Testing Thresholds for High-Dimensional Sparse Random Geometric Graphs

Siqi Liu
slu18@berkeley.edu
University of California, Berkeley
United States

Sidhanth Mohanty
sidhanth@cs.berkeley.edu
University of California, Berkeley
United States

Tselil Schramm
tselil@stanford.edu
Stanford University
United States

ABSTRACT
The random geometric graph model $\text{Geo}_d(n, p)$ is a distribution over graphs in which the edges capture a latent geometry. To sample $G \sim \text{Geo}_d(n, p)$, we identify each of our $n$ vertices with an independently and uniformly sampled vector from the $d$-dimensional unit sphere $S^{d-1}$, and we connect pairs of vertices whose vectors are “sufficiently close,” such that the marginal probability of an edge is $p$. Because of the underlying geometry, this model is natural for applications in data science and beyond.

We investigate the problem of testing for this latent geometry, or in other words, distinguishing an Erdős-Rényi graph $G(n, p)$ from a random geometric graph $\text{Geo}_d(n, p)$. It is not too difficult to show that if $d \to \infty$ while $n$ is held fixed, the two distributions become indistinguishable; we wish to understand how fast $d$ must grow as a function of $n$ for indistinguishability to occur.

When $p = \frac{d}{n}$ for constant $\sigma$, we prove that if $d \geq \text{poly}(\log(n))$, the total variation distance between the two distributions is close to 0; this improves upon the best previous bound of Brennan, Bresler, and Nagaraj (2020), which required $d \gg n^{3/2}$, and further our result is nearly tight, resolving a conjecture of Bubeck, Ding, Eldan, & Rácz (2016) up to logarithmic factors. We also obtain improved upper bounds on the statistical indistinguishability thresholds in $d$ for the full range of $p$ satisfying $\frac{1}{n} \leq p \leq \frac{1}{2}$, improving upon the previous bounds by polynomial factors.

Our analysis uses the Belief Propagation algorithm to characterize the distributions of (subsets of) the random vectors conditioned on producing a particular graph. In this sense, our analysis is connected to the “cavity method” from statistical physics. To analyze this process, we rely on novel sharp estimates for the area of the intersection of a random sphere cap with an arbitrary subset of $S^{d-1}$, which we prove using optimal transport maps and entropy-transport inequalities on the unit sphere. We believe these techniques may be of independent interest.

1 INTRODUCTION
The study of random graphs has been incredibly influential, not only in modeling applications, but also in the development of algorithms and in the study of mathematics. For example, consider the simple Erdős-Rényi random graph model $G(n, p)$, in which $n$ nodes are each connected independently with probability $p$. This model was introduced by Erdős and Rényi in 1959 [13], and since then its study has blossomed. To list only a few of the many fruits of this line of research: $G(n, p)$ graphs have been used to demonstrate the existence of many combinatorial objects via the probabilistic method (e.g. [1]); they have been used in algorithm design both as a benchmark and as a starting point for analysis in worst-case or semirandom settings (e.g. [2, 15]); they have been useful for proving conditional and unconditional lower bounds in complexity theory (e.g. [14, 19, 21, 31, 32]); and finally $G(n, p)$ graphs are a common model in statistical physics [26, 28] and network science [8, 20]. Though deep questions remain, decades of intensive study of the $G(n, p)$ distribution have been rewarded with a rich understanding and many unforeseen insights.

Our primary object of study is the random geometric graph model on the sphere, $\text{Geo}_d(n, p)$. In this model, a graph is sampled by choosing $n$ vectors $v_1, \ldots, v_n$ uniformly at random from the $d$-dimensional sphere $S^{d-1}$, identifying $v_i$ with vertex $i$, then connecting each pair $i, j$ whose vectors $v_i, v_j$ have inner product exceeds a threshold $\tau(p)$, chosen so that $\Pr[\langle v_i, v_j \rangle > \tau(p)] = p$. A compelling aspect of this model is that the graph structure is derived from a geometric representation of the vertices; this makes it suitable for modeling applications in, say, data science, where...
we think of network nodes as representable by a feature vector in some $d$-dimensional space, with neighboring nodes sharing similar features.

Though random geometric graphs on the sphere (and also on other domains, such as $[0,1]^d$) have been studied extensively (see the monograph [29]), the focus has been on the low-dimensional setting, where we think of $d$ is fixed as $n \rightarrow \infty$. However, the high-dimensional setting, in which $d \rightarrow \infty$ as a function of $n$, is still poorly understood. In low dimensions, one can understand $\text{Geo}_d(n,p)$ as a fine random discretization of $S^{d-1}$, but we no longer expect $\text{Geo}_d(n,p)$ to behave in this way in the high-dimensional setting.

