edu

Barna Saha
University of California, Berkeley, USA
barnas@berkeley.edu

Abstract

Random geometric graph (Gilbert, 1961) is a basic model of random graphs for spatial networks proposed shortly after the introduction of the Erdős-Rényi random graphs. The *geometric block model* (GBM) is a probabilistic model for community detection defined over random geometric graphs (RGG) similar in spirit to the popular *stochastic block model* which is defined over Erdős-Rényi random graphs. The GBM naturally inherits many desirable properties of RGGs such as transitivity ("friends having common friends') and has been shown to model many real-world networks better than the stochastic block model. Analyzing the properties of a GBM requires new tools and perspectives to handle correlation in edge formation. In this paper, we study the necessary and sufficient conditions for community recovery over GBM in the connectivity regime. We provide efficient algorithms that recover the communities exactly with high probability and match the lower bound within a small constant factor. This requires us to prove new connectivity results for *vertex-random graphs* or *random annulus graphs* which are natural generalizations of random geometric graphs.

A vertex-random graph is a model of random graphs where the randomness lies in the vertices as opposed to an Erdős-Rényi random graph where the randomness lies in the edges. A vertex-random graph $G(n, [r_1, r_2]), 0 \leq r_1 < r_2 \leq 1$ with n vertices is defined by assigning a real number in $[0, 1]$ randomly and uniformly to each vertices and adding an edge between two vertices if the "distance" between the corresponding two random numbers is between r_1 and r_2. For the special case of $r_1 = 0$, this corresponds to random geometric graph in one dimension. We can extend this model naturally to higher dimensions; these higher dimensional counterparts are referred to as *random annulus graphs*. Random annulus graphs appear naturally whenever the well-known Goldilocks principle ("not too close, not too far') holds in a network. In this paper, we study the connectivity properties of such graphs, providing both necessary and sufficient conditions. We show a surprising *long edge phenomena* for vertex-random graphs: the minimum gap for connectivity between r_1 and r_2 is significantly less when $r_1 > 0$ vs when $r_1 = 0$ (RGG). We then extend the connectivity results to high dimensions. These results play a crucial role in analyzing the GBM.

2012 ACM Subject Classification Mathematics of computing → Random graphs

Keywords and phrases random graphs, geometric graphs, community detection, block model

Digital Object Identifier 10.4230/LIPIcs.APPROX-RANDOM.2019.53

Category RANDOM

Related Version A full version of the paper is available at [16], https://arxiv.org/abs/1804.05013.

Funding S. Galhotra and B. Saha are supported in part by NSF 1652303, a Google award and a Sloan fellowship. A. Mazumdar and S. Pal are supported in part by NSF 1642658 and 1642550.

Approximation, Randomization, and Combinatorial Optimization. Algorithms and Techniques (APPROX/RANDOM 2019).
Editors: Dimitris Achlioptas and László A. Végh; Article No. 53; pp. 53:1–53:23
Leibniz International Proceedings in Informatics
Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

1 Introduction

Models of random graphs are ubiquitous with Erdős-Rényi graphs [12, 17] at the forefront. Studies of the properties of random graphs have led to many fundamental theoretical observations as well as many engineering applications. In an Erdős-Rényi graph $G(n, p), n \in \mathbb{Z}_+, p \in [0, 1]$, the randomness lies in how the edges are chosen: each possible pair of vertices forms an edge independently with probability p. It is also possible to consider models of graphs where randomness lies in the vertices.

Keeping up with the simplicity of the Erdős-Rényi model, one can define a vertex-random graph (VRG) in the following way. Given two reals $0 \le r_1 \le r_2 \le 1/2$, the vertex-random graph $\mathrm{VRG}(n, [r_1, r_2])$ is a random graph with n vertices. Each vertex u is assigned a random point X_u selected uniformly from the circumference of a circle of perimeter 1. Two vertices u and v are connected by an edge, if and only if the distance of the corresponding points on the circle (the geodesic distance) is between r_1 and r_2. This definition is by no means new. For the case of $r_1 = 0$, this is the random geometric graphs (RGG) in one dimension. Random Geometric graphs were defined first by [18] and constitute the first and simplest model of spatial networks. Since then, they have found wide-spread applications in modeling wireless (ad-hoc) communication networks [9, 19], information propagation in social networks [13, 31] etc., and have been studied extensively [4, 5, 6]. The definition of VRG has been previously mentioned in [9]. The interval $[r_1, r_2]$ is called the connectivity interval in VRGs.

Vertex random graphs inherit many desirable properties of RGGs such as vertices with high modularity and the degree associativity property (high degree nodes tend to connect), which in turn led to the popularity of RGGs [13, 31]. In addition, VRGs naturally arise whenever the Goldilocks principle (not too close, not too far) is applicable in networks. For example, in a co-purchase network, a person who bought a bike may buy similar products like a helmet along with it, but not another bike [15]. Understanding connectivity properties of VRGs can shed light in co-purchasing behavior and product recommendation. Interestingly, the connectivity properties of VRGs turn out to be crucial to develop and analyze community detection algorithms for the *geometric block model* [15].

Connectivity of Vertex Random Graph (VRG). Threshold properties of Erdős-Rényi graphs have been at the center of much interest, and in particular it is known that many graph properties exhibit sharp phase transition phenomena [14]. Random geometric graphs also exhibit similar threshold properties [26]. Our first contribution in this work is to identify such connectivity threshold for VRGs. Consider a $\mathrm{VRG}(n, [0, r])$ defined above with $r = \frac{a \ln n}{n}$. It is known that $\mathrm{VRG}(n, [0, r])$ is connected with high probability if and only if $a > 1$ (I.e., $\mathrm{VRG}(n, [0, \frac{(1+\epsilon)\ln n}{n}])$ is connected for any $\epsilon > 0$. We will ignore this ϵ and just mention connectivity threshold as $\frac{\ln n}{n}$). Now let us consider the graph $\mathrm{VRG}(n, [\frac{b \ln n}{n}, \frac{\ln n}{n}])$, $b > 0$. Clearly this graph has less edges than $\mathrm{VRG}(n, [0, \frac{\ln n}{n}])$. **Is this graph still connected?** Surprisingly, we show that the modified graph remains connected as long as $b \le 0.5$. Therefore, $\mathrm{VRG}(n, [\frac{0.5 \ln n}{n}, \frac{\ln n}{n}])$ is connected, but $\mathrm{VRG}(n, [0, \frac{(1-\epsilon)\ln n}{n}])$ is not $\forall \epsilon > 0$.

Can we explain this striking shift in connectivity interval, when one goes from $b = 0$ to $b > 0$? Note that the $\mathrm{VRG}(n, [\frac{0.50 \ln n}{n}, \frac{\ln n}{n}])$ is obtained from the $\mathrm{VRG}(n, [0, \frac{\ln n}{n}])$ by deleting all “short-distance” edges. It turns out the “long-distance” edges are sufficient to maintain connectivity, because they can connect points over multiple hops in the graph. Another possible explanation is that connectivity threshold for VRG is not dictated by isolated nodes as is the case in Erdős-Rényi graphs. Thus, after the connectivity threshold has been achieved, removing short edges still retains connectivity.

The Geometric Block Model. We are motivated to study the threshold phenomena of vertex-random graphs, because it appears naturally in the analysis of the geometric block model (GBM) [15]. The geometric block model is a probabilistic generative model of communities and is a spatial analogue to the popular stochastic block model (SBM) [22, 10, 8, 2, 1, 20, 7, 24]. The SBM generalizes the Erdős-Rényi graphs in the following way. Consider a graph $G(V, E)$, where $V = V_1 \sqcup V_2 \sqcup \cdots \sqcup V_k$ is a disjoint union of k clusters denoted by $V_1, \ldots, V_k$. The edges of the graph are drawn randomly: there is an edge between $u \in V_i$ and $v \in V_j$ with probability $q_{i,j}, 1 \le i, j \le k$. Given the adjacency matrix of such a graph, the task is to find the partition $V_1 \sqcup V_2 \sqcup \cdots \sqcup V_k$ of V.

This model has been incredibly popular both in theoretical and practical domains of community detection. Recent theoretical works focus on characterizing sharp threshold of recovering the partition in the SBM. For example, when there are only two communities of exactly equal sizes, and the inter-cluster edge probability is $\frac{b \ln n}{n}$ and intra-cluster edge probability is $\frac{a \ln n}{n}$, it is known that exact recovery is possible if and only if $\sqrt{a} - \sqrt{b} > \sqrt{2}$ [1, 24]. The regime of the probabilities being $\Theta\left(\frac{\ln n}{n}\right)$ has been put forward as one of most interesting ones, because in an Erdős-Rényi random graph, this is the threshold for graph connectivity [4]. Note that the results are not only of theoretical interest, many real-world networks exhibit a "sparsely connected" community feature [23], and any efficient recovery algorithm for sparse SBM has many potential applications.

While the SBM is a popular model (because of its apparent simplicity), there are many aspects of real social networks, such as "transitivity rule" ("friends having common friends') and other community structures that are not accounted for in SBM. Defining a block model over a random geometric graph, the geometric block model (GBM), circumvents this since GBMs naturally inherit the transitivity property of random geometric graphs. In a previous work [15], we showed GBMs model community structures better than an SBM in many real world networks (e.g. DBLP collaboration network, Amazon co-purchase network etc.). The GBM depends on the basic definition of the random geometric graph in the same way the SBM depends on Erdős-Rényi graphs. The two-cluster GBM with vertex set $V = V_1 \sqcup V_2$, $V_1 = V_2$ is a random graph defined in the following way. Suppose, $0 \le r_d < r_s \le 1/2$ be two real numbers. For each vertex $u \in V$ randomly and independently choose a point X_u from the circumference of a circle of unit perimeter. There will be an edge between u and v if and only if,

$$d_L(X_u, X_v) \le r_s \text{ when } u, v \in V_1 \text{ or } u, v \in V_2$$
$$d_L(X_u, X_v) \le r_d \text{ when } u \in V_1, v \in V_2 \text{ or } u \in V_2, v \in V_1,$$

where d_L denotes the geodesic distance. Let us denote this random graph as $\text{GBM}(r_s, r_d)$. Given this graph $\text{GBM}(r_s, r_d)$, the main problem of community detection is to recover the partition (i.e., V_1 and V_2). The GBM provides a systematic way to introduce correlation during edge formation, an important aspect in real networks that often renders a problem theoretically intractable. The tool set needed to recover communities under a GBM thus differs significantly than what has been used to analyze the SBM.

Motivated by the SBM literature, we here also look at the GBM in the connectivity regime, i.e., when $r_s = \frac{a \ln n}{n}, r_d = \frac{b \ln n}{n}$. Our first contribution in this part is to provide a lower bound that shows that it is impossible to recover the parts from $\text{GBM}(\frac{a \ln n}{n}, \frac{b \ln n}{n})$ when $a - b < 1/2$. No lower bound for recovery was known before. We also derive a relation between a and b that defines a sufficient condition of recovery in $\text{GBM}(\frac{a \ln n}{n}, \frac{b \ln n}{n})$, closely matching the lower bound. The analysis crucially exploits the connectivity properties of vertex-random graphs.

