Electronic Journal of Statistics
Vol. 17 (2023) 3512–3574
ISSN: 1935-7524
https://doi.org/10.1214/23-EJS2162

Phase transition in noisy high-dimensional random geometric graphs

Suqi Liu

Department of Operations Research and Financial Engineering, Princeton University, Sherrerd Hall, Charlton Street, Princeton, NJ 08544
e-mail: suqil@princeton.edu

Miklós Z. Rácz

Department of Operations Research and Financial Engineering, Princeton University, Sherrerd Hall, Charlton Street, Princeton, NJ 08544
e-mail: mracz@princeton.edu

Abstract: We study the problem of detecting latent geometric structure in random graphs. To this end, we consider the soft high-dimensional random geometric graph $\mathcal{G}(n, p, d, q)$, where each of the n vertices corresponds to an independent random point distributed uniformly on the sphere $\mathbb{S}^{d-1}$, and the probability that two vertices are connected by an edge is a decreasing function of the Euclidean distance between the points. The probability of connection is parametrized by $q \in [0, 1]$, with smaller q corresponding to weaker dependence on the geometry; this can also be interpreted as the level of noise in the geometric graph. In particular, the model smoothly interpolates between the hard spherical random geometric graph $\mathcal{G}(n, p, d)$ (corresponding to $q = 1$) and the Erdős–Rényi model $\mathcal{G}(n, p)$ (corresponding to $q = 0$). We focus on the dense regime (i.e., p is a constant).

We show that if $nq \to 0$ or $d \gg n^3 q^2$, then geometry is lost: $\mathcal{G}(n, p, d, q)$ is asymptotically indistinguishable from $\mathcal{G}(n, p)$. On the other hand, if $d \ll n^3 q^6$, then the signed triangle statistic provides an asymptotically powerful test for detecting geometry. These results generalize those of Bubeck, Ding, Eldan, and Rácz (2016) for $\mathcal{G}(n, p, d)$, and give quantitative bounds on how the noise level affects the dimension threshold for losing geometry. We also prove analogous results under a related but different distributional assumption which corresponds to the random dot product graph.

MSC2020 subject classifications: Primary 60K35; secondary 05C80, 60B20, 60C05.
Keywords and phrases: Random graph, random geometric graph, hypothesis testing, high-dimensional geometric structure, signed triangle.

Received September 2022.

1. Introduction

Random graphs emerge as canonical models for many real-life applications, including social networks, wireless communications, and in the biological sciences.

arXiv: 2103.15249

Among them, the simplest yet structurally rich model is the Erdős–Rényi random graph, which has been studied extensively (e.g., [19, 7]). However, in many scenarios the independence of edges is an oversimplified assumption and is often insufficient to capture the subtle relations in complex networks.

A natural extension is to assume an underlying geometric structure. The graph is then generated according to some dependence on this structure. Due to their wide applicability, random graphs of this kind have various incarnations in different fields: random geometric graphs, latent space models, spatial networks, random connection models, to name a few. We refer the reader to [33] for a comprehensive theoretical treatment of the subject.

In real-world networks, the geometric space often remains unobserved. Most of the time, only the graph structure can be obtained rather than the latent factors that generated the graph. For example, in a social network, it is usually easy to access the interactions between people but not how these connections are established. This brings up the natural question of understanding the extent to which a latent space model is an accurate description. As a first step, it is crucial to understand when the presence of geometry is even detectable assuming that the network follows a specific generative model. Mathematically this was first studied by Devroye, György, Lugosi, and Udina [16] for a particular random geometric graph equipped with the spherical geometry. They showed that this random geometric graph becomes indistinguishable from an Erdős–Rényi graph when the dimension of the sphere goes to infinity. In other words, geometry is lost in high dimensions. Subsequently, Bubeck, Ding, Eldan, and Rácz [12] pinpointed the phase transition for testing high-dimensional geometry in dense random graphs. Our paper builds upon and generalizes this result.

A caveat of the aforementioned works is that the model is restricted to a "hard geometry" setting, where the existence of an edge is a deterministic function of the distance between the latent points corresponding to the two vertices. This assumption overlooks the fact that in reality connections can often have stochastic dependence on the latent variables. Consequently, the phase transition in the hard geometry setting happens at dimensions as high as the cube of the number of vertices, seemingly much larger than what many high-dimensional statistics theories would consider [43].

Our focus in this paper is to understand the above question in the setting of *soft* random geometric graphs, in which the softness can be viewed as noise in the geometric graph. We are particularly interested in the interplay between dimensionality and noise in affecting the phenomena of losing geometry in random graphs. To this end, we study a particular type of soft random geometric graph where there is a parameter $q \in [0, 1]$ that naturally encodes the level of noise, correspondingly the strength of geometry. This model is an interpolation between the hard spherical random geometric graph (corresponding to $q = 1$) and the Erdős–Rényi model (corresponding to $q = 0$). Our main results provide bounds, as a function of both dimension and geometry strength, of where the phase transition lies. In particular, these results quantitatively demonstrate the qualitative phenomenon that the dimension threshold for losing geometry is smaller for soft random geometric graphs and decreases as a function of noise level. We next specify the precise setting of our work, before describing our main results.

1.1. Random geometric graphs

We first describe the spherical random geometric graph model $\mathcal{G}(n,p,d)$ [16, 12]. For a set of n vertices $V = [n] \coloneqq \{1, 2, \ldots, n\}$, associate each vertex i with a point represented by a d-dimensional random vector $\boldsymbol{x}_i$. We assume that $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$ are independently uniformly distributed on the sphere $\mathbb{S}^{d-1} \coloneqq \{\boldsymbol{x} \in \mathbb{R}^d : \|\boldsymbol{x}\| = 1\}$, where $\|\cdot\|$ stands for the Euclidean norm. For a fixed value of $p \in [0,1]$ that parametrizes the edge density, the graph is defined as follows: There is an undirected edge between distinct vertices i and j if and only if

$$\langle \boldsymbol{x}_i, \boldsymbol{x}_j \rangle \geq t_{p,d}, \tag{1}$$

where $\langle \cdot, \cdot \rangle$ denotes the inner product of two vectors. Equivalently, i and j are connected by an edge if and only if their Euclidean distance satisfies $\|\boldsymbol{x}_i - \boldsymbol{x}_j\| \leq \sqrt{2(1 - t_{p,d})}$. The threshold $t_{p,d}$, which may depend on p and d, is determined by the equation

$$\mathbb{P}(\langle \boldsymbol{x}_i, \boldsymbol{x}_j \rangle \geq t_{p,d}) = p,$$

so that the probability of an edge existing between any pair of distinct vertices is p. Given the latent vectors $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$, the edges in the graph are deterministic. The only source of randomness in $\mathcal{G}(n,p,d)$ comes from the random points. For this reason, the random graph defined above is referred to as a *hard* random geometric graph.

More generally, the model above may be extended by adding additional randomness to the edge generating process, given the latent positions. That is, we connect i and j with probability $\phi(\langle \boldsymbol{x}_i, \boldsymbol{x}_j \rangle)$ for a *connection function* $\phi : \mathbb{R} \to [0,1]$. Formally, let $i \sim j$ denote the event that there exists an undirected edge between i and j. Then,

$$\mathbb{P}(i \sim j \mid \boldsymbol{x}_1, \ldots, \boldsymbol{x}_n) = \phi(\langle \boldsymbol{x}_i, \boldsymbol{x}_j \rangle). \tag{2}$$

This is equivalent to connecting i and j with probability $\phi(1 - \|\boldsymbol{x}_i - \boldsymbol{x}_j\|^2/2)$. When the connection function is an indicator $\phi(x) = \mathbb{1}\{x \geq t_{p,d}\}$, we obtain the hard random geometric graph defined previously. For general connection functions, which are typically nondecreasing,[1] such random graphs are referred to as *soft* random geometric graphs.