This opens an intriguing possibility: what are the properties of $\text{Geo}_d(n,p)$ in the high-dimensional setting? And, what are the undiscovered applications of these random objects in algorithm design and in mathematics? One reason to study this question is that the high-dimensional setting is a more faithful model for graphs arising from modern high-dimensional data (e.g. high-dimensional feature vectors). Thus, high-dimensional geometric random graphs could constitute a more useful benchmark for algorithmic methods, or a more useful starting point for designing algorithms for the semirandom setting. Another reason to study this distribution is that it may yield graphs with properties that we have not yet realized are possible; this has historical precedent, for example, Erdős-Rényi graphs witnessing the existence of certain Ramsey graphs [1].

Devroye, György, Lugosi, and Udina [10] made the first exploration of $\text{Geo}_d(n,p)$ in the high-dimensional regime. They study the classic property of the chromatic number in random geometric graphs, and also raise the following fundamental question: for which $d = d(n,p)$ is it possible to test for the presence of an underlying geometry? When can one distinguish $\text{Geo}_d(n,p)$ from $G(n,p)$?

For intuition, one can see that if $d \rightarrow \infty$ for fixed $p$, the $\text{Geo}_d(n,p)$ distribution approaches $G(n,p)$: the vectors $\mathbf{v}_1, \ldots, \mathbf{v}_n$ are effectively mutually orthogonal, $r(p)$ is very small, and conditioning on the presence of the edge $(i,j)$ (or equivalently on $|\langle \mathbf{v}_i, \mathbf{v}_j \rangle| > r(p)$) has little impact on the probability that the edges $(i,k)$ and/or $(j,k)$ are present. How fast must $d$ grow as a function of $n$ to realize this limiting behavior? In [10] the authors observe that if $d \gg \exp(n^2)$, a central limit theorem implies that $\text{Geo}_d(n,p)$ and $G(n,p)$ are indistinguishable; they also note that this bound is likely far from tight.

Determining the asymptotic threshold in $d$ when geometric graphs become indistinguishable from Erdős-Rényi graphs is a most basic question that we must address if we wish to make a serious study of high-dimensional random geometric graphs. Bubeck, Ding, Eldan, and Rácz [5] were the first to tackle this question, showing that in dense graphs (when $p = \Theta(1)$), the threshold occurs at $d \gg n^2$; when $d \gg n^3$ the total variation distance goes to 0, and when $d \ll n^3$ the “signed triangle count” provides a good test statistic. But in the arguably more interesting sparse case $p = \Theta\left(\frac{1}{n}\right)$, [5] are only able to establish that signed triangle counts distinguish when $d = O(\log^3 n)$. They conjecture that when $d = \Omega(\log^3 n)$, the distributions are indistinguishable.

The current best bound, due to Brennan, Bresler, and Nagaraj [4], asserts that in the regime $p = \Theta\left(\frac{1}{n}\right)$, $d_{\text{TV}}(\text{Geo}_d(n,p), G(n,p)) \rightarrow 0$ so long as $d \gg n^{3/2}$. In essence, their bound relies on the fact that independent random vectors $\mathbf{v}_i, \mathbf{v}_j \sim S^{d-1}$ have $|\langle \mathbf{v}_i, \mathbf{v}_j \rangle| = O\left(\frac{1}{\sqrt{d}}\right)$ with high probability; when $d \gg n^{3/2}$, these inner products are small enough relative to $n$ that $\mathbf{v}_i$ has negligible projection (of order $\approx \sqrt{\sum_{j \neq i} |\langle \mathbf{v}_i, \mathbf{v}_j \rangle|^2} = O(\sqrt{n/d})$) into span$\{\mathbf{v}_j\}_{j \neq i}$, which is enough to guarantee approximate independence of edges. This argument is carried out with technical sophistication in [4], but clearly, this technique cannot be extended to $d < n$, much less to $d = \text{poly} \log n$, as the vectors in $\{\mathbf{v}_j\}_{j \neq i}$ then span all of $\mathbb{R}^d$.

In this work, we come close to closing this gap in the sparse regime, confirming the conjecture of [5] up to polylogarithmic factors: we show that if $p = \Theta\left(\frac{1}{n}\right)$, the total variation distance goes to zero when $d \gg \log^{36} n$. We also give a separate improved bound on the total variation distance for the entire parameter regime $p \ll 1$, showing that the total variation distance goes to 0 when $d \gg n^3 p^2$, improving upon previous bounds by polynomial factors. Our proof relies on a number of novel technical contributions. Crucially, we must understand, for “typical” $G \sim \text{Geo}_d(n,p)$, the distribution of (subsets of the) vectors $\mathbf{v}_1, \ldots, \mathbf{v}_n$ sampled from $S^{d-1}$ conditioned on giving a vector embedding for $G$. In order to understand this complicated conditional distribution, we use the Belief Propagation (BP) algorithm, where our variables are the vertices $[n]$ and their “labels” are vectors in $S^{d-1}$. To analyze BP, we rely on a (to our knowledge) novel concentration inequality, which we prove using optimal transport, for the area of the intersection of a random spherical cap with any subset $L \subseteq S^{d-1}$. We also demonstrate a coupling of $G \sim \text{Geo}_d(n,p)$ and $G_* \sim G(n,p + o(p))$ which produces $G \subseteq G_*$ (meaning $G$ is a subgraph of $G_*$) with high probability. We feel that these techniques may find other applications, and be of independent interest.