It is possible to generalize the GBM to include different distributions, different metric spaces and multiple parts. It is also possible to construct other type of spatial block models such as the one very recently being put forward in [28] which rely on the random dot product graphs [30]. In [28], edges are drawn between vertices randomly and independently as a function of the distance between the corresponding vertex random variables. In contrast, in GBM edges are drawn deterministically given the vertex random variables, and edges are dependent unconditionally. Moreover [28] only considers the recovery scenario where in addition to the graph, values of the vertex random variables are provided. Note that in GBM, we only observe the graph. In particular, it will be later clear that if we are given the corresponding random variables (locations) to the vertices in addition to the graph, then recovery of the partitions in $\mathrm{GBM}(\frac{a \ln n}{n}, \frac{b \ln n}{n})$ is possible if and only if $a - b > 0.5$ and $a > 1$, that is we can identify the recovery threshold exactly.

VRG in Higher Dimension: The Random Annulus Graphs. It is natural to ask similar question of connectivity for VRGs in higher dimension. In a VRG at dimension t, we may assign t-dimensional random vectors to each of the vertices, and use a standard metric such as the Euclidean distance to decide whether there should be an edge between two vertices. Formally, let us define the t-dimensional sphere as $S^t \equiv \{x \in \mathbb{R}^{t+1} \mid ||x||_2 = 1\}$. Given two reals $0 \le r_1 \le r_2 \le 2$, the random annulus graph $\mathrm{RAG}_t(n, [r_1, r_2])$ is a random graph with n vertices. Each vertex u is assigned a random vector X_u selected randomly and uniformly from S^t. Two vertices u and v are connected by an edge, if and only if $r_1 \le d(u, v) \equiv \|X_u - X_v\|_2 \le r_2$. Note that for $t = 1$ an $\mathrm{RAG}_1(n, [r_1, r_2])$ is nothing but a VRG as defined above, where we need to convert the Euclidean distance to the geodesic distance and scale the probabilities by a factor of 2π. The $\mathrm{RAG}_t(n, [0, r])$ gives the standard definition of random geometric graphs in t dimensions (for example, see [6] or [26]).

We refer to high-dimensional VRGs as the random annulus graph (RAG) since here two vertices are connected iff one is within an “annulus” centered at the other. For the random annulus graphs, we extend our connectivity results of $t = 1$ to general t. In particular, we show that there exists an isolated vertex in the $\mathrm{RAG}_t(n, [b(\frac{\ln n}{n})^{\frac{1}{t}}, a(\frac{\ln n}{n})^{\frac{1}{t}}])$ with high probability if and only if

$$a^t - b^t < \frac{\sqrt{\pi}(t+1)\Gamma(\frac{t+2}{2})}{\Gamma(\frac{t+3}{2})} \equiv \psi(t),$$

where $\Gamma(\cdot)$ is the gamma function. Computing the connectivity threshold of RAG exactly is highly challenging, and we have to use several approximations of high dimensional geometry. Our arguments crucially rely on VC dimensions of sets of geometric objects such as intersections of high dimensional annuluses and hyperplanes. Overall we find that the $\mathrm{RAG}_t(n, [b(\frac{\ln n}{n})^{\frac{1}{t}}, a(\frac{\ln n}{n})^{\frac{1}{t}}])$ is connected with high probability if

$$(a/2)^t - b^t \ge 8(t+1)\psi(t) \text{ and } a > 2b.$$

Using the connectivity result for RAG_t, the results for the geometric block model can be extended to high dimensions. The latent feature space of nodes in most networks are high-dimensional. For example, road networks are two-dimensional whereas the number of features used in a social network may have much higher dimensions. In a “high-dimensional” GBM: for any $t > 1$, instead of assigning a random variable from $[0, 1]$ we assign a random vector $X_u \in S^t$ to each vertex u; and two vertices in the same part is connected if and only if their Euclidean distance is less than r_s, whereas two vertices from different parts are connected if and only if their distance is less than r_d. We show the algorithm developed for one dimension extends to higher dimensions with nearly tight lower and upper bounds.

In this paper, we consistently refer to the $t = 1$ case for RAG as the vertex-random graph.

The paper is organized as follows. In Section 2, we provide the formal definitions and the main results of the paper. In Section 3, the sharp connectivity phase transition results for vertex-random graphs are proven. In Section 4, the connectivity results are proven for high dimensional random annulus graphs (details in full version [16]). Finally, in Section 5, a lower bound for the geometric block model as well as the main recovery algorithm are presented (details for the high-dimensional case in full version [16]).

2 Main Results

We formally define the random graph models, and state our results here.

▶ **Definition 1** (Vertex-Random Graph). *A vertex-random graph* $\mathrm{VRG}(n, [r_1, r_2])$ *on n vertices has parameters n, and a pair of real numbers $r_1, r_2 \in [0, 1/2], r_1 \leq r_2$. It is defined by assigning a number $X_i \in \mathbb{R}$ to vertex $i, 1 \leq i \leq n$, where X_i's are independent and identical random variables uniformly distributed in $[0, 1]$. There will be an edge between vertices i and $j, i \neq j$, if and only if $r_1 \leq d_L(X_i, X_j) \leq r_2$ where $d_L(X_i, X_j) \equiv \min\{|X_i - X_j|, 1 - |X_i - X_j|\}$.*

We choose $d_L(X_i, X_j) = \min\{|X_i - X_j|, 1 - |X_i - X_j|\}$ to ignore the boundary effect, although the results extend identically to the scenario when $d_L(X_i, X_j) = |X_i - X_j|$. One can also interpret $X_i, 1 \leq i \leq n$, to be uniformly distributed on the perimeter of a circle with radius $\frac{1}{2\pi}$ and the distance $d_L(\cdot, \cdot)$ to be the geodesic distance. As a shorthand, for any two vertices u, v, let $d(u, v)$ denote $d_L(X_u, X_v)$ where X_u, X_v are the random variables corresponding to the vertices. We also use $d(u, v)$ to denote the distance between a vertex u (or the embedding of that vertex in $[0, 1]$) and a point $v \in [0, 1]$ naturally. Our main result regarding VRGs is summarized in the following theorem.

▶ **Theorem 2** (Connectivity threshold of vertex-random graphs). *The* $\mathrm{VRG}(n, [\frac{b \ln n}{n}, \frac{a \ln n}{n}])$ *is connected with probability $1 - o(1)$ if $a > 1$ and $a - b > 0.5$. On the other hand, the* $\mathrm{VRG}(n, [\frac{b \ln n}{n}, \frac{a \ln n}{n}])$ *is not connected with probability $1 - o(1)$ if $a < 1$ or $a - b < 0.5$.*

Only for the special case of $b = 0$, the connectivity result was known before [25, 26]. See also [27]. Generalization to $b > 0$ is both nontrivial and counter-intuitive (the minimum connectivity gap is no longer $a - b \geq 1$). Indeed, our analysis also leads to an alternate simple proof of connectivity for one-dimensional RGGs.

▶ **Definition 3** (The Random Annulus Graph). *Let us define the t-dimensional unit sphere as $S^t \equiv \{x \in \mathbb{R}^{t+1} \mid ||x||_2 = 1\}$. A random annulus graph* $\mathrm{RAG}_t(n, [r_1, r_2])$ *on n vertices has parameters $n, t \in \mathbb{Z}_+$, and a pair of real numbers $r_1, r_2 \in [0, 2], r_1 \leq r_2$. It is defined by assigning a number $X_i \in S^t$ to vertex $i, 1 \leq i \leq n$, where X_i's are independent and identical random vectors uniformly distributed in S^t. There will be an edge between vertices i and $j, i \neq j$, if and only if $r_1 \leq \|X_i - X_j\|_2 \leq r_2$ where $\| \cdot \|_2$ denote the ℓ_2 norm.*

When from the context it is clear that we are in high dimensions, we use $d(u, v)$ to denote $\|X_u - X_v\|_2$ or just the ℓ_2 distance between the arguments[1]. The following result summarizes the condition for the existence of isolated vertices in RAGs.

[1] If we substitute $t = 1$, then $\mathrm{RAG}_1(n, [r_1, r_2])$ is a random graph where each vertex is associated with a random variable uniformly distributed in the unit circle. The distance between two vertices is the length of the chord connecting the random variables corresponding to the two vertices. If the length of the chord is $r \leq 2$, then the length of the corresponding (smaller) chord length of the corresponding arc between the vertices along the circumference of the circle is $2 \sin^{-1} \frac{r}{2}$. If we normalize the circumference of the circle by 2π, we obtain a random graph model that is equivalent to our definition of the vertex-random graphs. Since handling geodesic distances is more cumbersome in the higher dimensions, we resorted to Euclidean distance.

▶ **Theorem 4** (Zero-One law for Isolated Vertex in RAG). *For a random annulus graph* $\mathrm{RAG}_t(n,[r_1,r_2])$ *where* $r_2 = a\left(\frac{\ln n}{n}\right)^{\frac{1}{t}}$ *and* $r_1 = b\left(\frac{\ln n}{n}\right)^{\frac{1}{t}}$*, there exists isolated nodes with probability* $1-o(1)$ *if*

$$a^t - b^t < \frac{\sqrt{\pi}(t+1)\Gamma(\frac{t+2}{2})}{\Gamma(\frac{t+3}{2})} \equiv \psi(t),$$

where $\Gamma(x) = \int_0^\infty y^{x-1}e^{-y}dy$ *is the gamma function, and there does not exist an isolated vertex with probability* $1-o(1)$ *if* $a^t - b^t > \psi(t)$.

As a corollary of the above, we observe an $\mathrm{RAG}_t(n,[b\left(\frac{\ln n}{n}\right)^{\frac{1}{t}}, a\left(\frac{\ln n}{n}\right)^{\frac{1}{t}}])$ is not connected with probability $1-o(1)$ if $a^t - b^t < \psi(t)$. Our main result provides a sufficient condition for the connectivity.

▶ **Theorem 5.** *A* t *dimensional random annulus graph* $\mathrm{RAG}_t(n,[b\left(\frac{\ln n}{n}\right)^{\frac{1}{t}}, a\left(\frac{\ln n}{n}\right)^{\frac{1}{t}}])$ *is connected with probability* $1-o(1)$ *if* $(a/2)^t - b^t \geq 8(t+1)\psi(t)$ *and* $a > 2b$.

These connectivity results find immediate application in analyzing the geometric block model (GBM), a generative model for networks with underlying community structure.

▶ **Definition 6** (Geometric Block Model). *Given* $V = V_1 \sqcup V_2, |V_1| = |V_2| = \frac{n}{2}$*, choose a random variable* X_u *uniformly distributed in* $[0,1]$ *for all* $u \in V$*. The geometric block model* $\mathrm{GBM}(r_s, r_d)$ *with parameters* $1/2 \geq r_s > r_d$ *is a random graph where an edge exists between vertices* u *and* v *if and only if,*

$$\begin{aligned} &d_L(X_u,X_v) \leq r_s \text{ when } u,v \in V_1 \text{ or } u,v \in V_2 \\ &d_L(X_u,X_v) \leq r_d \text{ when } u \in V_1, v \in V_2 \text{ or } u \in V_2, v \in V_1. \end{aligned}$$

As a consequence of the connectivity lower bound on VRG, we are able to show community recovery lower bound, that is we show the recovery of the partition is not possible with high probability in $\mathrm{GBM}(\frac{a\ln n}{n}, \frac{b\ln n}{n})$ whenever $a-b < 0.5$ or $a < 1$ (see, Theorem 18). If in addition the vertex locations are known, then we can show a matching lower and upper bounds: the recovery is possible if and only if $a-b > 0.5$ or $a > 1$ (formal statement in full version [16]).

Coming back to the actual recovery problem, our main contribution for GBM is to provide a simple and efficient algorithm that performs well in the connectivity regime and recovers the clusters exactly. The following theorem provides a weaker (but simpler to understand) bound.