Denote by $\boldsymbol{K} = [k_{i,j}]$ the *connection matrix* defined by $k_{i,j} \coloneqq \phi(\langle \boldsymbol{x}_i, \boldsymbol{x}_j \rangle)$. For a simple graph $G = (V, E)$, where V is the set of vertices and E is the set of edges, let $\boldsymbol{A} = [a_{i,j}]$ be its adjacency matrix, where $a_{i,j} = 1$ if $i \sim j$ and $a_{i,j} = 0$ otherwise for all $i \neq j$. Denote by $\boldsymbol{X} \in \mathbb{R}^{n \times d}$ the matrix whose rows are the random vectors $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$. For a soft random geometric graph with connection

[1] In the literature, connection functions often take as their argument the distance $\|\boldsymbol{x}_i - \boldsymbol{x}_j\|$ and hence are a nonincreasing function. Here it is more convenient to take the inner product $\langle \boldsymbol{x}_i, \boldsymbol{x}_j \rangle$ as the argument of ϕ and hence this is a nondecreasing function.

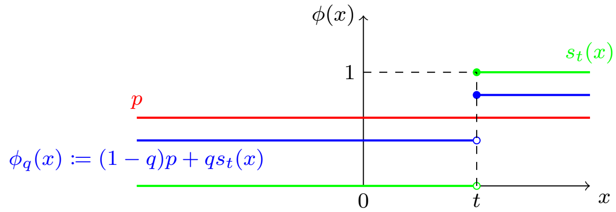

FIG 1. *A comparison of connection functions.*

matrix $\boldsymbol{K}$, conditioned on $\boldsymbol{X}$, $\mathbb{1}\{i \sim j\}$ is an independent Bernoulli random variable with parameter $k_{i,j}$. The distribution of the soft random geometric graph is then specified by

$$\mathbb{P}(G) = \mathbb{E}_{\boldsymbol{X}}\left[\prod_{i<j} {k_{i,j}}^{a_{i,j}} (1 - k_{i,j})^{1-a_{i,j}}\right]. \tag{3}$$

We focus on a particular family of soft random geometric graphs parametrized by the level of dependence on the underlying points. The connection function in these models is a linear interpolation between a constant $p \in [0, 1]$ and a step function $s_t : \mathbb{R} \to \{0, 1\}$ defined as $s_t(x) := \mathbb{1}\{x \geq t\}$. Concretely, we consider the following connection function with a parameter $q \in [0, 1]$:

$$\phi_q(x) := (1 - q)p + qs_t(x), \tag{4}$$

where q controls the strength of geometry in the graph. The threshold t is similarly determined by setting the probability of an edge to be p:

$$\mathbb{P}(i \sim j) = \mathbb{E}[\phi_q(\langle \boldsymbol{x}_i, \boldsymbol{x}_j \rangle)] = p,$$

which gives the same threshold $t = t_{p,d}$ as in the definition of $\mathcal{G}(n, p, d)$. We denote by $\mathcal{G}(n, p, d, q)$ the soft random geometric graph equipped with the connection function ϕ_q. When $q = 1$, the graph becomes the hard random geometric graph $\mathcal{G}(n, p, d)$. When $q = 0$, each edge in the graph is generated independently with probability p, corresponding to the well-known Erdős–Rényi graph $\mathcal{G}(n, p)$ that does not possess latent geometric structure. As an illustration, $\phi_q(x)$ is plotted against the connection functions of $\mathcal{G}(n, p)$ and $\mathcal{G}(n, p, d)$ in Figure 1.

$\mathcal{G}(n, p, d, q)$ can also be viewed as a hard random geometric graph with independent edge resampling. Starting with a sample graph from $\mathcal{G}(n, p, d)$, for each pair of vertices in the graph, we flip a biased coin with the probability of heads equal to q. If the coin shows heads, we keep the edge/non-edge between them; otherwise, we resample the connection, creating an edge independently with probability p. The resampled graph then follows $\mathcal{G}(n, p, d, q)$.

1.2. Main results

As the scope of this paper, we are interested in whether it is possible to detect the underlying geometric structure, which we formulate as the following hypothesis testing problem. The null hypothesis is that the observed graph G is a sample from the Erdős–Rényi model with edge density p:

$$H_0 : G \sim \mathcal{G}(n, p).$$

The alternative hypothesis is that the graph is a (matching edge density) soft random geometric graph with dimension d and geometry strength q:

$$H_1 : G \sim \mathcal{G}(n, p, d, q).$$

The hypothesis testing problem can be understood through guarantees for the total variation distance between the two distributions. Our findings are summarized in the following theorem.

Theorem 1.1 (Detecting geometry)**.** *Let $p \in (0, 1)$ be fixed.*

(a) (Impossibility) If $nq \to 0$ or $n^3q^2/d \to 0$, then

$$\mathrm{TV}(\mathcal{G}(n, p), \mathcal{G}(n, p, d, q)) \to 0.$$

(b) (Possibility) If $n^3q^6/d \to \infty$, then

$$\mathrm{TV}(\mathcal{G}(n, p), \mathcal{G}(n, p, d, q)) \to 1.$$

Theorem 1.1(a) specifies a lower bound for detection: If $nq \to 0$ or $d \gg n^3q^2$ then no test can detect the latent geometric structure; $\mathcal{G}(n, p, d, q)$ is asymptotically indistinguishable from $\mathcal{G}(n, p)$. On the other hand, Theorem 1.1(b) provides an upper bound: If $d \ll n^3q^6$ then there exists an asymptotically powerful test for detecting the latent geometric structure. Specifically, we will show that the signed triangle statistic of Bubeck et al. [12] (which in particular is computationally efficient) works in this regime to distinguish the two models.

Recall that $\mathcal{G}(n, p, d, q)$ becomes $\mathcal{G}(n, p, d)$ in the special case when $q = 1$. In this case Theorem 1.1 recovers the results of Bubeck et al. [12], showing that $d \asymp n^3$ is the dimension threshold for losing geometry in $\mathcal{G}(n, p, d)$.

In general, both the impossibility and possibility results in Theorem 1.1 depict a polynomial dependency on q. However, the polynomials have different powers of q, which implies that there exists a gap between the lower and upper bounds. We believe that the phase transition for losing geometry in $\mathcal{G}(n, p, d, q)$ happens at a certain power of q. We conjecture that the bound in Theorem 1.1(b) specified by the signed triangle statistic is tight, that is, when $d \gg n^3q^6$, it is impossible to distinguish between the two models; in other words, the signed triangle statistic is (nearly) optimal. In particular, the dimension threshold $d \asymp (nq^2p\log(1/p))^3$ determined by the signed triangle statistic is consistent with the conjectured threshold for the original random geometric graph in the sparse regime (see [12, Conjecture 1] and [26, after Theorem 1.2]). Specifically, the authors of [26] showed that signed triangles can in fact distinguish $\mathcal{G}(n, p, d)$ from $\mathcal{G}(n, p)$ whenever $d \ll (np\log(1/p))^3$ for all $p = \Omega(1/n)$ and conjectured this is optimal for all $p \leq 1/2$.

Theorem 1.1 can be displayed graphically by a phase diagram of when the latent geometric structure can be detected and when it cannot in $\mathcal{G}(n, p, d, q)$ regarding dimension d and geometry strength q. We further introduce a more convenient parametrization that allows us to visualize the phase diagram.

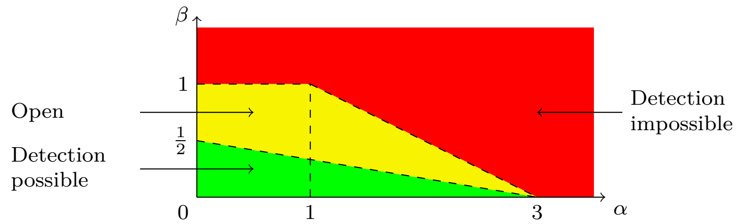

FIG 2. *Phase diagram for detecting geometry in the soft random geometric graph $\mathcal{G}(n,p,d,q)$. Here $d = n^{\alpha}$ and $q = n^{-\beta}$ for some $\alpha, \beta > 0$.*

Corollary 1.2 (Phase diagram)**.** *Suppose that $d \asymp n^{\alpha}$ and $q \asymp n^{-\beta}$ for some $\alpha, \beta > 0$.*

(a) If $\beta > 1$ or $\alpha + 2\beta > 3$, then as $n \to \infty$,

$$\mathrm{TV}(\mathcal{G}(n,p), \mathcal{G}(n,p,d,q)) \to 0.$$

(b) If $\alpha + 6\beta < 3$, then as $n \to \infty$,

$$\mathrm{TV}(\mathcal{G}(n,p), \mathcal{G}(n,p,d,q)) \to 1.$$

The resulting phase diagram is plotted in the two-dimensional space of α and β in Figure 2.