### 1.1 Our Results

Our main result is an indistinguishability result for sparse Random Geometric Graphs and Erdős-Rényi graphs when the dimension $d$ exceeds $\text{poly} \log n$.

**Theorem 1.1.** For any fixed constant $\alpha > 0$, if $d = \Omega(\log^{36} n)$, then

$$
\lim_{n \to \infty} d_{\text{TV}}(\text{Geo}_d(n, \frac{n^\alpha}{n}), G(n, \frac{n^\alpha}{n})) = 0.
$$

Our result settles the conjecture of [5] up to logarithmic factors, an exponential improvement over the previous bound of [4], which required $d \gg n^{3/2}$. We remark that we have not made an effort to optimize the logarithmic factors; it is possible that our current proofs in combination with chaining-style arguments will yield $\log^3 n$, matching their conjecture. We also obtain an improved result for general $p = \Omega\left(\frac{1}{n}\right)$:

**Theorem 1.2.** For any fixed constant $\alpha > 0$, if $\frac{n^\alpha}{n} < p < \frac{1}{2}$ and $d = \tilde{\Omega}(p^2 n^3)$,

$$
\lim_{n \to \infty} d_{\text{TV}}(\text{Geo}_d(n, p), G(n, p)) = 0.
$$

This improves by polynomial factors (in $p$ and $n$) on the previous bound of [4], which required $d \gg \min\{pn^3 \log \frac{1}{p}, p^2 n^{7/2} \text{poly} \log n\}$.
and $d \gg n \text{poly log } n$. However, this result is not tight (at least for small $p$) since in particular it does not recover Theorem 1.1. Given that we have come close to establishing the conjecture of [5] in the sparse case, it is tempting to interpolate between the upper and lower bounds of [5] in the $p = \Theta(1)$ regime and conjecture for the $p = \Theta\left(\frac{1}{n}\right)$ regime and speculate that for all $p \leq \frac{1}{2}$, the testing threshold occurs at $d \asymp (nH(p))^3 = O\left(n^3 p^3 \log^3 \frac{1}{p}\right)$, for $H(p)$ the binary entropy function. In Appendix A of the full version, we show that the "signed triangle count" test statistic analyzed by [5] in the $p = \Theta(1)$ regime can in fact distinguish whenever $d \ll (nH(p))^3$ for all $p = \Omega\left(\frac{1}{n}\right)$, establishing that the testing threshold occurs at some $d = \Omega\left((nH(p))^3\right)$. If $d \gg (nH(p))^3$ is indeed the threshold, our Theorem 1.2 is tight up to a factor of $O(p)$.

Relaxed stochastic dominance by Erdős-Rényi graphs. En route to proving Theorem 1.1 and Theorem 1.2, we establish a result which may be of independent interest.

Proposition 1.3. For any constant $\alpha > 0$ there exist constants $C_1, C_2 > 0$ such that if $\frac{2}{n} < p < \frac{1}{4}$ and $d \geq C_1 (n^2 p^2 + \log^3 n) \log^4 n$, for any $\varepsilon \geq C_2 \sqrt{\frac{1}{n}} (np + \log n) \log^4 n$, one can simultaneously sample $G_\varepsilon \sim G(n, (1 - \varepsilon)p)$, $G \sim Geo(n, p),$ and $G_{\varepsilon} \sim G(n, (1 + \varepsilon)p)$ in a correlated manner so that with probability at least $1 - n^{-\Omega(\log n)}$, $G_\varepsilon \subseteq G \subseteq G_{\varepsilon}$.

The notation $G \subseteq G_\varepsilon$ means that $G$ is a subgraph of $G_\varepsilon$. We find it intriguing that this coupling succeeds whenever $d = \Omega(p^2 n^2)$, which is well below the speculative "interpolated" threshold $d = \Omega(p^2 n^2)$ for large $p$.

1.2 Prior and Related Work

Prior Works. Random geometric graphs in fixed dimension are a well-studied model, with connections to Poisson processes and continuum percolation. The tools used to study fixed-dimensional random graphs are of a very different flavor from ours; for example, in low dimensions it is often useful to compare $Geo_d(n, p)$ to a random process on an appropriate low-dimensional infinite lattice which discretizes the space. We refer the reader to the survey [35] and the monograph [29].