▶ **Theorem 7** (Recovery algorithm for GBM). *Suppose we have a graph* $G(V,E)$ *generated according to* $\mathrm{GBM}(r_s \equiv \frac{a\ln n}{n}, r_d \equiv \frac{b\ln n}{n})$ *and* $b > \frac{1}{4\ln 2 - 2}$*, then there exists an efficient algorithm (see Algorithm 1) which recovers the correct partition in* G *with probability* $1-o(1)$ *if* $a - 8b > 1$.

For the full range of parameter b, the (stronger) recovery guarantees for Algorithm 1 is discussed in Theorem 22 in Section 5. Table 1 lists some examples of the parameters when the proposed algorithm (Algorithm 1) can successfully recover the clusters. As can be anticipated, the connectivity results for RAGs apply to the "high dimensional" GBM.

Table 1 Minimum value of a, given b for which Algorithm 1 resolves clusters correctly in the setting for $\mathrm{GBM}(\frac{a\ln n}{n}, \frac{b\ln n}{n})$.

b	1	2	3	4	5	6	7
Minimum value of a	8.96	12.63	15.9	18.98	21.93	24.78	27.57

▶ **Definition 8** (The GBM in High Dimensions). *Given $V = V_1 \sqcup V_2, |V_1| = |V_2| = \frac{n}{2}$, choose a random vector X_u independently uniformly distributed in S^t for all $u \in V$. The geometric block model $\mathrm{GBM}_t(r_s, r_d)$ with parameters $r_s > r_d$ is a random graph where an edge exists between vertices u and v if and only if,*

$$||X_u - X_v||_2 \le r_s \text{ when } u, v \in V_1 \text{ or } u, v \in V_2$$
$$||X_u - X_v||_2 \le r_d \text{ when } u \in V_1, v \in V_2 \text{ or } u \in V_2, v \in V_1.$$

We extend the algorithmic results to high dimensions.

▶ **Theorem 9.** *There exists a polynomial time efficient algorithm that recovers the partition from $\mathrm{GBM}_t(r_s, r_d)$ with probability $1 - o(1)$ if $r_s = \Theta((\frac{\ln n}{n})^{\frac{1}{t}})$ and $r_s - r_d = \Omega((\frac{\ln n}{n})^{\frac{1}{t}})$. Moreover, any algorithm fails to recover the parts with probability at least $1/2$ if $r_s - r_d = o((\frac{\ln n}{n})^{\frac{1}{t}})$ or $r_s = o((\frac{\ln n}{n})^{\frac{1}{t}})$.*

3 Connectivity of Vertex-Random Graphs

In this section we give a proof of Theorem 2.

3.1 Sufficient condition for connectivity of VRG

▶ **Theorem 10.** *The vertex-random graph $\mathrm{VRG}(n, [\frac{b\ln n}{n}, \frac{a\ln n}{n}])$ is connected with probability $1 - o(1)$ if $a > 1$ and $a - b > 0.5$.*

To prove this theorem we use two main technical lemmas that show two different events happen with high probability simultaneously. First, we show that a $\mathrm{VRG}(n, [\frac{b\ln n}{n}, \frac{a\ln n}{n}])$ can be decomposed into union of cycles such that each of them cover $[0, 1]$. Second, we show there exists a vertex u_0 such that it has at least one neighbor in each cycle[2].

▶ **Lemma 11.** *A set of vertices $\mathcal{C} \subseteq V$ is called a cover of $[0, 1]$, if for any point y in $[0, 1]$ there exists a vertex $v \in \mathcal{C}$ such that $d(v, y) \le \frac{a\ln n}{2n}$. A $\mathrm{VRG}(n, [\frac{b\ln n}{n}, \frac{a\ln n}{n}])$ is a union of cycles such that every cycle forms a cover of $[0, 1]$ as long as $a - b > 0.5$ and $a > 1$ with probability $1 - o(1)$.*

Let us consider a weaker condition $a - b > 1$ than the statement of Lemma 11. This will be much easier to prove and already establishes the connectivity result for RGG in one dimension. Note that since the points are on a circle, it is natural to define a right (clockwise) and a left (counterclockwise) direction. When $a - b > 1$, we show each vertex has at least one neighbor on both directions. To see this for each vertex u , assign two indicator $\{0, 1\}$-random variables A_u^l and A_u^r, with $A_u^l = 1$ if and only if there is no node x to the left

[2] If the points are assumed to be present on a unit line [0,1], the same proof works with a difference that $\mathrm{VRG}(n, [\frac{b\ln n}{n}, \frac{a\ln n}{n}])$ can now be decomposed into a collection of paths that cover [0,1] and all these paths are connected through a vertex u_0. This analysis requires us to handle the nodes present in the boundary region – $[0, \frac{a\ln n}{n}]$ and $[1 - \frac{a\ln n}{n}, 1]$ separately.

of node u such that $d(u,x) \in [\frac{b\ln n}{n}, \frac{a\ln n}{n}]$. Similarly, let $A_u^r = 1$ if and only if there is no node x to the right of node u such that $d(u,x) \in [\frac{b\ln n}{n}, \frac{a\ln n}{n}]$. Now define $A = \sum_u (A_u^l + A_u^r)$. We have,

$$\Pr(A_u^l = 1) = \Pr(A_u^r = 1) = (1 - \frac{(a-b)\ln n}{n})^{n-1},$$

and,

$$\mathbb{E}[A] = 2n(1 - \frac{(a-b)\ln n}{n})^{n-1} \le 2n^{1-(a-b)}.$$

If $a - b > 1$ then $\mathbb{E}[A] = o(1)$ which implies, by invoking Markov inequality, that with high probability every node will have neighbors (connected by an edge in the VRG) on either side. Therefore every vertex will lie on a cycle that covers $[0,1]$. This is true for every vertex, hence the graph is simply a union of cycles each of which is a cover of $[0,1]$. The main technical challenge is to show that this conclusion remains valid even when $a - b > 0.5$, which is proved in Lemma 11 in Appendix A. Indeed, when $a - b > 0.5$, not every vertex will have neighbors on both sides; rather we need to analyze the connectivity via multi-hops to establish the desired result.

▶ **Lemma 12.** *Set two real numbers $k \equiv \lceil \frac{b}{(a-b)} \rceil + 1$ and $\epsilon < \frac{1}{2k}$. In an $\mathrm{VRG}(n, [\frac{b\ln n}{n}, \frac{a\ln n}{n}])$, $0 < b < a$, with probability $1 - o(1)$ there exists a vertex u_0 and k nodes $\{u_1, u_2, \dots, u_k\}$ to the right of u_0 such that $d(u_0, u_i) \in [\frac{(i(a-b)-2i\epsilon)\ln n}{n}, \frac{(i(a-b)-(2i-1)\epsilon)\ln n}{n}]$ and another set of k nodes $\{v_1, v_2, \dots, v_k\}$ also to the right of u_0 such that $d(u_0, v_i) \in [\frac{((i(a-b)+b-(2i-1)\epsilon)\ln n}{n}, \frac{(i(a-b)+b-(2i-2)\epsilon)\ln n}{n}]$, for $i = 1, 2, \dots, k$. The arrangement of the vertices is shown in Figure 1.*

We delegate the proof of this lemma to Appendix A.

Proof of Theorem 10. We have shown that the two events mentioned in Lemmas 11 and 12 happen with high probability. Therefore they simultaneously happen under the condition $a > 1$ and $a - b > 0.5$. Now we will show that these events together imply that the graph is connected. To see this, consider the vertices $u_0, \{u_1, u_2, \dots, u_k\}$ and $\{v_1, v_2, \dots, v_k\}$ that satisfy the conditions of Lemma 12. We can observe that each vertex v_i has an edge with u_i and u_{i-1}, $i = 1, \dots, k$. This is because (see Figure 1 for a depiction)

$$d(u_i, v_i) \ge \frac{((i(a-b)+b-(2i-1)\epsilon)\ln n}{n} - \frac{i(a-b)-(2i-1)\epsilon)\ln n}{n} = \frac{b\ln n}{n} \quad \text{and}$$

$$d(u_i, v_i) \le \frac{i(a-b)+b-(2i-2)\epsilon\ln n}{n} - \frac{(i(a-b)-2i\epsilon)\ln n}{n} = \frac{(b+2\epsilon)\ln n}{n}.$$

Similarly,

$$d(u_{i-1}, v_i) \ge \frac{((i(a-b)+b-(2i-1)\epsilon)\ln n}{n} - \frac{(i-1)(a-b)-(2i-3)\epsilon)\ln n}{n}$$
$$= \frac{(a-2\epsilon)\ln n}{n} \quad \text{and}$$

$$d(u_{i-1}, v_i) \le \frac{i(a-b)+b-(2i-2)\epsilon\ln n}{n} - \frac{((i-1)(a-b)-2(i-1)\epsilon)\ln n}{n} = \frac{a\ln n}{n}.$$

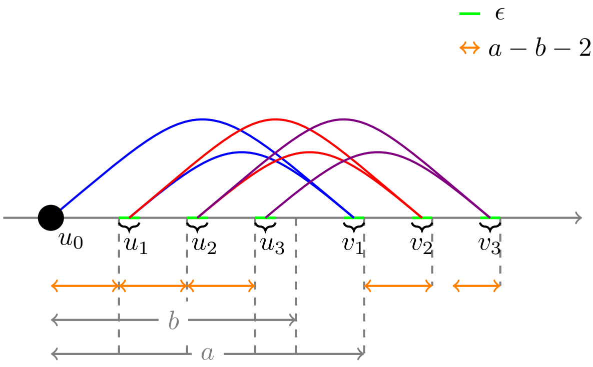

Figure 1 The location of u_i and v_i relative to u scaled by $\frac{\ln n}{n}$ in Lemma 12. Edges stemming put of v_1, v_2, v_3 are shown as blue, red and violet respectively.

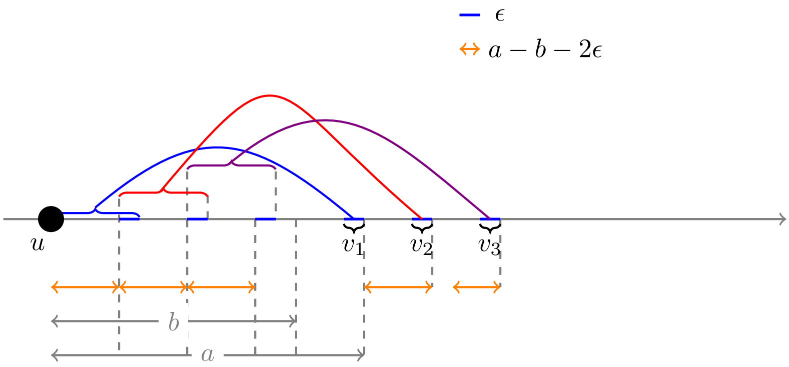

Figure 2 The line segments where v_1, v_2, v_3 can have neighbors (scaled by $\frac{\log n}{n}$) in the proof of Theorem 10. The point t has to lie in one of these regions.