1.3. Related work

The study of high-dimensional random geometric graphs originates from the work of Devroye et al. [16], who showed via the multivariate central limit theorem that geometry is lost in high dimensions. Subsequently, Bubeck et al. [12] determined that the phase transition of losing geometry happens asymptotically at dimension $d \asymp n^3$ in the dense setting (for fixed $0 < p < 1$). This work also points out the connection to classical random matrix ensembles, showing that the Wishart to GOE transition also happens at $d \asymp n^3$ (see also [23, 11, 37, 36, 14, 28]). Eldan and Mikulincer [17] further extended the results to an anisotropic setting of the underlying distribution of points; recent follow-up work of Brennan, Bresler, and Huang [10] precisely determined the detection threshold in this setting. In the sparse setting, when p vanishes as a function of n, it is conjectured in [12] that geometry should be lost at much lower dimensions. Progress towards this conjecture, which in particular breaks the n^3 barrier, was made by Brennan, Bresler, and Nagaraj [8]. This is an active line of research; after we finished this work, we learned about a new preprint by Brennan, Bresler, and Huang [9] on understanding the Wishart to GOE transition when only a subset of entries is revealed. A breakthrough in the sparse case was made by Liu, Mohanty, Schramm, and Yang [26], who showed that when $p = c/n$ for a constant c, if $d \geq \mathrm{polylog}(n)$, the total variation distance between the two distributions is close to 0. This resolves the conjecture of [12] up to logarithmic factors.

Soft random geometric graphs arise as natural models in many areas, including wireless communications [20], social networks [22], and biological networks [41]. Penrose [34] studied the connectivity of soft random geometric graphs from a modern probability-theoretic perspective, determining the asymptotic probability of connectivity in fixed dimensions and for a broad class of connection functions. Dettmann and Georgiou [15] discussed the same questions from a statistical physics viewpoint in two and three dimensions, and provided a comprehensive list of connection functions widely used in practice. Connectivity in one-dimensional soft random geometric graphs was considered in [45], where the authors showed that the reason for connectivity is vastly different from the hard case. Parthasarathy et al. [32] studied a model of perturbed networks, which is similar to the setting under consideration in our work. A phase transition in soft random geometric graphs with a critical value of chemical potential was demonstrated in [31], where a related model was also considered. Our paper can be viewed as a first step towards understanding the questions described in the previous paragraph for soft random geometric graphs. In a follow-up paper to this one [25], we showed a similar phase diagram for detecting high-dimensional geometry for *smooth* connection functions. A recent preprint by Bangachev and Bresler [5] extends the results to geometric and algebraic settings and makes several advances in this direction.

Following up detecting geometry, a natural next question is how to recover it; indeed, a line of research focuses on recovering the underlying latent positions of soft random geometric graphs. In [40], it is shown that latent positions for random dot product graphs can be estimated consistently using the eigendecomposition of the adjacency matrix, when the dimension d is fixed. Several subsequent works [3, 2, 18] applied similar approaches to kernels and general connection functions on spheres satisfying certain eigengap conditions.

1.4. Open problems

The most immediate problem that our work leaves open is to understand the intermediate region not covered by Theorem 1.1. Specifically, the main question is to determine the exact boundary between the two phases where the limiting total variation distance transitions from 1 to 0. The existence of an intermediate phase where detection is information-theoretically possible while no efficient algorithm exists is also worth studying.

More broadly, a natural direction of future research is to consider these questions for other connection functions or underlying latent spaces, in order to understand how the dimension threshold for losing geometry depends on them.

1.5. Outline of the paper

The rest of the paper is organized as follows. In Section 2, we introduce some notations used throughout the paper and several standard definitions. The impossibility of detection is presented afterwards in Section 3, where the results of Theorem 1.1(a) in the two regimes are proved. Section 4 consists of the proof of Theorem 1.1(b) using the signed triangle statistic. Finally, in Section 5 we show that similar results also hold under a different distributional assumption, when the underlying latent vectors are i.i.d. standard normal, and we consider the corresponding random dot product graph.

2. Notations and preliminaries

We use boldface capitals to denote matrices, and their corresponding lower cases with subscript indices separated by a comma to denote the entries. For example, a matrix $\boldsymbol{A} = [a_{i,j}]$ has entry $a_{i,j}$ in its ith row and jth column. We use $\circ$ to denote the Hadamard product, which is the entrywise product of two matrices. Vectors are represented by boldface lowercase letters, and the corresponding normal font with a subscript index denotes each entry. All vectors are treated as columns.

Definition 2.1 (f-divergence)**.** Let $\mathcal{P}$ and $\mathcal{Q}$ be two probability measures on the same measurable space $(\Omega, \mathcal{F})$. Suppose that $\mathcal{P}$ is absolutely continuous with respect to $\mathcal{Q}$. For a convex function f such that $f(1) = 0$, the f-divergence of $\mathcal{P}$ and $\mathcal{Q}$ is defined as

$$\mathrm{D}_f(\mathcal{P} \parallel \mathcal{Q}) \coloneqq \mathbb{E}_{\mathcal{Q}}\left[f\left(\frac{d\mathcal{P}}{d\mathcal{Q}}\right)\right] = \int_\Omega f\left(\frac{d\mathcal{P}}{d\mathcal{Q}}\right) d\mathcal{Q},$$

where $\frac{d\mathcal{P}}{d\mathcal{Q}}$ is the Radon–Nikodym derivative of $\mathcal{P}$ with respect to $\mathcal{Q}$.

In Definition 2.1, by choosing $f(t) = t \log t$, we have the *Kullback–Leibler (KL) divergence*, which we simply refer to as the *divergence*. Throughout the paper, log stands for the natural logarithm.

Definition 2.2 (Kullback–Leibler (KL) divergence)**.** Let $\mathcal{P}$ and $\mathcal{Q}$ be two probability measures on the same measurable space $(\Omega, \mathcal{F})$. Suppose that $\mathcal{P}$ is absolutely continuous with respect to $\mathcal{Q}$. The (KL) divergence of $\mathcal{P}$ and $\mathcal{Q}$ is defined as

$$\mathrm{KL}(\mathcal{P} \parallel \mathcal{Q}) \coloneqq \mathbb{E}_{\mathcal{Q}}\left[\frac{d\mathcal{P}}{d\mathcal{Q}} \log \frac{d\mathcal{P}}{d\mathcal{Q}}\right] = \int_\Omega \log \frac{d\mathcal{P}}{d\mathcal{Q}}\, d\mathcal{P}.$$

Definition 2.3 (Total variation distance)**.** Let $\mathcal{P}$ and $\mathcal{Q}$ be two probability measures on a measurable space $(\Omega, \mathcal{F})$. The total variation distance between $\mathcal{P}$ and $\mathcal{Q}$ is defined as

$$\mathrm{TV}(\mathcal{P}, \mathcal{Q}) \coloneqq \sup_{A \in \mathcal{F}} |\mathcal{P}(A) - \mathcal{Q}(A)|.$$

The total variation distance can also be viewed as an f-divergence with $f(x) = \frac{1}{2}|x-1|$. The total variation distance is simply referred to as the *distance* where no confusion is caused. From the definition, it is clear that the distance between $\mathcal{P}$ and $\mathcal{Q}$ is symmetric. That is,

$$\mathrm{TV}(\mathcal{P}, \mathcal{Q}) = \mathrm{TV}(\mathcal{Q}, \mathcal{P}).$$

We use several inequalities concerning f-divergences in the proofs; we state these later before they are applied. The divergence and the distance are connected through Pinsker's inequality.