The study of random geometric graphs in high dimension was initiated by [10]. The main result of [10] is a bound on the clique number of $Geo_d(n, p)$: for example in the dense case $p = \Theta(1)$, they show that the clique numbers of $Geo_d(n, p)$ and $G(n, p)$ become indistinguishable when $d = \Omega(\log^3 n)$. The authors of [10] also note that if $d \rightarrow \infty$ fast enough as a function of $n$ (which require $d \gg \exp(\frac{n^2}{\log^2 n})$), $Geo_d(n, p)$ and $G(n, p)$ become indistinguishable. This naturally raises the question: for which $d$ is it possible to test for the underlying geometry?

The authors of [5] are the first to directly study the testing phase transition in $d$. They show that at any density, if $d \gg n^3$, then the total variation distance between $Geo_d(n, p)$ and $G(n, p)$ goes to zero with $n$, and conversely, if $d \ll n^3$, then in the dense regime $p = \Theta(1)$, the signed triangle count statistic furnishes a hypothesis test between the distributions. However in the sparse regime $p = \Theta\left(\frac{1}{n^2}\right)$, their results are not as conclusive; they are able to show that triangle counts furnish a test if and only if $d \ll \log^3 n$, and they conjecture that if $d \gg \log^3 n$ the total variation distance goes to zero. Briefly, their bound on the TV distance is via a reduction to the indistinguishability of Wishart matrices and matrices from the Gaussian Orthogonal Ensemble (GOE); the idea is that one can obtain a sample from $G(n, p)$ by thresholding the off-diagonal entries of a GOE matrix $B$ with independent entries sampled from $N(0, \frac{1}{n})$ at threshold $\approx \tau(p)$, and similarly obtain a sample from $Geo_d(n, p)$ by thresholding the off-diagonal entries of a Wishart matrix $AA^T$ at $\approx \tau(p)$, where $A \in \mathbb{R}^{n \times d}$ with rows sampled independently from $N(0, \frac{1}{n}I_d)$. Hence, if $B$ and $AA^T$ are indistinguishable from their off-diagonal entries, one can conclude that $G(n, p)$ and $Geo_d(n, p)$ are indistinguishable as well. It makes sense, then, that the result may not be tight in the sparse regime, as thresholding at a higher value $\tau(p)$ (corresponding to a sparser graph) intuitively reveals less information about the original GOE or Wishart matrix. The proof in [5] directly compares the Wishart and GOE densities to obtain the TV bound. Independently, [22] obtain the same bounds on the TV distance of Wishart and GOE matrices.

Following the work of [5], [4] study the question of detecting underlying geometry in greater generality. The authors show that for any $\frac{\log n}{n} < p < \frac{1}{2}$, if $d \gg \min\left\{n^2 \log \frac{1}{p}, p^2 n^{7/2} \text{poly log } n\right\}$ and $d \gg n \log^4 n$, then $d_{TV}(Geo_d(n, p), G(n, p)) \to 0$. In particular, they match the bound appearing in [5] for dense graphs, and in sparse graphs with $p = \Theta\left(\frac{1}{n}\right)$ they improve the bound to $d \gg n^{3/2} \text{poly log } n$. At a high level, their proof applies information-theoretic inequalities to reduce the question to bounding the Chi-square divergence of the marginal of a single edge in $Geo_d(n, p)$ vs. $G(n, p)$; they then bound this Chi-squared divergence by showing that if one conditions $v_1, \ldots, v_n \sim \left(\mathbb{S}^{d-1}\right)^\otimes n$ on the event that they produce a "typical" random geometric graph $G$ excluding edge $(n-1, n)$, then the vectors $v_n, v_{n-1}$ remain sufficiently independent that $Pr\left(v_{n-1}, v_n > r\right) \approx p$. They achieve this by showing that $v_{n-1}$ and $v_n$ each have a small projection onto span$(v_1, \ldots, v_{n-2})$, even after conditioning on their adjacency into $[n-2]$. Their argument is essentially oblivious to the particular choice of $G$, and merely uses properties that hold with high probability over independently sampled vectors in $\mathbb{S}^{d-1}$. This argument relies on $v_1, \ldots, v_{n-2}$ not spanning the entirety of $\mathbb{R}^d$, and the authors of [4] explicitly state that improving their bound to any $d < n$ requires a different approach.

Techniques. We describe our techniques in detail in Section 2, but here we discuss some connections in the literature. To improve upon the bounds of [4], we draw upon tools from several areas. At the heart of the proofs of both Theorem 1.1 and Theorem 1.2 are new sharp concentration of measure results for intersections.
of random spherical caps with arbitrary subsets of $S^{d-1}$; to prove these bounds, we make use of optimal transport inequalities in the Wasserstein metric (see e.g. [34]). To our knowledge this is the first application of optimal transport in this context, and may be of independent interest. These tail bounds (in combination with Pinsker’s inequality and some simple arguments) are enough to yield Theorem 1.2, proving that the total variation distance goes to zero if $d \gg n^p$.