This implies that u_0 is connected to u_i and v_i for all $i = 1, \ldots, k$. Using Lemma 11, the first event implies that the connected components are cycles spanning the entire line $[0, 1]$. Now consider two such disconnected components, one of which consists of the nodes $u_0, \{u_1, u_2, \ldots, u_k\}$ and $\{v_1, v_2, \ldots, v_k\}$. There must exist a node t in the other component (cycle) such that t is on the right of u_0 and $d(u_0, t) \equiv \frac{x \ln n}{n} \leq \frac{a \ln n}{n}$. If $x \leq b$, then there exists an i such that $i \leq k$ and $i(a-b) + b - a - (2i-2)\epsilon \leq x \leq i(a-b) - (2i-1)\epsilon$ (see Figure 2). Thus, when $x \leq b$, we can calculate the distance between t and v_i as

$$d(t, v_i) \geq \frac{(i(a-b) + b - (2i-1)\epsilon) \ln n}{n} - \frac{(i(a-b) - (2i-1)\epsilon) \ln n}{n} = \frac{b \ln n}{n}$$

and

$$d(t, v_i) \leq \frac{(i(a-b) + b - (2i-2)\epsilon) \ln n}{n} - \frac{(i(a-b) + b - a - (2i-2)\epsilon) \ln n}{n} = \frac{a \ln n}{n}.$$

Therefore t is connected to v_i when $x \leq b$. If $x > b$ then t is already connected to u_0. Therefore the two components (cycles) in question are connected.This is true for all cycles and hence there is only a single component in the entire graph. Indeed, if we consider the cycles to be disjoint super-nodes, then we have shown that there must be a star configuration. ◀

The following result is an immediate corollary of the connectivity upper bound.

▶ **Corollary 13.** *Consider a random graph $G(V, E)$ is being generated as a variant of the VRG where each $u, v \in V$ forms an edge if and only if $d(u,v) \in \left[0, c\frac{\ln n}{n}\right] \cup \left[b\frac{\ln n}{n}, a\frac{\ln n}{n}\right], 0 < c < b < a$. This graph is connected with probability $1 - o(1)$ if $a - b + c > 1$ or if $a - b > 0.5, a > 1$.*

3.2 Necessary condition for connectivity of VRG

▶ **Theorem 14** (VRG connectivity lower bound)**.** *The* $\mathrm{VRG}(n, [\frac{b\ln n}{n}, \frac{a\ln n}{n}])$ *is not connected with probability $1 - o(1)$ if $a < 1$ or $a - b < 0.5$.*

Proof. First of all, it is known that $\mathrm{VRG}(n, [0, \frac{a\ln n}{n}])$ is not connected with high probability when $a < 1$ [25, 26]. Therefore $\mathrm{VRG}(n, [\frac{b\ln n}{n}, \frac{a\ln n}{n}])$ must not be connected with high probability when $a < 1$ as the connectivity interval is a strict subset of the previous case, and $\mathrm{VRG}(n, [\frac{b\ln n}{n}, \frac{a\ln n}{n}])$ can be obtained from $\mathrm{VRG}(n, [0, \frac{a\ln n}{n}])$ by deleting all the edges that has the two corresponding random variables separated by distance less than $\frac{b\ln n}{n}$.

Next we will show that if $a - b < 0.5$ then there exists an isolated vertex with high probability. It would be easier to think of each vertex as a uniform random point in $[0, 1]$. Define an indicator variable A_u for every node u which is 1 when node u is isolated and 0 otherwise. We have,

$$\Pr(A_u = 1) = \Big(1 - \frac{2(a-b)\ln n}{n}\Big)^{n-1}.$$

Define $A = \sum_u A_u$, and hence

$$\mathbb{E}[A] = n\Big(1 - \frac{2(a-b)\ln n}{n}\Big)^{n-1} = n^{1-2(a-b)-o(1)}.$$

Therefore, when $a - b < 0.5$, $\mathbb{E}[A] = \Omega(1)$. To prove this statement with high probability we can show that the variance of A is bounded. Since A is a sum of indicator random variables, we have that

$$\begin{aligned}\mathrm{Var}(A) &\le \mathbb{E}[A] + \sum_{u\neq v} \mathrm{Cov}(A_u, A_v)\\ &= \mathbb{E}[A] + \sum_{u\neq v} (\Pr(A_u = 1 \cap A_v = 1) - \Pr(A_u = 1)\Pr(A_v = 1)).\end{aligned}$$

Now, consider the scenario when the vertices u and v are at a distance more than $\frac{2a\ln n}{n}$ apart (happens with probability $1 - \frac{4a\ln n}{n}$). Then the region in $[0, 1]$ that is between distances $\frac{b\ln n}{n}$ and $\frac{a\ln n}{n}$ from both of the vertices is empty and therefore $\Pr(A_u = 1 \cap A_v = 1) = \Big(1 - \frac{4(a-b)\ln n}{n}\Big)^{n-2}$. When the vertices are within distance $\frac{2a\ln n}{n}$ of one another, then $\Pr(A_u = 1 \cap A_v = 1) \le \Pr(A_u = 1)$. Therefore,

$$\begin{aligned}\Pr(A_u = 1 \cap A_v = 1) \le& (1 - \frac{4a\ln n}{n})\Big(1 - \frac{4(a-b)\ln n}{n}\Big)^{n-2} + \frac{4a\ln n}{n}\Pr(A_u = 1)\\ &\le (1 - \frac{4a\ln n}{n})n^{-4(a-b)+o(1)} + \frac{4a\ln n}{n}n^{-2(a-b)+o(1)}.\end{aligned}$$

Consequently for large enough n,

$$\begin{aligned}\Pr(A_u = 1 \cap A_v = 1) - \Pr(A_u = 1)\Pr(A_v = 1) \le (1 - \frac{4a\ln n}{n})n^{-4(a-b)+o(1)}\\ + \frac{4a\ln n}{n}n^{-2(a-b)+o(1)} - n^{-4(a-b)+o(1)} \le \frac{8a\ln n}{n}\Pr(A_u = 1).\end{aligned}$$

Now,

$$\mathrm{Var}(A) \leq \mathbb{E}[A] + \binom{n}{2}\frac{8a\ln n}{n}\Pr(A_u = 1) \leq \mathbb{E}[A](1 + 4a\ln n).$$

By using Chebyshev bound, with probability at least $1 - \frac{1}{\ln n}$,

$$A > n^{1-2(a-b)} - \sqrt{n^{1-2(a-b)}(1 + 4a\ln n)\ln n},$$

which imply for $a - b < 0.5$, there will exist isolated nodes with high probability. ◀

4 Connectivity of High Dimensional Random Annulus Graphs: Proof of Theorem 5

In this section we provide a proof sketch of Theorem 5 to establish the sufficient condition of connectivity of random annulus graphs. The details of the proof and the necessary conditions are provided in the full version [16].

Note, here $r_1 \equiv b\left(\frac{\ln n}{n}\right)^{1/t}$ and $r_2 \equiv a\left(\frac{\ln n}{n}\right)^{1/t}$. To show the upper bound for connectivity, the very first step is to define a *pole* which is a vertex that is connected to all vertices within a distance of r_2 from itself. We show such a pole exists with high probability in Lemma 15. This is a significant generalization of Lemma 12 from Section 3. We prove there exist annuli of suitably small radii around a node u_0 such that they are each non-empty and the vertices in these annuli are connected to each other along with u_0. Moreover the center of the annuli are collinear. Every point within distance r_2 from u_0 is then shown to be connected to at least one vertex in these constructed annuli.

▶ **Lemma 15.** *In a* $\mathrm{RAG}_t\left(n, \left[b\left(\frac{\ln n}{n}\right)^{1/t}, a\left(\frac{\ln n}{n}\right)^{1/t}\right]\right), 0 < b < a$, *with probability* $1 - o(1)$ *there exists a pole.*

Next, Lemma 16 shows that for every vertex u and every hyperplane L passing through u and not too close to the tangent hyperplane at u, there will be a neighbor of u on either side of the plane. Therefore, there should be a neighbor towards the direction of the pole. In order to formalize this, let us define a few regions associated with a node u and a hyperplane $L : w^T x = \beta$ passing through u.

$$\mathcal{R}_L^1 \equiv \{x \in S^t \mid r_1 \leq d(u,x) \leq r_2, w^T x \leq \beta\}$$
$$\mathcal{R}_L^2 \equiv \{x \in S^t \mid r_1 \leq d(u,x) \leq r_2, w^T x \geq \beta\}$$
$$\mathcal{A}_L \equiv \{x \mid x \in \mathcal{S}^t, \quad w^T x = \beta\}.$$

Informally, $\mathcal{R}_L^1$ and $\mathcal{R}_L^2$ represent the partition of the annulus on either side of the hyperplane L and $\mathcal{A}_L$ represents the region on the sphere lying on L.

▶ **Lemma 16.** *If we sample n nodes from S^t according to* $\mathrm{RAG}_t\left(n, \left[b\left(\frac{\ln n}{n}\right)^{1/t}, a\left(\frac{\ln n}{n}\right)^{1/t}\right]\right)$, *then for every node u and every hyperplane L passing through u such that $\mathcal{A}_L$ is not all within distance r_2 of u, node u has a neighbor on both sides of the hyperplane L with probability at least $1 - \frac{1}{n}$ provided $(a/2)^t - b^t \geq \frac{8\sqrt{\pi}(t+1)^2\Gamma(\frac{t+2}{2})}{\Gamma(\frac{t+3}{2})}$ and $a > 2b$.*

The proof of this lemma is quite challenging. Since, we do not know the location of the pole, we need to show that every point has a neighbor on both sides of the plane L no matter what the orientation of the plane. Since the number of possible orientations is uncountably infinite, we cannot use a union-bound type argument. To show this we have to rely on the VC Dimension of the family of sets $\{x \in S^t \mid r_1 \le \|u-x\|_2 \le r_2, w^Tx \ge \beta, \mathcal{A}_{L:w^Tx=\beta}$ not all within r_2 of $u\}$ for all hyperplanes L (which can be shown to be less than $t+1$). We rely on the celebrated result of [21] (we derive a continuous version of it), see full version [16], to deduce our conclusion.

For a node u and its corresponding location $X_u = (u_1, u_2, \dots, u_{t+1})$, define the particular hyperplane $L_u^\star : x_1 = u_1$ which is normal to the line joining $u_0 \equiv (1, 0, \dots, 0)$ and the origin and passes through u. We now need one more lemma that will help us prove Theorem 5.

▶ **Lemma 17.** *For a particular node u and corresponding hyperplane $L_u^\star$, if every point in $\mathcal{A}_{L_u^\star}$ is within distance r_2 from u, then u must be within r_2 of u_0.*

Proof of Theorem 5. We consider an alternate (rotated but not shifted) coordinate system by multiplying every vector by an orthonormal matrix such that the new position of the pole is the $t+1$-dimensional vector $(1, 0, \dots, 0)$ where only the first coordinate is non-zero. Let the $t+1$ dimensional vector describing any node u in this new coordinate system be $\hat{u} = (\hat{u}_1, \hat{u}_2, \dots, \hat{u}_{t+1})$. Now consider the hyperplane $L : x_1 = \hat{u}_1$ and if u is not connected to the pole already, then by Lemma 16 and Lemma 17, the node u has a neighbor u_2 which has a higher first coordinate ($\hat{u}_2 > \hat{u}_1$). The same analysis applies for u_2 and hence we have a path where the first coordinate of every node is higher than the previous node. Since the number of nodes is finite, this path cannot go on indefinitely and at some point, one of the nodes is going to be within r_2 of the pole and will be connected to the pole. Therefore every node is going to be connected to the pole and hence our theorem is proved. ◀

5 The Geometric Block Model

In this section, we prove a necessary condition for exact cluster recovery of the GBM and give an efficient algorithm that matches that within a constant factor. Very interestingly, our algorithm is based on a simple triangle counting method, whose variants are used as popular heuristics for community recovery in many real networks [3, 29, 11]. This further validates the suitability of GBMs as a community detection model.

5.1 Immediate consequence of VRG connectivity

The following lower bound for GBM can be obtained as a consequence of Theorem 2.