Proposition 2.4 (Pinsker's inequality)**.** *For probability measures $\mathcal{P}$ and $\mathcal{Q}$, we have that*

$$\mathrm{TV}(\mathcal{P},\mathcal{Q}) \leq \sqrt{\frac{1}{2}\,\mathrm{KL}(\mathcal{P} \parallel \mathcal{Q})}.$$

3. Impossibility of detecting geometry

In this section we prove the impossibility results stated in Theorem 1.1(a). We start our discussion with some weaker bounds obtained through a simple mixture argument, and then proceed to the proof of the main impossibility results.

We may view the soft random geometric graph $\mathcal{G}(n,p,d,q)$ as an edge-wise mixture between $\mathcal{G}(n,p)$ and $\mathcal{G}(n,p,d)$, in the following way. First, we draw two sample graphs $G_1 \sim \mathcal{G}(n,p)$ and $G_2 \sim \mathcal{G}(n,p,d)$. We next construct a graph G using G_1, G_2, and additional coin flips. Specifically, for every pair of distinct vertices i and j, we flip an independent biased coin which comes up heads with probability q. If the coin flip is heads, connect i and j with an edge in G if and only if they are connected with an edge in G_2; otherwise, connect i and j with an edge in G if and only if they are connected with an edge in G_1. This construction guarantees that $G \sim \mathcal{G}(n,p,d,q)$.

We can obtain two simple bounds directly from this construction. Bubeck et al. [12, Theorem 1(c)] showed that $\mathcal{G}(n,p)$ and $\mathcal{G}(n,p,d)$ become indistinguishable when $d \gg n^3$. Thus, as an edge-wise mixture of these two models, $\mathcal{G}(n,p,d,q)$ also cannot be distinguished from $\mathcal{G}(n,p)$ in this regime.

Meanwhile, when $q \ll 1/n^2$, for any pair (i,j) the probability that the connection between i and j is sampled from G_2 is $o(1/n^2)$. This implies that the expected number of edges that are sampled from G_2 is $o(1)$. Therefore, by Markov's inequality, the probability that there exists an edge which is sampled from G_2 is $o(1)$. Hence, with probability $1 - o(1)$ we have that $G = G_1$ in the construction above. Therefore, $\mathcal{G}(n,p)$ and $\mathcal{G}(n,p,d,q)$ are indistinguishable when $q \ll 1/n^2$. These two arguments are summarized in the following claim.

Claim 3.1. *If $n^3/d \to 0$ or $n^2 q \to 0$, then*

$$\sup_{p \in [0,1]} \mathrm{TV}(\mathcal{G}(n,p), \mathcal{G}(n,p,d,q)) \to 0.$$

Our main result in Theorem 1.1(a) improves on Claim 3.1 by relaxing both of the conditions and thus proving the impossibility result for a larger parameter regime. First, notice that the condition $n^3/d \to 0$ does not take q into consideration at all; we improve this to the condition $n^3 q^2/d \to 0$. We also improve the condition $n^2 q \to 0$ to $nq \to 0$.

As before, by choosing a convenient parametrization, Claim 3.1 translates into the following corollary picturing a region of a phase diagram.

Corollary 3.2. *Suppose that $d \asymp n^\alpha$ and $q \asymp n^{-\beta}$ for some $\alpha, \beta > 0$. If $\alpha > 3$ or $\beta > 2$, then as $n \to \infty$,*

$$\sup_{p \in [0,1]} \mathrm{TV}(\mathcal{G}(n,p), \mathcal{G}(n,p,d,q)) \to 0.$$

3.1. Impossibility of detection under large noise

In this subsection we show that $\mathcal{G}(n,p)$ and $\mathcal{G}(n,p,d,q)$ are asymptotically indistinguishable when $nq \to 0$, thus proving Theorem 1.1(a) under this regime. This shows that when the noise is large enough (i.e., q is small enough), detecting geometry becomes impossible, regardless of the dimensionality.

For a graph G with adjacency matrix $\boldsymbol{A} = [a_{i,j}]$, the probability $\mathbb{P}_{\mathcal{G}(n,p,d,q)}(G)$ is given in (3) with $k_{i,j} = \phi_q(\langle \boldsymbol{x}_i, \boldsymbol{x}_j \rangle)$. We can write the probability $\mathbb{P}_{\mathcal{G}(n,p)}(G)$ similarly:

$$\mathbb{P}_{\mathcal{G}(n,p)}(G) = \prod_{i<j} p^{a_{i,j}} (1-p)^{1-a_{i,j}}. \tag{5}$$

The divergence of $\mathcal{G}(n,p)$ and $\mathcal{G}(n,p,d,q)$ can then be written as

$$\begin{aligned}\mathrm{KL}(\mathcal{G}(n,p) \,\|\, \mathcal{G}(n,p,d,q)) &= \mathbb{E}_{\mathcal{G}(n,p)}\left[-\log \frac{\mathbb{P}_{\mathcal{G}(n,p,d,q)}(G)}{\mathbb{P}_{\mathcal{G}(n,p)}(G)}\right] \\ &= \mathbb{E}_{\boldsymbol{A}}\left[-\log \mathbb{E}_{\boldsymbol{X}}\left[\prod_{i<j}\left(\frac{k_{i,j}}{p}\right)^{a_{i,j}}\left(\frac{1-k_{i,j}}{1-p}\right)^{1-a_{i,j}}\right]\right],\end{aligned}$$

where the $a_{i,j}$'s are independent Bernoulli random variables with parameter p since the expectation is taken under $G \sim \mathcal{G}(n,p)$. Since $-\log$ is convex, by Jensen's inequality we have that

$$\begin{aligned}&\mathrm{KL}(\mathcal{G}(n,p) \,\|\, \mathcal{G}(n,p,d,q)) \\ &\quad \le \mathbb{E}_{\boldsymbol{A},\boldsymbol{X}}\left[-\log \prod_{i<j}\left(\frac{k_{i,j}}{p}\right)^{a_{i,j}}\left(\frac{1-k_{i,j}}{1-p}\right)^{1-a_{i,j}}\right] \\ &\quad = \mathbb{E}_{\boldsymbol{A},\boldsymbol{X}}\left[-\sum_{i<j}\left(a_{i,j}\log\frac{k_{i,j}}{p} + (1-a_{i,j})\log\frac{1-k_{i,j}}{1-p}\right)\right] \\ &\quad = -\sum_{i<j}\left(\mathbb{E}_{\boldsymbol{A}}[a_{i,j}]\,\mathbb{E}_{\boldsymbol{X}}\left[\log\frac{k_{i,j}}{p}\right] + \mathbb{E}_{\boldsymbol{A}}[1-a_{i,j}]\,\mathbb{E}_{\boldsymbol{X}}\left[\log\frac{1-k_{i,j}}{1-p}\right]\right),\end{aligned} \tag{6}$$

where the last line is due to linearity of expectation and independence.