Then, to break the $d \times n$ barrier for sparse graphs and prove Theorem 1.1, we must answer the following question: if $v_1, \ldots, v_n$ are sampled uniformly conditioned on producing a vector embedding of a fixed graph $G$, how strong are the correlations in the marginal distributions on $\{v_i\}_{i \in S}$ for small subsets $S \subseteq [n]$? This is similar to establishing decay of correlations, as in the analysis of Gibbs sampling (c.f. [11, 36]). To achieve this, we carry out a rigorous analysis in the style of the “cavity method” from statistical physics (c.f. [25]): we compute the marginal over the depth $\log \log n$ neighborhood of vertices in $G$ via the Belief Propagation (BP) algorithm, under arbitrary boundary conditions on the rest of the graph. The cavity method has been previously used to compute solution geometry phase transitions for a number of prominent discrete spin systems, such as coloring, Ising and Potts models, and more (c.f. [7, 9, 27, 33]): cavity-style arguments have also been employed in a similar way to establish decay of correlation properties in the analysis of Glauber dynamics, as in [18]. In our instantiation of Belief Propagation, the variables are the vertices in the graph, and their “labels” or “assignments” are vectors in $S^{d-1}$, the constraints are that vertices corresponding to edges have inner product at least $\tau$, and hence the “messages” passed from variable to variable are convolutions of marginal distributions with spherical caps; the concentration of measure for intersections of sets with random spherical caps is again useful in the analysis of this BP.

**Applications of high-dimensional geometric graphs in Theoretical CS.** A line of work including [16, 17] has utilized a distribution similar to $\text{Geo}_q(n, p)$, to obtain integrality gaps for semidefinite relaxations of max-cut or graph coloring. In their setting, a graph $G$ is sampled by first placing $n$ vertices in a “regular” configuration on a $d$-dimensional sphere, after which $n' = \exp(\Theta(d))$ vertices are randomly subsampled independently with some probability $q$, and the sample is a graph induced on these $n'$ vertices, in which vertices are connected if their inner product lies in some range of distances $(\tau_1, \tau_2)$ (as opposed to the $\text{Geo}_q(n, p)$ case in which the connectivity criteria is inner product at least $\tau$). The resulting $G$ comes with a natural embedding into $S^{d-1}$, which is utilized in constructing the semidefinite programming certificate. Similar constructions are also used in [23] to give the optimal approximation ratio of constraint satisfaction problems assuming the unique games conjecture. Though these graphs are not sampled from $\text{Geo}_q(n, p)$, they are sampled from a distribution over a high-dimensional sphere which is qualitatively similar. We feel that this points to the promise of high-dimensional random geometric graphs for applications in theoretical computer science.

**Phase transitions between Wishart and GOE matrices.** The works mentioned above are those most closely related to our results. We mention as well additional works concerning the phase transition between Wishart and GOE matrices: the work of [30] studies the phase transition between Wishart and GOE at a higher resolution in the dense regime, deriving the expression for the total variation distance as a function of $c = \sqrt{d/n^3}$ when $\lim_n d/n^3$ is finite. The result [6] generalizes the $d \gg n^3$ bound to Wishart matrices in vectors drawn from log-concave measures, and in [12] the authors study the phase transition for Wishart matrices in vectors drawn from non-isotropic Gaussian measures, drawing conclusions for hypothesis testing dense geometric graphs derived from these ensembles as well. Other similar questions that have been studied are hypothesis testing in noisy Wishart matrices [24], in which the entries are with some probability independently resampled from the Gaussian distribution, and masked hypothesis testing between Wishart and GOE matrices, in which only a subset of the matrix’s entries are revealed [3]. In both of these cases, it has been demonstrated the presence of noise or masking can shift the threshold in interesting ways.

## 2 TECHNICAL OVERVIEW

### 2.1 Relative Entropy Tensorization

Our goal is to determine $d$ at which the total variation distance $d_{TV}(G(n, p), \text{Geo}_q(n, p))$ goes to zero as $n \to \infty$. Like the authors of [4], we relate the TV distance between these two distributions to their relative entropy $D(\text{Geo}_q(n, p) || G(n, p))$ (Definition 3.1 of the full version) via Pinsker’s inequality (Theorem 3.3 of the full version), and then apply the tensorization of the relative entropy (Claim 8.1 of the full version). Roughly, the tensorization says that given a decomposition of $G(n, p)$ as a product distribution, we can reduce the problem of bounding $D(\text{Geo}_q(n, p) || G(n, p))$ to bounding the relative entropy over (potentially simpler) distributions with smaller support.