▶ **Theorem 18** (Impossibility in GBM). *Any algorithm to recover the partition in* $\mathrm{GBM}(\frac{a \ln n}{n}, \frac{b \ln n}{n})$ *will give incorrect output with probability* $1 - o(1)$ *if* $a - b < 0.5$ *or* $a < 1$.

Proof. Consider the scenario that not only the geometric block model graph $\mathrm{GBM}(\frac{a \ln n}{n}, \frac{b \ln n}{n})$ was provided to us, but also the random values $X_u \in [0,1]$ for all vertex u in the graph were provided. We will show that we will still not be able to recover the correct partition of the vertex set V with probability at least 0.5 (with respect to choices of X_u, $u, v \in V$ and any randomness in the algorithm).

In this situation, the edge (u, v) where $d_L(X_u, X_v) \le \frac{b \ln n}{n}$ does not give any new information than X_u, X_v. However the edges (u, v) where $\frac{b \ln n}{n} \le d_L(X_u, X_v) \le \frac{a \ln n}{n}$ are informative, as existence of such an edge will imply that u and v are in the same part. These edges constitute a vertex-random graph $\mathrm{VRG}(n, [\frac{b \ln n}{n}, \frac{a \ln n}{n}])$. But if there are more than two components in this vertex-random graph, then it is impossible to separate out the vertices into the correct two parts, as the connected components can be assigned to any of the two parts and the VRG along with the location values $(X_u, u \in V)$ will still be consistent.

What remains to be seen that $\mathrm{VRG}(n,[\frac{b\ln n}{n},\frac{a\ln n}{n}])$ will have $\omega(1)$ components with high probability if $a-b<0.5$ or $a<1$. This is certainly true when $a-b<0.5$ as we have seen in Theorem 14, there can indeed be $\omega(1)$ isolated nodes with high probability. On the other hand, when $a<1$, just by using an analogous argument it is possible to show that there are $\omega(1)$ vertices that do not have any neighbors on the left direction (counterclockwise). We delegate the proof of this claim as Lemma 19. If there are k such vertices, there must be at least $k-1$ disjoint candidates. This completes the proof. ◀

▶ **Lemma 19.** *A random geometric graph $G(n,\frac{a\ln n}{n})$ will have $\omega(1)$ disconnected components for $a<1$.*

Proof. Define an indicator random variable A_u for a node u which is 1 if it does not have a neighbor on its left. We must have that $\Pr(A_u)=\left(1-\frac{a\ln n}{n}\right)^{n-1}$. Therefore we must have that $\sum_u \mathbb{E}A_u = n^{1-a}=\Omega(1)$ if $a<1$. This statement also holds true with high probability. To show this we need to prove that the variance of $\sum_u \mathbb{E}A_u$ is bounded. We have

$$\begin{aligned}\mathrm{Var}(A) &< \mathbb{E}[A]+\sum_{u\neq v}\mathrm{Cov}(A_u,A_v)\\ &= \mathbb{E}[A]+\sum_{u\neq v}\Pr(A_u=1\cap A_v=1)-\Pr(A_u=1)\Pr(A_v=1)\end{aligned}$$

Now, consider the scenario when the vertices u and v are at a distance more than $\frac{2a\ln n}{n}$ apart (happens with probability at least $1-\frac{4a\ln n}{n}$). Then the region in $[0,1]$ that is within distance $\frac{a\ln n}{n}$ from both of the vertices is empty and therefore $\Pr(A_u=1\cap A_v=1)=\Pr(A_u=1)\Pr(A_v=1|A_u=1)\leq\Pr(A_u=1)\Pr(A_v=1)=(\Pr(A_u=1))^2$. When the vertices are within distance $\frac{2a\ln n}{n}$ of one another, then $\Pr(A_u=1\cap A_v=1)\leq\Pr(A_u=1)$. Therefore,

$$\Pr(A_u=1\cap A_v=1)\leq(1-\frac{4a\ln n}{n})(\Pr(A_u=1))^2+\frac{4a\ln n}{n}\Pr(A_u=1).$$

Consequently,

$$\begin{aligned}\Pr(A_u=1\cap A_v=1)-\Pr(A_u=1)\Pr(A_v=1)&\leq(1-\frac{4a\ln n}{n})(\Pr(A_u=1))^2\\ &+\frac{4a\ln n}{n}\Pr(A_u=1)-(\Pr(A_u=1))^2\leq\frac{4a\ln n}{n}\Pr(A_u=1).\end{aligned}$$

Now,

$$\mathrm{Var}(A)\leq\mathbb{E}[A]+\binom{n}{2}\frac{4a\ln n}{n}\Pr(A_u=1)\leq\mathbb{E}[A](1+2a\ln n).$$

By using Chebyshev bound, with probability at least $1-\frac{1}{\ln n}$,

$$A>n^{1-a}-\sqrt{n^{1-a}(1+2a\ln n)\ln n},$$

Now, observe that if there exist k nodes with no neighbor on one side, then there must exist $k-1$ disconnected components. Hence the number of components in $G(n,\frac{a\ln n}{n})$ is $\omega(1)$. ◀

Indeed, when the locations X_u associated with every vertex u is provided, it is also possible to recover the partition when $a-b>0.5$ and $a>1$, matching the above lower bound exactly. Similar impossibility result extends to higher dimensional GBM from the necessary condition on connectivity of RAG.

5.2 A Recovery Algorithm for GBM

We now turn our attention to an efficient recovery algorithm for GBM. Intriguingly, we show a simple triangle counting based algorithm works well for GBM and recovers the communities in the connectivity regime.

Algorithm 1 Community recovery in GBM.

```
Require: GBM $G = (V, E)$, $r_s, r_d$
for $(u, v) \in E$ do
    if process$(u, v, r_s, r_d)$=false then
        $E.remove((u, v))$
    end if
end for
return connectedComponent$(V, E)$
```

Algorithm 2 `process`.

```
Require: $u$,$v$, $r_s$, $r_d$
Ensure: true/false
    {Comment: When $a > 2b$, $t_1 = \min\{t : (2b + t) \ln \frac{2b+t}{2b} - t > 1\}, t_2 = \min\{t : (2b - t) \ln \frac{2b-t}{2b} + t > 1$ and $E_S = (2b + t_1) \frac{\ln n}{n}$ and $E_D = (2b - t_2) \frac{\ln n}{n}\}$
count $\leftarrow |\{z : (z, u) \in E, (z, v) \in E\}|$
if $\frac{\text{count}}{n} \geq E_S(r_d, r_s)$ or $\frac{\text{count}}{n} \leq E_D(r_d, r_s)$ then
    return true
end if
return false
```

Suppose we are given a graph $G = (V : |V| = n, E)$ with two disjoint parts, $V_1, V_2 \subseteq V$ generated according to GBM(r_s, r_d). The algorithm (Algorithm 1) goes over all edges $(u, v) \in E$. It counts the number of triangles containing the edge (u, v) by calling the `process` function that counts the number of common neighbors of u and v.

`process` outputs "true" if it is confident that the nodes u and v belong to the same cluster and "false" otherwise. More precisely, if the count is within some prescribed values E_S and E_D, it returns "false". Note that the thresholds E_S and E_D refer to the maximum and minimum value of triangle-count for an "inter-cluster" edge. The algorithm removes the edge on getting a "false" from `process` function. After processing all the edges of the network, the algorithm is left with a reduced graphs (with certain edges deleted from the original). It then finds the connected components in the graph and returns them as the parts V_1 and V_2.

It would have been natural to consider two thresholds E_D and E_S and if the triangle count of an edge is closer to E_S than E_D, then the two end-points are assigned to the same cluster and otherwise in separate clusters. Indeed such a natural algorithm has been analyzed in [15]. On the other hand, here we remove an edge if the triangle count lies in an interval. This is apparently non-intuitive, but gives a significant improvement over the previously known bound (see Figure 3).

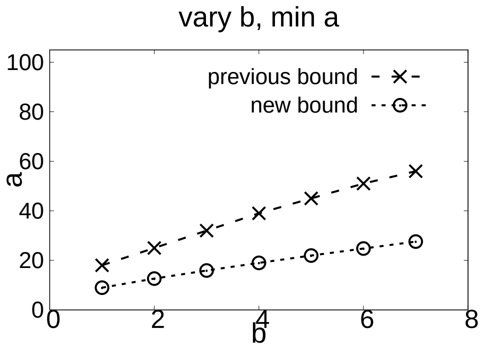

Figure 3 The minimum gap between a and b permitted by our algorithm vs the previously known bound of [15].

5.3 Analysis of Algorithm 1

Given a graph $G(V,E) \equiv GBM(r_s \equiv \frac{a \ln n}{n}, r_d = \frac{b \ln n}{n})$ with two clusters $V = V_1 \sqcup V_2$, and a pair of vertices $u, v \in V$, the events $\mathcal{E}_z^{u,v}, z \in V$ of any other vertex z being a common neighbor of both u and v given $(u,v) \in E$ are dependent; however given the distance between the corresponding random variables $d_L(X_u, X_v) = x$, the events are independent. This is a crucial observation that lets us overcome the difficulty of handling correlated edge formation.

Moreover, given the distance between two nodes u and v are the same, the probabilities of $\mathcal{E}_z^{u,v} \mid (u,v) \in E$ are different when u and v are in the same cluster and when they are in different clusters. Therefore the count of the common neighbors are going to be different, and substantially separated with high probability for two vertices in cases when they are from the same cluster or from different clusters. However, this may not be the case, if we do not restrict the distance to be the same and look at the entire range of possible distances.

The distribution of the number of common neighbors given $(u,v) \in E$ and $d(u,v) = x$ is given in Table 2 (follows from Lemma 23 and Lemma 24 from Appendix). As throughout this paper, we have assumed that there are only two clusters of equal size. In the table, $u \sim v$ means u and v are in the same cluster and $\text{Bin}(n,p)$ denotes a binomial random variable with mean np.

Table 2 Distribution of triangle count for an edge (u,v) conditioned on the distance between them $d(u,v) = d_L(X_u, X_v) = x$, when there are two equal sized clusters.

$(u,v) \in E$	Distribution of count $(r_s > 2r_d)$		Distribution of count $(r_s \le 2r_d)$	
$d(u,v) = x$	$u \sim v, x \le r_s$	$u \nsim v, x \le r_d$	$u \sim v, x \le r_s$	$u \nsim v, x \le r_d$
Motif : $z \mid (z,u) \in E, (z,v) \in E$	$\text{Bin}(\frac{n}{2} - 2, 2r_s - x) + \mathbb{1}\{x \le 2r_d\}\text{Bin}(\frac{n}{2}, 2r_d - x)$	$\text{Bin}(n-2, 2r_d)$	$\text{Bin}(\frac{n}{2} - 2, 2r_s - x) + \text{Bin}(\frac{n}{2}, 2r_d - x)$	$\text{Bin}(n-2, \min(r_s + r_d - x, 2r_d))$

At this point in a $\text{GBM}(r_s, r_d)$ for any edge u, v that does not belong to the same part, the expected total number of common neighbors of u and v does not depend on their distance. In Lemma 20, we show that in this case the normalized total number of common neighbors is concentrated around $2r_d$.