Since $a_{i,j} \sim \mathrm{Bern}(p)$, we have $\mathbb{E}[a_{i,j}] = p$ and $\mathbb{E}[1-a_{i,j}] = 1-p$. By the definition of the connection function in (2) and (4), we have that

$$k_{i,j} = \phi_q(\langle \boldsymbol{x}_i, \boldsymbol{x}_j \rangle) = \begin{cases} (1-q)p + q & \text{if } \langle \boldsymbol{x}_i, \boldsymbol{x}_j \rangle \ge t_{p,d}, \\ (1-q)p & \text{otherwise.}\end{cases}$$

Recall that $t_{p,d}$ is chosen such that $\mathbb{P}(\langle \boldsymbol{x}_i, \boldsymbol{x}_j \rangle \geq t_{p,d}) = p$. Hence, the marginal distribution of $k_{i,j}$ satisfies $k_{i,j} = (1-q)p + q$ with probability p and $k_{i,j} = (1-q)p$ with probability $1-p$. Therefore, we have that

$$\mathbb{E}_{\boldsymbol{X}}\left[\log \frac{k_{i,j}}{p}\right] = p \log\left(1 + \frac{1-p}{p} q\right) + (1-p)\log(1-q)$$

and

$$\mathbb{E}_{\boldsymbol{X}}\left[\log \frac{1-k_{i,j}}{1-p}\right] = p \log(1-q) + (1-p)\log\left(1 + \frac{p}{1-p} q\right).$$

By the elementary inequality $\log(1+x) \geq x - x^2$ for $x \geq -1/2$, we obtain that for $0 \leq q \leq 1/2$,

$$\mathbb{E}_{\boldsymbol{X}}\left[\log \frac{k_{i,j}}{p}\right] \geq -\frac{1-p}{p} q^2 \quad \text{and} \quad \mathbb{E}_{\boldsymbol{X}}\left[\log \frac{1-k_{i,j}}{1-p}\right] \geq -\frac{p}{1-p} q^2.$$

Inserting the above estimates into (6), we conclude that for $0 \leq q \leq 1/2$,

$$\mathrm{KL}(\mathcal{G}(n,p) \,\|\, \mathcal{G}(n,p,d,q)) \leq \binom{n}{2} q^2 \leq \frac{1}{2} n^2 q^2. \tag{7}$$

To be consistent with the main discussion, we turn this upper bound on the divergence into an upper bound on the distance. An application of Pinsker's inequality (Proposition 2.4) combined with (7) proves the following theorem.

Theorem 3.3. *For $0 \leq q \leq 1/2$ we have that*

$$\sup_{p \in [0,1]} \mathrm{TV}(\mathcal{G}(n,p), \mathcal{G}(n,p,d,q)) \leq \frac{1}{2} nq.$$

The $nq \to 0$ regime of Theorem 1.1(a) directly follows from Theorem 3.3.

3.2. Impossibility of detecting weak high-dimensional geometry

In this subsection we show that $\mathcal{G}(n,p)$ and $\mathcal{G}(n,p,d,q)$ are asymptotically indistinguishable when $n^3 q^2/d \to 0$, thus proving Theorem 1.1(a) in this regime. This result thus highlights the interplay between noise and dimensionality in determining when it is possible to detect geometry.

In order to capture this interplay between noise and dimensionality, we use several inequalities concerning f-divergences, and we start by recalling these. The distance (divergence) between two random variables is understood as the distance (divergence) between their corresponding probability measures. Since we focus our attention on random graphs without self-loops, the diagonal entries of real symmetric matrices are usually set to zero unless specified. For the distance (divergence) between two real symmetric random matrices, only the lower triangular part is considered. We also place notations for distributions inside the operators $\mathbb{E}, \mathbb{V}\mathrm{ar}$ to denote a sample from the corresponding distribution.

Proposition 3.4 (Conditioning increases divergence). *Let $\mathcal{P}_X$ and $\mathcal{Q}_X$ be two probability measures. Let Y be a random variable on the same space and denote by $\mathcal{P}_{X|Y}$ and $\mathcal{Q}_{X|Y}$ the conditional laws. Then, the f-divergence satisfies*

$$\mathrm{D}_f(\mathcal{P}_X \parallel \mathcal{Q}_X) \le \mathbb{E}_Y \mathrm{D}_f(\mathcal{P}_{X|Y} \parallel \mathcal{Q}_{X|Y}).$$

Proposition 3.4 is usually referred to as "conditioning increases divergence" in standard texts (e.g., Theorem 2.2(5) and Remark 4.2 in [35]). Following the convention widely adopted in the information theory community, we write $\mathrm{KL}(\mathcal{P}_{X|Y} \parallel \mathcal{Q}_{X|Y} \mid \mathcal{P}_Y) \coloneqq \mathbb{E}_Y \mathrm{KL}(\mathcal{P}_{X|Y} \parallel \mathcal{Q}_{X|Y})$ and call it the *conditional divergence*.

Proposition 3.5 (Data processing inequality). *For two probability distributions $\mathcal{P}_X$ and $\mathcal{Q}_X$, consider the joint distributions $\mathcal{P}_{X,Y} = \mathcal{P}_{Y|X}\mathcal{P}_X$ and $\mathcal{Q}_{X,Y} = \mathcal{P}_{Y|X}\mathcal{Q}_X$ with the same conditional law $\mathcal{P}_{Y|X}$. Then, the f-divergence of the marginal distributions $\mathcal{P}_Y \coloneqq \mathbb{E}_{\mathcal{P}_X}[\mathcal{P}_{Y|X}]$ and $\mathcal{Q}_Y \coloneqq \mathbb{E}_{\mathcal{Q}_X}[\mathcal{P}_{Y|X}]$ satisfies*

$$\mathrm{D}_f(\mathcal{P}_Y \parallel \mathcal{Q}_Y) \le \mathrm{D}_f(\mathcal{P}_X \parallel \mathcal{Q}_X).$$

A simple proof of Proposition 3.5 using Jensen's inequality can be found in most texts (see, e.g., [35, Theorem 6.2]). For a measurable function $g\colon E \to F$, by choosing $\mathcal{P}_{Y|X}(y \mid x) = \mathbb{1}\{y = g(x)\}$, we have the following corollary.

Corollary 3.6 (Data processing inequality). *Let $X, Y \in E$ be two random variables and let $g : E \to F$ be a measurable function. Then, the f-divergence of the pushforward measures satisfies*

$$\mathrm{D}_f(g(X) \parallel g(Y)) \le \mathrm{D}_f(X \parallel Y).$$

With these preliminaries in place, we now turn to our question of interest. Let $\boldsymbol{Z} \in \mathbb{R}^{n\times d}$ be a random matrix with independent standard normal entries. Then, $\boldsymbol{W} \coloneqq \boldsymbol{Z}\boldsymbol{Z}^\top \in \mathbb{R}^{n\times n}$ has a Wishart distribution $\mathcal{W}_n(\boldsymbol{I}, d)$. Let $\boldsymbol{Y}$ contain the off-diagonals of $\boldsymbol{W}$ scaled by the square root of the dimension, that is, $y_{i,j} \coloneqq w_{i,j}/\sqrt{d}$. By the central limit theorem, $y_{i,j}$ converges to a standard normal random variable as $d \to \infty$. Let $\boldsymbol{V} \coloneqq \frac{1}{d}\operatorname{diag}(\boldsymbol{W})$ consist of the scaled diagonals; by the law of large numbers, each nonzero entry $v_{i,i}$ converges to 1. For compactness of presentation, denote by $\boldsymbol{v}$ the vectorized diagonal of $\boldsymbol{V}$, that is, $v_i \coloneqq w_{i,i}/d$. Let $\boldsymbol{M} = [m_{i,j}]$ be a zero-diagonal symmetric random matrix with off-diagonal entries $m_{i,j}$ following independent standard normal distributions for $1 \le i < j \le n$. Let $\boldsymbol{M}'$ and $\boldsymbol{M}''$ be two independent copies of $\boldsymbol{M}$.