$G(n, p)$ is conveniently a product distribution over edges. However, unlike [4], we do not use this straightforward decomposition of $G(n, p)$ by edge. Instead, let $\mu_i$ be the distribution of vertex $i$’s edges to $[t-1]$. Similarly, let $v_i$ be the marginal distribution of vertex $i$’s edges to $[t-1]$ over the graph being sampled from $\text{Geo}_q(n, p)$. Our bound via tensorization now becomes

$$D(\text{Geo}_q(n, p) || G(n, p)) = \sum_{t=1}^{n} \frac{E}{\mathcal{G}_{t-1} \sim \text{Geo}_q(t-1, p)} \left[ D(\nu_t (\cdot | \mathcal{G}_{t-1}) \parallel \mu_t) \right] \leq n \cdot \frac{E}{\mathcal{G}_{n-1} \sim \text{Geo}_q(n-1, p)} \left[ D(\nu_{n-1} (\cdot | \mathcal{G}_{n-1}) \parallel \mu_{n-1}) \right]$$

where the final inequality follows after applying the chain rule for relative entropy (Claim 8.2 of the full version).

The coupling view. The tensorization inequality reduces bounding the TV distance to comparing the probability distribution of the neighborhood of the “final” vertex in $G(n, p)$ and $\text{Geo}_q(n, p)$. Specifically, we study $\mathbb{E}_{\mathcal{G}_{n-1} \sim \text{Geo}_q(n-1, p)} [D(\nu_n (\cdot | \mathcal{G}_{n-1}) \parallel \mu_n)]$ by considering the following scenario: we already have a graph $\mathcal{G}_{n-1}$ sampled on $n-1$ vertices, and we want to incorporate vertex $n$ into our...
graph. By the definition of the Erdős–Rényi distribution, \( p \) will sample the neighbor set \( S \subseteq [n - 1] \) with probability \( p^{S}(1 - p)^{n - |S|} \). For a random geometric graph, we can sample a vector \( \mathbf{v}_n \sim \rho \), and take its dot products to vectors \( \mathbf{v}_1, \ldots, \mathbf{v}_{n-1} \) sampled uniformly from \( S^{d-1} \) conditioned on producing \( G_{n-1} \), to determine the neighbors of \( n \) in \( G \) (which we denote by \( N_G(n) \)). Our goal now is to compare \( \Pr_{G \sim \text{Geo}_d(n, p)}[N_G(n) = S] \) and \( \Pr_{G \sim \text{Geo}_d(n, p)}[N_G(n) = S] \) for \( S \subseteq [n - 1] \).

### 2.2 A Geometric Interpretation of Neighborhood Probability

For \( G \sim \text{Geo}_d(n, p) \), if vertex \( i \) is associated to a (random) vector \( \mathbf{v}_i \), and \((i, j)\) is an edge, we consequently know that \((\mathbf{v}_i, \mathbf{v}_j) \geq \tau \). On the sphere \( S^{d-1} \), the locus of points where \( \mathbf{v}_j \) can be, conditioned on \((i, j)\) being an edge, is a sphere cap centered at \( \mathbf{v}_i \) with a \( p \) fraction of the sphere’s surface area, which we denote by \( \text{cap}(\mathbf{v}_i) \). Similarly, if we know that \( i \) and \( j \) are not adjacent, the locus of points where \( \mathbf{v}_j \) can fall is the complement of a sphere cap, which we call an “anti-cap,” with measure \( 1 - p \).

Equipped with this geometric picture, we can view the probability that vertex \( n \)’s neighborhood is exactly equal to \( S \subseteq [n - 1] \) as the measure \( \rho(L_S) \), where \( L_S \subseteq S^{d-1} \) is a random set defined as:

\[
L_S := \left( \bigcap_{i \in S} \text{cap}(\mathbf{v}_i) \right) \cap \left( \bigcap_{i \notin S} \text{cap}(\mathbf{v}_i) \right)
\]

To show that the distance between \( G \sim \text{Geo}_d(n, p) \) and \( G(n, p) \) is small, we must show that \( \rho(L_S) \) concentrates around \( p^{|S|}(1 - p)^{|n - 1 - |S|}| \), which is the probability that \( n \)’s neighborhood is equal to \( S \) under the Erdős–Rényi model.

**Optimal transport.** The backbone of our result is (to our knowledge) novel application of optimal transport. In Section 4 of the full version, we prove for a generic distribution \( \mathcal{L} \) supported on \( S^{d-1} \), and \( z \) sampled independently and uniformly at random over \( S^{d-1} \), that

\[
\Pr_{x \sim \mathcal{L}}[(x, z) \geq \tau] \geq (1 + \epsilon) \cdot p \quad \forall \epsilon \ll 1 \quad \text{for } z \subseteq S^d - L \text{ with high probability over } z.
\]

In other words, how tightly the random variable \( X_{\mathcal{L}}(z) = \Pr_{x \sim \mathcal{L}}[(x, z) \geq \tau] \) concentrates is directly related to the maximum value of its relative density, \( \|\mathcal{L}\|_{\infty} \).