▶ **Lemma 20.** *Suppose we are given a graph $G(V,E)$ generated according to* $\mathrm{GBM}(r_s \equiv \frac{a\ln n}{n}, r_d \equiv \frac{b\ln n}{n}), a \geq 2b$. *Our algorithm with $E_S = (2b+t_1)\frac{\ln n}{n}$ and $E_D = (2b-t_2)\frac{\ln n}{n}$, deletes all the edges $(u,v) \in E$ such that u and v are in different parts with probability at least $1-o(1)$, where*

$$t_1 = \min\{t : (2b+t)\ln\frac{2b+t}{2b} - t > 1\}, \quad t_2 = \min\{t : (2b-t)\ln\frac{2b-t}{2b} + t > 1\}.$$

Therefore, when Algorithm 1 finishes processing all the edges, all the "inter-cluster" edges are removed with high probability. However some of the "in-cluster" edges are also deleted, namely, those that have a count of common neighbors between E_S and E_D. In the next lemma, we show the necessary condition on the "in-cluster" edges such that they do not get removed by Algorithm 1.

▶ **Lemma 21.** *Suppose we are given a graph $G(V,E)$ generated according to* $\mathrm{GBM}(r_s \equiv \frac{a\ln n}{n}, r_d \equiv \frac{b\ln n}{n}), a \geq 2b$. *Define t_1, t_2, E_D, E_S as in Lemma 20. Consider an edge $(u,v) \in E$ where u,v belong to the same part of the GBM and let $d(u,v) \equiv x \equiv \frac{\theta \ln n}{n}$. Suppose θ satisfies either of the following conditions:*

1. $\frac{1}{2}\Big((4b+2t_1)\ln\frac{4b+2t_1}{2a-\theta} + 2a - \theta - 4b - 2t_1\Big) > 1$ *and* $0 \leq \theta \leq 2a - 4b - 2t_1$

2. $\frac{1}{2}\Big((4b-2t_2\ln\frac{4b-2t_2}{2a-\theta} + 2a - \theta - 4b + 2t_2\Big) > 1$ *and* $a \geq \theta \geq \max\{2b, 2a-4b+2t_2\}..$

Then Algorithm 1 with $E_S = (2b+t_1)\frac{\ln n}{n}$ and $E_D = (2b-t_2)\frac{\ln n}{n}$ will not remove this edge with probability at least $1 - O(\frac{1}{n(\ln n)^2})$.

Now we are in a position to prove our main theorem from this part.

▶ **Theorem 22.** *Suppose we are given a graph $G(V,E)$ generated according to* $\mathrm{GBM}(r_s \equiv \frac{a\ln n}{n}, r_d \equiv \frac{b\ln n}{n}), a \geq 2b$. *Define t_1, t_2, E_S and E_D as in Lemma 20, and θ_1 and θ_2 as in Lemma 21. Then Algorithm 1 recovers the correct partition in G with probability $1-o(1)$ if $a - \theta_2 + \theta_1 > 2$ OR $a - \theta_2 > 1, a > 2$.*

Proof. From Lemma 20, we know that after Algorithm 1 has processed all the edges, the edges with end-points in different parts of the GBM are all deleted with probability $1-o(1)$. Moreover, from Lemma 21, an intra-cluster edge (u,v) will continue to exist if $d(u,v) \in [0,\theta_1] \cup [\theta_2, a]$ (by simply applying a union bound over at most $O(n\log n)$ edges). From Corollary 13, it is evident that each of the two parts of size $\frac{n}{2}$ each will be connected if either $a - \theta_2 + \theta_1 > 2$ or $a - \theta_2 > 1$ and $a > 2$. ◀

Theorem 7 is a weaker version of Theorem 22 which we obtain by setting specific values.

Proof of Theorem 7. Following the proof of Theorem 22, when $E_D = 0$ and $E_S = (2b+t_1)\frac{\ln n}{n}$, after Algorithm 1 processes all the edges, an edge between a pair u and v will continue to exist if $d(u,v) \in [0,\theta_1]$ which is equivalent to setting $\theta_2 \leq a$. Consider the case when $b > \frac{1}{4\ln 2 - 2}$. Note that from Theorem 22, $t_1 = \min\{t : (2b+t)\ln\frac{2b+t}{2b} - t > 1\}$. We see that $t = 2b$ satisfies the above condition since, $(2b+t)\ln\frac{2b+t}{2b} - t = 4b\ln 2 - 2b > 1$. This shows that $t_1 \leq 2b$. Similarly, from Theorem 22,

$$\theta_1 = \max\{\theta : \frac{1}{2}\Big((4b+2t_1)\ln\frac{4b+2t_1}{2a-\theta} + 2a - \theta - 4b - 2t_1\Big) > 1 \text{ and}$$
$$0 \leq \theta \leq 2a - 4b - 2t_1.\}$$

When $t_1 \leq 2b$, the expression $\theta \leq 2a - 4b - 2t_1$ is satisfied for all values of $\theta \leq 2a - 8b$. Hence, we choose $\theta = 2a - 16b$ to simplify the other expression and get the following chain of equations:

$$\begin{aligned}
&\frac{1}{2}\Big((4b + 2t_1)\ln\frac{4b + 2t_1}{2a - \theta} + 2a - \theta - 4b - 2t_1\Big)\\
&\geq \frac{1}{2}\Big((4b + 2t_1)\ln\frac{4b + 2t_1}{2a - \theta} + 2a - \theta - 8b\Big) = \frac{1}{2}\Big((4b + 2t_1)\ln\frac{4b + 2t_1}{2a - \theta}\Big) + 4b\\
&\geq \frac{1}{2}\Big((4b)\ln\frac{4b}{2a - \theta}\Big) + 4b \geq \frac{1}{2}\Big((4b)\ln\frac{4b}{16b}\Big) + 4b = -2b\ln 4 + 4b
\end{aligned}$$

which is greater than 1 whenever b satisfies $b > \frac{1}{4-4\ln 2}$. However, since we assumed that $b > \frac{1}{2(2\ln 2-1)}$, the condition $b > \frac{1}{4-4\ln 2}$ is automatically satisfied as $\frac{1}{2(2\ln 2-1)} > \frac{1}{4-4\ln 2}$. This implies that $\theta_1 > 2a - 16b$.

Using, $\theta_1 > 2a - 16b$ and $\theta_2 = a$, the final condition of Theorem 22, $a - \theta_2 + \theta_1 > 2$ is satisfied whenever $\theta_1 > 2$ that is, $2a - 16b > 2$. Hence, whenever $2a - 16b > 2$, or, $a - 8b > 1$, Algorithm 1 will recover the correct partition with probability $1 - o(1)$. ◀

References

1 Emmanuel Abbe, Afonso S. Bandeira, and Georgina Hall. Exact Recovery in the Stochastic Block Model. *IEEE Trans. Information Theory*, 62(1):471–487, 2016.

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

3 Austin R Benson, David F Gleich, and Jure Leskovec. Higher-order organization of complex networks. *Science*, 353(6295):163–166, 2016.

4 Béla Bollobás. Random Graphs. *Cambridge Press*, 2001.

5 Béla Bollobás. Percolation. *Cambridge Press*, 2006.

6 Sébastien Bubeck, Jian Ding, Ronen Eldan, and Miklós Z Rácz. Testing for high-dimensional geometry in random graphs. *Random Structures & Algorithms*, pages 503–532, 2016.

7 Peter Chin, Anup Rao, and Van Vu. Stochastic block model and community detection in sparse graphs: A spectral algorithm with optimal rate of recovery. In *Conference on Learning Theory (COLT)*, pages 391–423, 2015.

8 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.

9 Carl P Dettmann and Orestis Georgiou. Random geometric graphs with general connection functions. *Physical Review E*, 93(3):032313, 2016.

10 Martin E. Dyer and Alan M. Frieze. The solution of some random NP-hard problems in polynomial expected time. *Journal of Algorithms*, 10(4):451–489, 1989.

11 David Easley and Jon Kleinberg. Networks, crowds, and markets. *Cambridge Books*, 2012.

12 Paul Erdös and Alfréd Rényi. On random graphs, I. *Publicationes Mathematicae (Debrecen)*, 6:290–297, 1959.

13 Stephen Eubank, Hasan Guclu, VS Anil Kumar, Madhav V Marathe, Aravind Srinivasan, Zoltan Toroczkai, and Nan Wang. Modelling disease outbreaks in realistic urban social networks. *Nature*, 429(6988):180, 2004.

14 Ehud Friedgut and Gil Kalai. Every monotone graph property has a sharp threshold. *Proceedings of the American mathematical Society*, 124(10):2993–3002, 1996.

15 Sainyam Galhotra, Arya Mazumdar, Soumyabrata Pal, and Barna Saha. The Geometric Block Model. In *The Thirty-Second AAAI Conference on Artificial Intelligence (AAAI-18)*, 2018.

16 Sainyam Galhotra, Arya Mazumdar, Soumyabrata Pal, and Barna Saha. Connectivity in random annulus graphs and the geometric block model. *arXiv preprint*, 2019. `arXiv:1804.05013`.

17 Edgar N Gilbert. Random graphs. *The Annals of Mathematical Statistics*, 30(4):1141–1144, 1959.

18 Edward N Gilbert. Random plane networks. *Journal of the Society for Industrial and Applied Mathematics*, 9(4):533–543, 1961.

19 Martin Haenggi, Jeffrey G Andrews, François Baccelli, Olivier Dousse, and Massimo Franceschetti. Stochastic geometry and random graphs for the analysis and design of wireless networks. *IEEE Journal on Selected Areas in Communications*, 27(7):1029–1046, 2009.

20 Bruce E. Hajek, Yihong Wu, and Jiaming Xu. Computational Lower Bounds for Community Detection on Random Graphs. In *Proceedings of The 28th Conference on Learning Theory (COLT)*, pages 899–928, 2015. URL: `http://proceedings.mlr.press/v40/Hajek15.html`.

21 David Haussler and Emo Welzl. epsilon-nets and simplex range queries. *Discrete & Computational Geometry*, 2(2):127–151, 1987.

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

23 Jure Leskovec, Kevin J Lang, Anirban Dasgupta, and Michael W Mahoney. Statistical properties of community structure in large social and information networks. In *17th international conference on World Wide Web*, pages 695–704, 2008.

24 Elchanan Mossel, Joe Neeman, and Allan Sly. Consistency thresholds for the planted bisection model. In *47th Annual ACM Symposium on Theory of Computing (STOC)*, pages 69–75, 2015.

25 S Muthukrishnan and Gopal Pandurangan. The bin-covering technique for thresholding random geometric graph properties. In *Proceedings of the sixteenth annual ACM-SIAM symposium on Discrete algorithms (SODA)*, pages 989–998, 2005.

26 Mathew Penrose. Random geometric graphs. *Oxford University Press*, 2003.

27 Mathew D Penrose. Connectivity of soft random geometric graphs. *The Annals of Applied Probability*, 26(2):986–1028, 2016.

28 Abishek Sankararaman and François Baccelli. Community Detection on Euclidean Random Graphs. In *Proceedings of the Twenty-Ninth Annual ACM-SIAM Symposium on Discrete Algorithms (SODA)*, pages 2181–2200, 2018.

29 Charalampos E Tsourakakis, Jakub Pachocki, and Michael Mitzenmacher. Scalable motif-aware graph clustering. In *Proceedings of the 26th International Conference on World Wide Web (WWW)*, pages 1451–1460, 2017.

30 Stephen J Young and Edward R Scheinerman. Random dot product graph models for social networks. In *International Workshop on Algorithms and Models for the Web-Graph*, pages 138–149, 2007.

31 Weituo Zhang, Chjan C Lim, Gyorgy Korniss, and Boleslaw K Szymanski. Opinion dynamics and influencing on random geometric graphs. *Scientific reports, Nature Publishing Group*, 4:5568, 2014.

A Proof of Lemma 11 and Lemma 12

Proof of Lemma 11. The proof of this lemma is somewhat easily explained if we consider a weaker result (a stronger condition) with $a - b > 2/3$. Let us first briefly describe this case.