A standard method to create uniform random vectors on the sphere (which goes back to [27] and before) is as follows: If $\boldsymbol{z}$ is a standard normal vector, then $\hat{\boldsymbol{z}} \coloneqq \boldsymbol{z}/\|\boldsymbol{z}\|$ is uniformly distributed in $\mathbb{S}^{d-1}$. Thus, we can create the random vectors $\{\boldsymbol{x}_i\}_{i=1}^n$ using $\{\boldsymbol{z}_i\}_{i=1}^n$. With the random matrices defined above, the inner product of $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$ for $i \ne j$ can be expressed as

$$\begin{aligned}\langle \boldsymbol{x}_i, \boldsymbol{x}_j\rangle = \langle \hat{\boldsymbol{z}}_i, \hat{\boldsymbol{z}}_j\rangle = \frac{\langle \boldsymbol{z}_i, \boldsymbol{z}_j\rangle}{\|\boldsymbol{z}_i\|\|\boldsymbol{z}_j\|} = \frac{w_{i,j}}{\sqrt{w_{i,i}w_{j,j}}} = \frac{1}{\sqrt{d}}\frac{y_{i,j}}{\sqrt{v_i v_j}} \\ = \frac{1}{\sqrt{d}}(\boldsymbol{V}^{-1/2}\boldsymbol{Y}\boldsymbol{V}^{-1/2})_{i,j}.\end{aligned}$$

Define the step function $s_{t_p}(x) := \mathbb{1}\{x \geq t_p\}$, where t_p is determined by $\mathbb{E}[s_{t_p}(\mathcal{N}(0,1))] = p$. That is, if we denote the cumulative distribution function of the standard normal distribution by Φ, then $t_p = \Phi^{-1}(1-p)$. In what follows, when we apply a univariate function to a matrix, it is applied entrywise, resulting in a matrix of the same shape. With this convention, we define two matrices $\boldsymbol{P} = [p_{i,j}]$ and $\boldsymbol{Q} = [q_{i,j}]$ as follows:

$$\boldsymbol{P} \coloneqq s_{t_p}(\boldsymbol{M}')$$

and

$$\boldsymbol{Q} \coloneqq (1-q)s_{t_p}(\boldsymbol{M}) + q s_{t_{p,d}}(\boldsymbol{X}\boldsymbol{X}^\top) = (1-q)s_{t_p}(\boldsymbol{M}) + q s_{t_{p,d}\sqrt{d}}(\boldsymbol{V}^{-1/2}\boldsymbol{Y}\boldsymbol{V}^{-1/2}),$$

where recall that $\boldsymbol{X} \in \mathbb{R}^{n \times d}$ is the matrix with ith row equal to $\boldsymbol{x}_i$ and that $t_{p,d}$ is the threshold in (1).

By the definition of $\mathcal{G}(n,p)$ and the independence of entries of $\boldsymbol{M}'$,

$$\mathbb{P}_{\mathcal{G}(n,p)}(G) = \prod_{i<j} p^{a_{i,j}}(1-p)^{1-a_{i,j}} = \mathbb{E}_{\boldsymbol{P}}\left[\prod_{i<j} p_{i,j}^{a_{i,j}}(1-p_{i,j})^{1-a_{i,j}}\right], \tag{8}$$

where we use the standard convention that $0^0 = 1$. By (4) and the definition of $m_{i,j}$, we can write

$$k_{i,j} = (1-q)\,\mathbb{E}[s_{t_p}(m_{i,j})] + q s_{t_{p,d}}(\langle \boldsymbol{x}_i, \boldsymbol{x}_j \rangle) = \mathbb{E}[q_{i,j} \mid \boldsymbol{x}_i, \boldsymbol{x}_j].$$

Then, by (3) and the independence of the $m_{i,j}$,

$$\mathbb{P}_{\mathcal{G}(n,p,d,q)}(G) = \mathbb{E}_{\boldsymbol{Q}}\left[\prod_{i<j} q_{i,j}^{a_{i,j}}(1-q_{i,j})^{1-a_{i,j}}\right].$$

Since $\mathcal{G}(n,p)$ and $\mathcal{G}(n,p,d,q)$ have the same conditional law given $\boldsymbol{P}$ and $\boldsymbol{Q}$, respectively, by Proposition 3.5 we have that

$$\mathrm{TV}(\mathcal{G}(n,p), \mathcal{G}(n,p,d,q)) \leq \mathrm{TV}(\boldsymbol{P}, \boldsymbol{Q}).$$

Recall that the total variation distance is only applied to off-diagonal entries. Equivalently, we can set the diagonals of $\boldsymbol{P}$ and $\boldsymbol{Q}$ to zeros.

Define the zero-diagonal symmetric random matrix $\boldsymbol{B} = [b_{i,j}]$ with $b_{i,j}, 1 \leq i < j \leq n$, following an independent Bernoulli distribution with parameter q; the matrix $\boldsymbol{B}$ is also independent of everything else defined previously. We can then rewrite $\boldsymbol{Q}$ as

$$\boldsymbol{Q} = \mathbb{E}_{\boldsymbol{B}}[(\boldsymbol{1}\boldsymbol{1}^\top - \boldsymbol{B}) \circ s_{t_p}(\boldsymbol{M}) + \boldsymbol{B} \circ s_{t_{p,d}\sqrt{d}}(\boldsymbol{V}^{-1/2}\boldsymbol{Y}\boldsymbol{V}^{-1/2})].$$

For step functions with parameters t and t', we have the simple relation: $s_{t'}(x) = s_t(x + t - t')$. Hence, we can further express $\boldsymbol{Q}$ as

$$\begin{aligned}\boldsymbol{Q} &= \mathbb{E}_{\boldsymbol{B}}[(\boldsymbol{1}\boldsymbol{1}^\top - \boldsymbol{B}) \circ s_{t_p}(\boldsymbol{M}) + \boldsymbol{B} \circ s_{t_p}(\boldsymbol{V}^{-1/2}\boldsymbol{Y}\boldsymbol{V}^{-1/2} + (t_p - t_{p,d}\sqrt{d})\boldsymbol{1}\boldsymbol{1}^\top)] \\ &= \mathbb{E}_{\boldsymbol{B}}[s_{t_p}((\boldsymbol{1}\boldsymbol{1}^\top - \boldsymbol{B}) \circ \boldsymbol{M} + \boldsymbol{B} \circ (\boldsymbol{V}^{-1/2}\boldsymbol{Y}\boldsymbol{V}^{-1/2} + (t_p - t_{p,d}\sqrt{d})\boldsymbol{1}\boldsymbol{1}^\top))].\end{aligned}$$

Let $\delta_{p,d} \coloneqq t_p - t_{p,d}\sqrt{d}$ and $\boldsymbol{H} \coloneqq (\mathbf{1}\mathbf{1}^\top - \boldsymbol{B}) \circ \boldsymbol{M} + \boldsymbol{B} \circ (\boldsymbol{V}^{-1/2}\boldsymbol{Y}\boldsymbol{V}^{-1/2} + \delta_{p,d}\mathbf{1}\mathbf{1}^\top)$. With this notation we have that $\boldsymbol{Q} = \mathbb{E}_{\boldsymbol{B}}[s_{t_p}(\boldsymbol{H})]$.

Applying Proposition 3.4 gives

$$\mathrm{TV}(\boldsymbol{P}, \boldsymbol{Q}) = \mathrm{TV}(s_{t_p}(\boldsymbol{M}'), \mathbb{E}_{\boldsymbol{B}}[s_{t_p}(\boldsymbol{H})]) \le \mathbb{E}_{\boldsymbol{B}} \mathrm{TV}(s_{t_p}(\boldsymbol{M}'), s_{t_p}(\boldsymbol{H})).$$

Further, by Corollary 3.6 we have that

$$\mathbb{E}_{\boldsymbol{B}} \mathrm{TV}(s_{t_p}(\boldsymbol{M}'), s_{t_p}(\boldsymbol{H})) \le \mathbb{E}_{\boldsymbol{B}} \mathrm{TV}(\boldsymbol{M}', \boldsymbol{H}).$$

Let $\boldsymbol{H}' \coloneqq (\mathbf{1}\mathbf{1}^\top - \boldsymbol{B}) \circ \boldsymbol{M}'' + \boldsymbol{B} \circ \boldsymbol{V}^{-1/2}\boldsymbol{M}''\boldsymbol{V}^{-1/2}$. By the triangle inequality of the distance (see, e.g., [24, (4.6)]),

$$\mathbb{E}_{\boldsymbol{B}} \mathrm{TV}(\boldsymbol{M}', \boldsymbol{H}) \le \mathbb{E}_{\boldsymbol{B}} \mathrm{TV}(\boldsymbol{M}', \boldsymbol{H}') + \mathbb{E}_{\boldsymbol{B}} \mathrm{TV}(\boldsymbol{H}', \boldsymbol{H}).$$