To give some intuition for this result, first consider the case when \( \mathcal{L} = \rho \), the uniform distribution over \( S^{d-1} \). Then, the variable \( X_{\rho}(z) = \Pr_{x \sim \rho}[(x, z) \geq \tau] = p \) deterministically. Now, when \( \mathcal{L} \neq \rho \), we can work with a transport map \( \mathcal{D} \) between \( \mathcal{L} \) and \( \rho \), and we can couple \( y \sim \mathcal{L} \) and \( x \sim \rho \) according to \( \mathcal{D} \), so that

\[
X_{\mathcal{L}}(z) = \Pr_{y \sim \mathcal{L}}[(y, z) \geq \tau] = \Pr_{(x, y) \sim (\rho, \mathcal{L})}[(x, z) \geq \tau - (e, z)].
\]

The smaller \( \|\mathcal{L}\|_{\infty} \) is, the smaller the average of the transport distance \( |\mathcal{L}| = ||x - y|| \); further when \( z \sim \rho \) the quantity \( (e, z) \) concentrates tightly around \( e \). In this way, we translate the concentration of transport distance into tail bounds on \( X_{\rho}(z) - X_{\mathcal{L}}(z) \).

To analyze \( \rho(L) \) for \( L \) the intersection of caps and anti-caps defined above, we will apply the above in sequence inside a martingale concentration argument, building up \( L \) one cap at a time (Lemma 5.1 and Corollary 6.1 of the full version). Using this approach, our transport result alone is enough to conclude Theorem 1.2. (The proof is assembled in Section 8.1.1 of the full version.)

**The need to resample vectors.** In the general \( p \) setting, we can think of our analysis of \( v_n[1 - 1 - 1] \) as considering a fixed vector embedding \( v_1, \ldots, v_{n-1} \) of \( G_{n-1} \), and then analyzing the probability that \( n \) connects to some \( S \subseteq [n - 1] \). When \( p = \frac{\log n}{\log \log n} \), this does not yield tight results; moreover, one can show that this is not due to loose tail bounds on \( \rho(L) \), as our concentration results have matching anti-concentration results.

Hence, in order to prove Theorem 1.1, we must additionally consider the concentration of \( \rho(L) \) on average over vector embeddings of \( G_{n-1} \) as well. To do this, we will use a “cavity-method” style argument: we will view all vectors at distance \( \ell = \frac{\log n}{\log \log n} \) from \( S \) as fixed and arbitrary, and then exactly compute the marginal distributions over \( u_i \) for \( i \) at distance \( \leq \ell \) from \( S \), conditional on forming \( G_{n-1} \).

### 2.3 Neighborhood Containment as a Constraint Satisfaction Problem

We first reduce the need for high-probability estimates for

\[
\Pr_{G \sim \text{Geo}_d(n, p)}[N(k) = S]
\]

to obtaining estimates for \( \Pr_{G \sim \text{Geo}_d(n, p)}[N(k) \geq S] \) instead. This simplification is possible because the measure of anti-cap intersections concentrates dramatically better than the measure of cap intersections. With this step, we eliminate the need to study anti-correlations between \( u_i \) that do not have an edge between them.

Given \( S \) and \( G_{n-1} \) (and its corresponding vectors), we fix all vectors except those corresponding to the depth \( \log \frac{n}{\log \log n} \) neighborhood of \( S \) in \( G_{n-1} \), which is with high probability a union of trees. To formally analyze the distribution of the unfixed vectors upon resampling them, we set up a 2-CSP (constraint satisfaction problem) instance over a continuous alphabet that encodes the edges of \( G_{n-1} \) within the trees around \( S \): each node has a vector-valued variable in \( S^{d-1} \), and the constraints are that nodes joined by an edge must have vectors with inner product at least \( \tau \).

**Belief propagation.** Since our 2-CSPs are over trees, the belief propagation (BP) algorithm exactly computes the marginal distribution of each variable vector (see Section 3.6 of the full version for the definition of BP). Using our results on the concentration of \( \Pr_{\mathcal{L}}[(x, z) \geq \tau] \over z \text{ uniform on } S^{d-1} \), we can quantify the TV distance between the marginal distributions of our resampled vectors and the uniform distribution over \( S^{d-1} \). At a high level, the farther some \( v_j \) is from a fixed vector in our 2-CSP, the closer its distribution is to uniform. The key insight is that the message from \( i \) to its neighbor \( j \) in our belief propagation algorithm correspond to a convolution of the marginal distribution of \( v_i \) with a spherical cap. We can then use our concentration of measure for spherical caps from Section 4 to show that convolutions of spherical caps mix to uniform rapidly, causing the correlations between far away vertices to decay. This can be seen as a form of the “decay of correlations”
phenomenon. This analysis gives us the finer-grained control over $\Pr[N_2(n) = S]$ needed to conclude Theorem 1.1.