Consider a node u and assume without loss of generality that the position of u is 0 (i.e. $X_u = 0$). Associate four indicator $\{0, 1\}$-random variables $A_u^i, i = 1, 2, 3, 4$ which take the value of 1 if and only if there does not exist any node x such that

1. $d(u, x) \in [b\frac{\ln n}{n}, a\frac{\ln n}{n}] \cup [0, \frac{a-b}{2}\frac{\ln n}{n}]\}$ for $i = 1$
2. $d(u, x) \in [b\frac{\ln n}{n}, a\frac{\ln n}{n}] \cup [\frac{-a-b}{2}\frac{\ln n}{n}, -b\frac{\ln n}{n}]\}$ for $i = 2$
3. $d(u, x) \in [-a\frac{\ln n}{n}, -b\frac{\ln n}{n}] \cup [\frac{-a+b}{2}\frac{\ln n}{n}, 0]\}$ for $i = 3$
4. $d(u, x) \in [-a\frac{\ln n}{n}, -b\frac{\ln n}{n}] \cup [b\frac{\ln n}{n}, \frac{a+b}{2}\frac{\ln n}{n}]\}$ for $i = 4$.

The intervals representing these random variables are shown in Figure 4.

Notice that $\Pr(A_u^i = 1) \le \max\{\left(1 - 1.5(a-b)\frac{\ln n}{n}\right)^{n-1}, \left(1 - a\frac{\ln n}{n}\right)^{n-1}\}$ and therefore $\sum_{i,u} \mathbb{E}A_u^i \le 4\max\{n^{1-1.5(a-b)}, n^{1-a}\} = 4n^{\min\{1-1.5(a-b),1-a\}}$. This means that for $a - b \ge 0.67$ and $a \ge 1$, $\sum_{i,u} \mathbb{E}A_u^i = o(1)$. Hence there exist vertices in all the regions described above for every node u with high probability.

Now, A_u^1 and A_u^2 being zero implies that either there is a vertex in $[b\frac{\ln n}{n}, a\frac{\ln n}{n}]$ or there exists two vertices v_1, v_2 in $[0, \frac{a-b}{2}\frac{\ln n}{n}]$ and $[\frac{-a-b}{2}\frac{\ln n}{n}, -b\frac{\ln n}{n}]$ respectively (see, Figure 4). In the second case, u is connected to v_2 and v_2 is connected to v_1. Therefore u has nodes on left (v_2) and right (v_1) and u is connected to both of them through one hop in the graph.

Similarly, A_u^3 and A_u^4 being zero implies that either there exists a vertex in $[-a\frac{\ln n}{n}, -b\frac{\ln n}{n}]$ or again u will have vertices on left and right and will be connected to them. So, when all the four $A_u^i, i = 1, 2, 3, 4$ are zero together:

- $A_u^1 = A_u^2 = 0$ implies there is a neighbor of u on either sides or there is a single node in $[b\frac{\ln n}{n}, a\frac{\ln n}{n}]$
- $A_u^3 = A_u^4 = 0$ implies there is a neighbor of u on either sides or there is a single node in $[-a\frac{\ln n}{n}, -b\frac{\ln n}{n}]$

This shows that when $A_u^1 = A_u^2 = 0$ and $A_u^3 = A_u^4 = 0$ guarantee a node on only one side of u, there are nodes in $[b\frac{\ln n}{n}, a\frac{\ln n}{n}]$ and $[-a\frac{\ln n}{n}, -b\frac{\ln n}{n}]$. But in that case u has direct neighbors on both its left and right. We can conclude that every vertex u is connected to a vertex v on its right and a vertex w on its left such that $d(u, v) \in [0, a\frac{\ln n}{n}]$ and $d(u, w) \in [-a\frac{\ln n}{n}, 0]$; therefore every vertex is part of a cycle that covers $[0, 1]$.

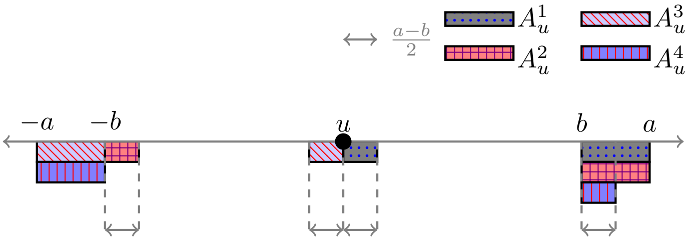

Figure 4 Representation of four different random variables for Lemma 11.

We can now extend this proof to the case when $a - b > 0.5$.

Let c be large number to be chosen specifically later. Consider a node u and assume that the position of u is 0. Now consider the four different regions $[-a\frac{\ln n}{n}, -b\frac{\ln n}{n}]$, $[-(a-b)\frac{\ln n}{n}, 0]$, $[b\frac{\ln n}{n}, a\frac{\ln n}{n}]$ and $[0, a - b\frac{\ln n}{n}]$ around u each divided into $L \equiv 2^c$ patches (intervals) of size $\theta = \frac{a-b}{2^c}$ in the following way:

1. $I_u^i = [\frac{(-a+(i-1)\theta)\ln n}{n}, \frac{(-a+i\theta)\ln n}{n}]$
2. $J_u^i = [\frac{(-(a-b)+(i-1)\theta)\ln n}{n}, \frac{(-(a-b)+i\theta)\ln n}{n}]$
3. $K_u^i = [\frac{(b+(i-1)\theta)\ln n}{n}, \frac{(b+i\theta)\ln n}{n}]$
4. $M_u^i = [\frac{((i-1)\theta)\ln n}{n}, \frac{i\theta\ln n}{n}]$

where $i = 1, 2, 3, \ldots, L$. Note that any vertex in $\cup I_u^i \cup K_u^i$ is connected to u. See, Figure 5 for a depiction.

Consider a $\{0, 1\}$-indicator random variable X_u that is 1 if and only if there does not exist any node in a region formed by union of any $2L - 1$ patches amongst the ones described above. Notice that when $a < 2b$, the patches do not overlap and the total size of $2L - 1$

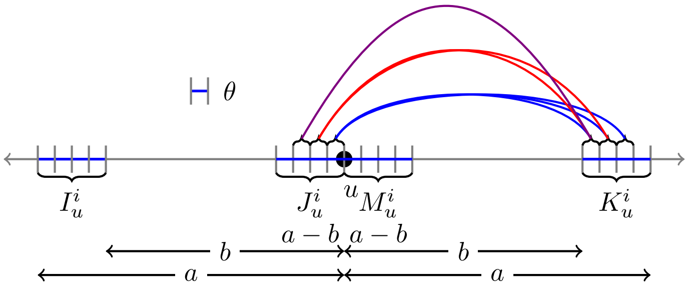

Figure 5 Pictorial representation of $I_u^i, J_u^i, K_u^i, M_u^i$ and their connectivity as described in Lemma 11. The colored lines show the regions that are connected to each other.

patches is $\frac{2^{c+1}-1}{2^c}\frac{(a-b)\ln n}{n}$ and when $a \geq 2b$, the patches can overlap and the total size of the $2L-1$ patches is going to be more than $\min\{\frac{2^{c+1}-1}{2^c}\frac{(a-b)\ln n}{n}, \frac{a\ln n}{n}\}$. Since there are $\binom{4L}{2L-1} \leq n^{\frac{4L}{\ln n}}$ possible regions that consists of $2L-1$ patches,

$$\sum_u \mathbb{E}X_u \leq n\binom{4L}{2L-1}\Big(1-\min\{\frac{2^{c+1}-1}{2^c}\frac{(a-b)\ln n}{n}, \frac{a\ln n}{n}\}\Big)^{n-1}$$
$$\leq \max\{n^{1-\frac{2^{c+1}-1}{2^c}(a-b)+\frac{4L}{\ln n}}, n^{1-a+\frac{4L}{\ln n}}\}.$$

At this point we can choose $c = c_n = o(\ln n)$ such that $\lim_n c_n = \infty$. Hence when $a-b > \frac{1}{2}$ and $a > 1$, for every vertex u there exists at least one patch amongst every $2L-1$ patches in $\cup I_u^i \cup J_u^j \cup K_u^k, i,j,k = 1,2,\ldots,L$ that contains a vertex.

Consider a collection of patches $\cup_i I_u^i \cup_j K_u^j, i,j = 1,2,\ldots,L$. We know that there exist two patches amongst these I_u^is and K_u^js that contain at least one vertices. If one of I_u^is and one of K_u^js contain two vertices, we found one neighbor of u on both left and right directions (see, Figure 5).

We consider the other case now. Without loss of generality assume that there are no vertex in all I_u^is and there exist at least two patches in K_u^is that contain at least one vertex each. Hence, there exists at least one of $\{K_u^i \mid i \in \{1,2,\ldots,L-1\}\}$ that contains a vertex. Similarly, we can also conclude in this case that there exists at least one of $\{J_u^i \mid i \in \{2,3\ldots,L\}\}$ which contain a node. Assume J_u^ϕ to be the left most patch in $\cup J_u^i \mid i \in \{1,2,\ldots,L\}$ that contains a vertex (see, Figure 5) . From our previous observation, we can conclude that $\phi \geq 2$.

We can observe that any vertex in J_u^j is connected to the vertices in patches $K_u^k, \forall k < j$. This is because for two vertices $v \in J_u^j$ and $w \in K_u^k$, we have

$$d(v,w) \geq \frac{(b+(k-1)\theta)\ln n}{n} - \frac{(-(a-b)+j\theta)\ln n}{n} = \frac{(a+(k-j-1)\theta)\ln n}{n};$$
$$d(v,w) \leq \frac{(b+k\theta)\ln n}{n} - \frac{(-(a-b)+(j-1)\theta)\ln n}{n} = \frac{(a+(k-j+1)\theta)\ln n}{n}.$$

Consider a collection of $2L-1$ patches $\{\cup I_u^i \cup J_u^j \cup K_u^k \mid i,j,k \in \{1,\ldots,L\}, j > \phi, k \leq \phi-1\}$ where $\phi \geq 2$. This is a collection of $2L-1$ patches out of which one must have a vertex and since none of $\{J_u^j \mid j > \phi\}$ and I_u^i can contain a vertex, one of $\{K_u^k \mid k \leq \phi-1\}$ must contain the vertex. Recall that the vertex in J_u^ϕ is connected to any node in K_u^k for any $k \leq \phi-1$ and therefore u has a node to the right direction and left direction that are connected to u. Therefore every vertex is part of a cycle and each of the circles covers $[0,1]$. ◀