By Proposition 3.4 again,

$$\mathbb{E}_{\boldsymbol{B}} \mathrm{TV}(\boldsymbol{H}', \boldsymbol{H}) \le \mathbb{E}_{\boldsymbol{B},\boldsymbol{V}} \mathrm{TV}(\boldsymbol{H}', \boldsymbol{H}).$$

For a fixed value of $\boldsymbol{B}$, both $\boldsymbol{H}$ and $\boldsymbol{H}'$ consist of entries from two matrices. For the (i,j)th entry, if $b_{i,j} = 0$, the entries are from $\boldsymbol{M}''$ and $\boldsymbol{M}$ respectively; if $b_{i,j} = 1$, they come from $\boldsymbol{V}^{-1/2}\boldsymbol{Y}\boldsymbol{V}^{-1/2} + \delta_{p,d}\mathbf{1}\mathbf{1}^\top$ and $\boldsymbol{V}^{-1/2}\boldsymbol{M}''\boldsymbol{V}^{-1/2}$ respectively. In the latter case, when $\boldsymbol{V}$ is fixed, we can multiply both entries by $\sqrt{v_{i,i}v_{j,j}}$, and the distance between the new matrices stays the same by definition. That is, if we let $\boldsymbol{H}'' \coloneqq (\mathbf{1}\mathbf{1}^\top - \boldsymbol{B}) \circ \boldsymbol{M} + \boldsymbol{B} \circ (\boldsymbol{Y} + \delta_{p,d}\boldsymbol{V}^{1/2}\mathbf{1}\mathbf{1}^\top\boldsymbol{V}^{1/2})$, then

$$\mathbb{E}_{\boldsymbol{B},\boldsymbol{V}} \mathrm{TV}(\boldsymbol{H}', \boldsymbol{H}) = \mathbb{E}_{\boldsymbol{B},\boldsymbol{V}} \mathrm{TV}(\boldsymbol{M}'', \boldsymbol{H}'').$$

Putting the bounds from above together, we obtain that

$$\mathrm{TV}(\mathcal{G}(n,p), \mathcal{G}(n,p,d,q)) \le \underbrace{\mathbb{E}_{\boldsymbol{B}} \mathrm{TV}(\boldsymbol{M}', \boldsymbol{H}')}_{E_1} + \underbrace{\mathbb{E}_{\boldsymbol{B},\boldsymbol{V}} \mathrm{TV}(\boldsymbol{M}'', \boldsymbol{H}'')}_{E_2}. \tag{9}$$

The first term on the right depicts the distance caused by normalization, while the second one characterizes the level of independence between edges. We deal with the two expectations E_1 and E_2 in (9) separately in the following two parts. Subsequently, we bring our estimates together to conclude at the end of the section.

3.2.1. Upper bound for the first expectation

By Pinsker's inequality and Proposition 3.4,

$$\begin{aligned} E_1 &\le \mathbb{E}_{\boldsymbol{B}} \sqrt{\frac{1}{2} \mathrm{KL}(\boldsymbol{M}' \parallel \boldsymbol{H}')} \le \mathbb{E}_{\boldsymbol{B}} \sqrt{\frac{1}{2} \mathbb{E}_{\boldsymbol{V}} \mathrm{KL}(\boldsymbol{M}' \parallel \boldsymbol{H}')} \\ &\le \sqrt{\frac{1}{2} \mathbb{E}_{\boldsymbol{B},\boldsymbol{V}} \mathrm{KL}(\boldsymbol{M}' \parallel \boldsymbol{H}')}, \end{aligned}$$

where the last inequality is by Jensen's inequality.

Given $\boldsymbol{B}$ and $\boldsymbol{V}$, the entries of $\boldsymbol{H}'$ are independent. Since the entries of $\boldsymbol{M}'$ are also independent, we have that

$$\mathrm{KL}(\boldsymbol{M}' \parallel \boldsymbol{H}') = \sum_{i<j} b_{i,j}\, \mathrm{KL}(m'_{i,j} \parallel (v_i v_j)^{-1/2} m''_{i,j}),$$

where we use the facts that $m'_{i,j}$ and $m''_{i,j}$ are identically distributed and that the divergence of identical distributions is zero. Therefore,

$$\begin{aligned}\mathbb{E}_{\boldsymbol{B},\boldsymbol{V}}\, \mathrm{KL}(\boldsymbol{M}' \parallel \boldsymbol{H}') &= \sum_{i<j} \mathbb{E}[b_{i,j}]\, \mathbb{E}_{\boldsymbol{V}}\, \mathrm{KL}(m'_{i,j} \parallel (v_i v_j)^{-1/2} m''_{i,j}) \\ &= q \sum_{i<j} \mathbb{E}_{\boldsymbol{V}}\, \mathrm{KL}(m'_{i,j} \parallel (v_i v_j)^{-1/2} m''_{i,j}).\end{aligned}$$

Since $m''_{i,j}$ is a standard normal random variable, $(v_i v_j)^{-1/2} m''_{i,j}$ is a mean zero normal random variable with variance $(v_i v_j)^{-1}$.

The divergence of two normal distributions has an explicit formula given by the following proposition, which appears in most standard texts (see, e.g., [6, exercise 1.30]).

Proposition 3.7. *For two normal distributions with means μ_1, μ_2 and variances σ_1^2, σ_2^2,*

$$\mathrm{KL}(\mathcal{N}(\mu_1, \sigma_1^2) \parallel \mathcal{N}(\mu_2, \sigma_2^2)) = \log \frac{\sigma_2}{\sigma_1} + \frac{\sigma_1^2 + (\mu_1 - \mu_2)^2}{2\sigma_2^2} - \frac{1}{2}.$$

Applying Proposition 3.7, we have that

$$\begin{aligned}\mathrm{KL}(m'_{i,j} \parallel (v_i v_j)^{-1/2} m''_{i,j}) &= \frac{1}{2}(-\log(v_i v_j) + v_i v_j - 1) \\ &= \frac{1}{2}(-\log v_i - \log v_j + v_i v_j - 1).\end{aligned}$$

Since $z_{i,j}$'s are independent standard normal random variables, $v_i d = \sum_{j=1}^{d} z_{i,j}^2$ has a $\chi^2(d)$ distribution. We utilize a lower bound on the expected logarithm of a chi-square random variable shown by the following proposition.

Proposition 3.8. *Suppose that X is a $\chi^2(k)$ random variable. Then,*

$$\mathbb{E}[\log X] \geq \log k - \frac{2}{k}.$$

Proof. We have the following explicit formula for the expected logarithm of X (see, e.g., [6, (B.30)]):

$$\mathbb{E}[\log X] = \psi\left(\frac{k}{2}\right) - \log \frac{1}{2},$$

where ψ is the digamma function defined by $\psi(x) \coloneqq \Gamma'(x)/\Gamma(x)$. The digamma function has the well-known upper and lower bounds (see [1, (2.2)] and references therein):

$$\log x - \frac{1}{x} \le \psi(x) \le \log x - \frac{1}{2x}. \tag{10}$$

By the lower bound in (10),

$$\mathbb{E}[\log X] \ge \log \frac{k}{2} - \frac{2}{k} - \log \frac{1}{2} = \log k - \frac{2}{k}.$$

The claim is hence proved. □

Using $\mathbb{E}[v_i] = \mathbb{E}[v_i d]/d = 1$ and the estimate in Proposition 3.8, we have that

$$\begin{aligned}
&\mathbb{E}_{\boldsymbol{V}} \operatorname{KL}(m'_{i,j} \parallel (v_i v_j)^{-1/2} m''_{i,j}) \\
&\quad = \frac{1}{2}(-\mathbb{E}[\log(v_i d)] + \log d - \mathbb{E}[\log(v_j d)] + \log d + \mathbb{E}[v_i]\,\mathbb{E}[v_j] - 1) \\
&\quad = \log d - \mathbb{E}[\log(v_i d)] \le \frac{2}{d}.
\end{aligned}$$

Therefore, we conclude that

$$E_1 \le \sqrt{\frac{1}{2} q \sum_{i<j} \mathbb{E}_{\boldsymbol{V}} \operatorname{KL}(m'_{i,j} \parallel (v_i v_j)^{-1/2} m''_{i,j})} \le \sqrt{\binom{n}{2} \frac{q}{d}} \le \sqrt{\frac{n^2 q}{2d}}. \tag{11}$$

3.2.2. Upper bound for the second expectation

We now turn to estimating E_2 from (9). We first bound the divergence of $\boldsymbol{M}''$ and $\boldsymbol{H}''$, assuming that $\boldsymbol{B}$ and $\boldsymbol{V}$ are fixed, and then provide an estimate for the distance between them through Pinsker's inequality (Proposition 2.4). The benefit of resorting to the divergence is the chain rule property. Our strategy resembles that of Bubeck and Ganguly [11].