ACKNOWLEDGMENTS

Much of this work was conducted when we were participating in the Simons Institute Fall 2020 program on “Probability, Geometry, and Computation in High Dimensions”. We thank the Simons Institute for their support, and the program for inspiration. S.M. would like to thank Abhishek Shetty for inspiring conversations on geometric concentration.

S.L. is supported in part by the Berkeley Haas Blockchain Initiative and a donation from the Ethereum Foundation. S.M. is supported by a Google PhD Fellowship. E.Y. is supported by the NSF GRFP under Grant No. DGE 1752814.

REFERENCES

[1] Noga Alon and Joel H. Spencer. 2004. The probabilistic method. John Wiley & Sons. https://doi.org/10.1002/0471722154

[2] Aditya Bhaskara, Moses Charikar, Eden Chlamtac, Uriel Feige, and Aravindan Vijayaraghavan. 2010. Detecting high log-densities: an $O(n^{1/4})$ approximation for densest $k$-subgraph. In Proceedings of the forty-second ACM symposium on Theory of computing. 201–210. https://doi.org/10.1145/1806689.1806719

[3] Matthew Brennan, Guy Bresler, and Brice Huang. 2021. De Finetti-Style Results to thank Abhishek Shetty for inspiring conversations on geometric concentration.

[4] Matthew Brennan, Guy Bresler, and Dheeraj Nagaraj. 2020. Phase transitions for detecting latent geometry in random graphs. Probability Theory and Related Fields 178, 3 (2020), 1215–1289. https://doi.org/10.1007/s00440-020-00998-3

[5] Sébastien Bubeck, Jian Ding, Ronen Eldan, and Miklós Z Rácz. 2016. Testing for vector chromatic numbers and huge chromatic numbers. SIAM J. Comput. 33, 6 (2004), 1338–1368. https://doi.org/10.1139/SJCS.2002.1181951

[6] Uriel Feige and Gideon Schechtman. 2002. On the optimality of the random hyperplane rounding technique for MAX CUT. Random Structures & Algorithms 20, 3 (2002), 403–440. https://doi.org/10.1002/rsa.10036

[7] Amir Dembo, Andrea Montanari, Allan Sly, and Nike Sun. 2014. The replica symmetry of strong approximation resistance. In Proceedings of the forty-sixth annual ACM symposium on Theory of computing. 634–643. https://doi.org/10.1145/2591796.2591817

[8] Siqi Liu and Miklós Z. Rácz. 2021. Phase transition in noisy high-dimensional random geometric graphs. arXiv preprint arXiv:2103.14011 (2021).

[9] Marc Mézard and Giorgio Parisi. 2003. The cavity method at zero temperature. J. Combin. Theory B 88, 2 (2003), 124–200. https://doi.org/10.1016/S0095-9972(02)00006-X

[10] Mark Jerrum. 1992. Large cliques elude the Metropolis process. Random Structures & Algorithms 3, 4 (1992), 347–359. https://doi.org/10.1002/rsa.3240030402

[11] Dmitry Panchenko. 2009. Cavity method in the spherical SK model. In Random Fields and Geometry. Springer, 273–324. https://doi.org/10.1007/978-3-030-36020-7_13

[12] Paul W. Holland, Kathryn Blackmond Laskey, and Samuel Leinhardt. 1983. Stochastic blockmodels: First steps. Social Networks 5, 2 (1983), 199–137. https://doi.org/10.1016/0378-8733(83)90021-7

[13] Uriel Feige and Joe Kilian. 2001. Heuristics for semirandom graph problems. J. Comput. System Sci. 63, 4 (2001), 639–671. https://doi.org/10.1006/jcss.2001.1773

[14] David Gamarnik, Dmitriy Katz, and Sudhansu Misra. 2015. Strong spatial mixing of list coloring of graphs. Random Structures & Algorithms 46, 4 (2015), 599–613. https://doi.org/10.1002/rsa.20518

[15] Dmitry Panchenko. 2001. Linear lower bound on degrees of Positivstellensatz calculus proofs for the parity. Theoretical Computer Science 259, 1-2 (2001), 613–622. https://doi.org/10.1016/S0304-3975(00)00157-2

[16]チノ三式本

[17] David Gamarnik, Dmitriy Katz, and Sidhant Misra. 2015. Strong spatial mixing for densest $k$-subgraph. In Proceedings of the thirty-eighth annual ACM symposium on Theory of computing. 721–730. https://doi.org/10.1145/2746539.2746640

[18] Daniel Cullina and Negar Kiyavash. 2016. Improved achievability and converse bounds for erdos-rényi graph matching. ACM SIGMETRICS Performance Evaluation Review 44, 1 (2016), 63–72. https://doi.org/10.1145/2896377.2901460

[19] Amit Coja-Oghlan, Florent Krzakala, Will Perkins, and Lenka Zdeborová. 2018. Entropic CLT and phase transition in noisy high-dimensional random geometric graphs. arXiv preprint arXiv:1802.08728 (2018).