Proof of Lemma 12. Recall that we want to show that there exists a node u_0 and k nodes $\{u_1, u_2, \dots, u_k\}$ to the right of u_0 such that $d(u_0, u_i) \in [\frac{(i(a-b)-2i\epsilon)\ln n}{n}, \frac{(i(a-b)-(2i-1)\epsilon)\ln n}{n}]$ and exactly k nodes $\{v_1, \dots, v_k\}$ to the right of u_0 such that $d(u_0, v_i) \in [\frac{((i(a-b)+b-(2i-1)\epsilon)\ln n}{n}, \frac{(i(a-b)+b-(2i-2)\epsilon)\ln n}{n}]$, for $i = 1, 2, \dots, k$ and ϵ is a constant less than $\frac{1}{2k}$ (see Figure 1 for a depiction). Let A_u be an indicator $\{0,1\}$-random variable for every node u which is 1 if u satisfies the above conditions and 0 otherwise. We will show $\sum_u A_u \geq 1$ with high probability. We have,

$$\begin{aligned}\Pr(A_u = 1) &= n(n-1)\dots(n-(2k-1))\Big(\frac{\epsilon \ln n}{n}\Big)^{2k}\Big(1 - 2k\epsilon\frac{\ln n}{n}\Big)^{n-2k} \\ &= c_0 n^{-2k\epsilon}(\epsilon \ln n)^{2k}\prod_{i=0}^{2k-1}(1-i/n) = c_1 n^{-2k\epsilon}(\epsilon \ln n)^{2k}\end{aligned}$$

where c_0, c_1 are just absolute constants independent of n (recall k is a constant). Hence,

$$\sum_u \mathbb{E}A_u = c_1 n^{1-2k\epsilon}(\epsilon \ln n)^{2k} \geq 1$$

as long as $\epsilon \leq \frac{1}{2k}$. Now, in order to prove $\sum_u A_u \geq 1$ with high probability, we will show that the variance of $\sum_u A_u$ is bounded from above. This calculation is very similar to the one in the proof of Theorem 14. Recall that if $A = \sum_u A_u$ is a sum of indicator random variables, we must have

$$\mathrm{Var}(A) \leq \mathbb{E}[A] + \sum_{u \neq v}\mathrm{Cov}(A_u, A_v) = \mathbb{E}[A] + \sum_{u \neq v}\Pr(A_u = 1 \cap A_v = 1) - \Pr(A_u = 1)\Pr(A_v = 1).$$

Now first consider the case when vertices u and v are at a distance of at least $\frac{2(a+b)\ln n}{n}$ apart (happens with probability $1 - \frac{4(a+b)\ln n}{n}$). Then the region in $[0,1]$ that is within distance $\frac{(a+b)\ln n}{n}$ from both u and v is the empty-set. In this case, $\Pr(A_u = 1 \cap A_v = 1) = n(n-1)\dots(n-(4k-1))\Big(\frac{\epsilon \ln n}{n}\Big)^{4k}\Big(1 - 4k\epsilon\frac{\ln n}{n}\Big)^{n-4k} = c_2 n^{-4k\epsilon}(\epsilon \ln n)^{4k}$, where c_2 is a constant.

In all other cases, $\Pr(A_u = 1 \cap A_v = 1) \leq \Pr(A_u = 1)$. Therefore,

$$\begin{aligned}\Pr(A_u = 1 \cap A_v = 1) \leq &\Big(1 - \frac{4(a+b)\ln n}{n}\Big)c_2 n^{-4k\epsilon}(\epsilon \ln n)^{4k} \\ &+ \frac{4(a+b)\ln n}{n}c_1 n^{-2k\epsilon}(\epsilon \ln n)^{2k}\end{aligned}$$

and

$$\begin{aligned}\mathrm{Var(A)} &\leq c_1 n^{1-2k\epsilon}(\epsilon \ln n)^{2k} + \binom{n}{2}\Big(\Pr(A_u = 1 \cap A_v = 1) - \Pr(A_u = 1)\Pr(A_v = 1)\Big) \\ &\leq c_1 n^{1-2k\epsilon}(\epsilon \ln n)^{2k} + c_3 n^{1-2k\epsilon}(\ln n)^{2k+1} \leq c_4 n^{1-2k\epsilon}(\ln n)^{2k+1}\end{aligned}$$

where c_3, c_4 are constants. Again invoking Chebyshev's inequality, with probability at least $1 - \frac{1}{\ln n}$

$$A > c_1 n^{1-2k\epsilon}(\epsilon \ln n)^{2k} - \sqrt{c_4 n^{1-2k\epsilon}(\ln n)^{2k+2}}.$$ ◀

B Missing Proofs of Section 5

▶ **Lemma 23.** *For any two vertices $u, v \in V_i : (u, v) \in E, i = 1, 2$ belonging to the same cluster with $d_L(X_u, X_v) = x$, the count of common neighbors $C_{u,v} \equiv |\{z \in V : (z, u), (z, v) \in E\}|$ is a random variable distributed according to $\text{Bin}(\frac{n}{2} - 2, 2r_s - x)$ when $r_s \geq x > 2r_d$ and according to $\text{Bin}(\frac{n}{2} - 2, 2r_s - x) + \text{Bin}(\frac{n}{2}, 2r_d - x)$ when $x \leq \min(2r_d, r_s)$, where $\text{Bin}(n, p)$ is a binomial random variable with mean np.*

Proof. Without loss of generality, assume $u, v \in V_1$. For any vertex $z \in V$, let $\mathcal{E}_z^{u,v} \equiv \{(u, z), (v, z) \in E\}$ be the event that z is a common neighbor. For $z \in V_1$,

$$\begin{aligned}&\Pr(\mathcal{E}_z^{u,v}) = \Pr((z, u) \in E, (z, v) \in E)\\&= 2r_s - x,\end{aligned}$$

since $d_L(X_u, X_v) = x$. For $z \in V_2$, we have,

$$\begin{aligned}&\Pr(\mathcal{E}_z^{u,v}) = \Pr((z, u), (z, v) \in E)\\&= \begin{cases} 2r_d - x & \text{if } x < 2r_d \\ 0 & \text{otherwise} \end{cases}.\end{aligned}$$

Now since there are $\frac{n}{2} - 2$ points in $V_1 \setminus \{u, v\}$ and $\frac{n}{2}$ points in V_2, we have the statement of the lemma. ◀

In a similar way, we can prove.

▶ **Lemma 24.** *For any two vertices $u \in V_1, v \in V_2 : (u, v) \in E$ belonging to different clusters with $d_L(X_u, X_v) = x$, the count of common neighbors $C_{u,v} \equiv |\{z \in V : (z, u), (z, v) \in E\}|$ is a random variable distributed according to $\text{Bin}(n - 2, 2r_d)$ when $r_s > 2r_d$ and according to $\text{Bin}(n - 2, \min(r_s + r_d - x, 2r_d))$ when $r_s \leq 2r_d$ and $x \leq r_d$.*

Proofs of Lemma 20 and Lemma 21

Proof of Lemma 20. Here we will use the fact that for $a \geq 1$, the number of edges in $\text{GBM}(r_s \equiv \frac{a \ln n}{n}, r_d \equiv \frac{b \ln n}{n})$ is $O(n \ln n)$ with probability $1 - \frac{1}{n^{\Theta(1)}}$. Consider any vertex $u \in V_1$ (symmetrically for $u \in V_2$), since the vertices are thrown uniformly at random in $[0, 1]$, the probability that a $v \in V_1$, $v \neq u$, is a neighbor of u is $\frac{a \ln n}{n}$, and for $v \in V_2$, the corresponding probability is $\frac{b \ln n}{n}$. Therefore, the expected degree of u is $\frac{(a+b)}{2} \ln n$. By a simple Chernoff bound argument, the degree of u is therefore $O(\ln n)$ with probability $1 - \frac{1}{n^c}$ for $c \geq 2$. By union bound over all the vertices, the total number of edges is $O(n \ln n)$ with probability $1 - \frac{1}{n}$.

Let Z denote the random variable that equals the number of common neighbors of two nodes $u, v \in V : (u, v) \in E$ such that u, v are from different parts of the GBM. Using Lemma 24, we know that Z is sampled from the distribution $\text{Bin}(n - 2, 2r_d)$, where $r_d = \frac{b \ln n}{n}$. Therefore,

$$\Pr(Z \geq nE_S) \leq \sum_{i=nE_S}^{n} \binom{n}{i} (2r_d)^i (n - 2r_d)^{n-i} \leq \exp\Big(-nD\Big((2b + t_1)\frac{\ln n}{n} \| \frac{2b \ln n}{n}\Big)\Big),$$

where $D(p\|q) \equiv p \ln \frac{p}{q} + (1 - p) \ln \frac{1-p}{1-q}$ is the KL divergence between Bernoulli(p) and Bernoulli(q) distributions. It is easy to see that,

$$nD(\frac{\alpha \ln n}{n} || \frac{\beta \ln n}{n}) = \Big(\alpha \ln \frac{\alpha}{\beta} + (\alpha - \beta)\Big) \ln n - o(\ln n).$$

Therefore $\Pr(Z \geq nE_S) \leq \frac{1}{n(\ln n)^2}$ because $(2b+t_1)\ln\frac{2b+t_1}{2b} - t_1 > 1$. Similarly, we have that

$$\begin{aligned}\Pr(Z \leq nE_D) &\leq \sum_{i=0}^{nE_D} \binom{n}{i} (2r_d)^i (n-2r_d)^{n-i} \leq \exp(-nD((2b-t)\frac{\ln n}{n} \| \frac{2b\ln n}{n}))\\ &\leq \frac{1}{n(\ln n)^2}.\end{aligned}$$

So all of the inter-cluster edges will be removed by Algorithm 1 with probability $1 - O(\frac{n\ln n}{n(\ln n)^2}) = 1 - o(1)$, as with probability $1 - o(1)$ the total number of edges in the graph is $O(n\ln n)$. ◀

Proof of Lemma 21. Let Z be the number of common neighbors of u, v. Recall that, u and v are in the same cluster. We know from Lemma 24 that Z is sampled from the distribution $\mathrm{Bin}(\frac{n}{2}-2, 2r_s - x) + \mathrm{Bin}(\frac{n}{2}, 2r_d - x)$ when $x \leq 2r_d$, and from the distribution $\mathrm{Bin}(\frac{n}{2}-2, 2r_s - x)$ when $x \geq 2r_d$. We have,

$$\begin{aligned}&\Pr(Z \leq nE_S)\\ &= \begin{cases} \sum_{i=0}^{nE_S} \binom{\frac{n}{2}-2}{i} (2r_s - x)^i (1-2r_s+x)^{\frac{n}{2}-i-2} \\ \times \sum_{j=0}^{nE_S - i} \binom{\frac{n}{2}}{j} (2r_d - x)^j (1-2r_d+x)^{\frac{n}{2}-j} & \text{if } x \leq 2r_d \\ \sum_{i=0}^{nE_s} \binom{\frac{n}{2}-2}{i} (2r_s - x)^i (1-2r_s+x)^{\frac{n}{2}-i} & \text{otherwise} \end{cases}\\ &\leq e^{-\frac{n}{2}D(2E_S || \frac{(2a-\theta)\ln n}{n})} \text{ since } 2a - \theta \geq 4b + 2t_1\\ &\leq e^{-\frac{n}{2}D(\frac{(4b+2t_1)\ln n}{n} || \frac{(2a-\theta)\ln n}{n})} \leq \frac{1}{n\ln^2 n},\end{aligned}$$

because of Condition 1 of this lemma. Therefore, this edge will not be deleted with high probability.

Similarly, let us find the probability of $Z \geq nE_D = (2b - t_2)\ln n$. Let us just assume the worst case when $\theta \leq 2b$: that the edge is being deleted (see Condition 2, this is prohibited if that condition is satisfied). Otherwise, $\theta > 2b$ and,

$$\begin{aligned}\Pr(Z \geq nE_D) &= \sum_{i=nE_D}^{n} \binom{\frac{n}{2}-2}{i} (2r_s - x)^i (1-2r_s+x)^{\frac{n}{2}-i-2}\\ &\leq e^{-\frac{n}{2}D(2E_D \| \frac{(2a-\theta)\ln n}{n})} \text{ if } 2a - \theta \leq 4b - 2t_2\\ &= e^{-\frac{n}{2}D(\frac{(4b-2t_2)\ln n}{n} \| \frac{(2a-\theta)\ln n}{n})} \leq \frac{1}{n\ln^2 n}\end{aligned}$$

because of Condition 2 of this lemma. ◀