We state the chain rule for the divergence as the following proposition.

Proposition 3.9 (Chain rule)**.** *For joint distributions $\mathcal{P}_{X,Y} = \mathcal{P}_{X|Y}\mathcal{P}_Y$ and $\mathcal{Q}_{X,Y} = \mathcal{Q}_{X|Y}\mathcal{Q}_Y$, the chain rule for the divergence reads*

$$\operatorname{KL}(\mathcal{P}_{X,Y} \parallel \mathcal{Q}_{X,Y}) = \operatorname{KL}(\mathcal{P}_Y \parallel \mathcal{Q}_Y) + \operatorname{KL}(\mathcal{P}_{X|Y} \parallel \mathcal{Q}_{X|Y} \mid \mathcal{P}_Y).$$

For an $n \times n$ matrix $\boldsymbol{A} = [a_{i,j}]$, denote its kth order leading principal submatrix by $\boldsymbol{A}_k$ and let $\boldsymbol{a}_k = (a_{k,1}, \dots, a_{k,k-1})$ be the vector of the first $k-1$ entries in the kth row. We also use $\boldsymbol{Z}_k$ to denote the matrix composed of the first k rows of $\boldsymbol{Z} \in \mathbb{R}^{n \times d}$ and $\boldsymbol{z}_k$ to denote the kth row of $\boldsymbol{Z}$.

Until the end of this section, we assume that $\boldsymbol{B}$ and $\boldsymbol{V}$ are fixed. We similarly use $\boldsymbol{H}''_k$ and $\boldsymbol{M}''_k$ to denote the first k rows of $\boldsymbol{H}''$ and $\boldsymbol{M}''$, and $\boldsymbol{m}''_k$ and $\boldsymbol{h}''_k$ to denote the kth row, respectively. Applying the chain rule to the divergence of $\boldsymbol{H}''_{k+1}$ and $\boldsymbol{M}''_{k+1}$, we obtain

$$\begin{aligned}\mathrm{KL}(\boldsymbol{H}''_{k+1} \parallel \boldsymbol{M}''_{k+1}) = \mathrm{KL}(\boldsymbol{H}''_k \parallel \boldsymbol{M}''_k) \\ + \mathbb{E}_{\boldsymbol{H}''_k} \mathrm{KL}(\boldsymbol{h}''_{k+1} \mid \boldsymbol{H}''_k \parallel \boldsymbol{m}''_{k+1} \mid \boldsymbol{M}''_k = \boldsymbol{H}''_k).\end{aligned}$$

Further, since $\boldsymbol{m}''_{k+1}$ is independent of $\boldsymbol{M}''_k$,

$$\mathbb{E}_{\boldsymbol{H}''_k} \mathrm{KL}(\boldsymbol{h}''_{k+1} \mid \boldsymbol{H}''_k \parallel \boldsymbol{m}''_{k+1} \mid \boldsymbol{M}''_k = \boldsymbol{H}''_k) = \mathbb{E}_{\boldsymbol{H}''_k} \mathrm{KL}(\boldsymbol{h}''_{k+1} \mid \boldsymbol{H}''_k \parallel \boldsymbol{m}''_{k+1}).$$

By Proposition 3.4,

$$\begin{aligned}\mathbb{E}_{\boldsymbol{H}''_k} \mathrm{KL}(\boldsymbol{h}''_{k+1} \mid \boldsymbol{H}''_k \parallel \boldsymbol{m}''_{k+1}) &\le \mathbb{E}_{\boldsymbol{H}''_k, \boldsymbol{Z}_k} \mathrm{KL}(\boldsymbol{h}''_{k+1} \mid \boldsymbol{H}''_k, \boldsymbol{Z}_k \parallel \boldsymbol{m}''_{k+1}) \\ &= \mathbb{E}_{\boldsymbol{Z}_k} \mathrm{KL}(\boldsymbol{h}''_{k+1} \mid \boldsymbol{Z}_k \parallel \boldsymbol{m}''_{k+1}).\end{aligned}$$

The equality holds since $\boldsymbol{h}''_{k+1}$ only depends on $\boldsymbol{Z}_k$ and is independent of other randomness in $\boldsymbol{H}''_k$. Since $\boldsymbol{z}_{k+1}$ is a standard normal random vector, conditioned on $\boldsymbol{Z}_k$, $\boldsymbol{Z}_k \boldsymbol{z}_{k+1}/\sqrt{d}$ is distributed as $\mathcal{N}(\boldsymbol{0}, \boldsymbol{Z}_k \boldsymbol{Z}_k^\top / d)$. By definition, $\boldsymbol{m}''_{k+1}$ has a $\mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_k)$ distribution. Let $\boldsymbol{D}_k := \mathrm{diag}(\boldsymbol{b}_{k+1})$ be the diagonal matrix whose entries are the elements of $\boldsymbol{b}_{k+1}$. Since

$$\begin{aligned}\boldsymbol{h}''_{k+1} &= (\boldsymbol{1} - \boldsymbol{b}_{k+1}) \circ \boldsymbol{m}_{k+1} + \boldsymbol{b}_{k+1} \circ (\boldsymbol{y}_{k+1} + \delta_{p,d}\sqrt{v_{k+1}} \boldsymbol{V}_k^{1/2} \boldsymbol{1}) \\ &= (\boldsymbol{1} - \boldsymbol{b}_{k+1}) \circ \boldsymbol{m}_{k+1} + \boldsymbol{b}_{k+1} \circ \left(\frac{\boldsymbol{Z}_k \boldsymbol{z}_{k+1}}{\sqrt{d}} + \delta_{p,d}\sqrt{v_{k+1}} \boldsymbol{V}_k^{1/2} \boldsymbol{1} \right),\end{aligned}$$

the distribution of $\boldsymbol{h}''_{k+1}$, given $\boldsymbol{Z}_k$, $\boldsymbol{D}_k$, and $\boldsymbol{V}_k$, is $\mathcal{N}(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$ with

$$\boldsymbol{\mu}_k = \delta_{p,d}\sqrt{v_{k+1}} \boldsymbol{D}_k \boldsymbol{V}_k^{1/2} \boldsymbol{1}$$

and

$$\boldsymbol{\Sigma}_k = \boldsymbol{D}_k \left(\frac{\boldsymbol{Z}_k \boldsymbol{Z}_k^\top}{d} \right) \boldsymbol{D}_k^\top + (\boldsymbol{I}_k - \boldsymbol{D}_k)(\boldsymbol{I}_k - \boldsymbol{D}_k)^\top = \boldsymbol{D}_k \left(\frac{\boldsymbol{Z}_k \boldsymbol{Z}_k^\top}{d} \right) \boldsymbol{D}_k + \boldsymbol{I}_k - \boldsymbol{D}_k.$$

As a general form of Proposition 3.7, we have an explicit formula for the divergence of two d-dimensional normal distributions (see, e.g., [43, Exercise 15.13(b)]), stated as follows.

Proposition 3.10. *For two d-dimensional multivariate normal distributions with means $\boldsymbol{\mu}_1, \boldsymbol{\mu}_2$ and covariance matrices $\boldsymbol{\Sigma}_1, \boldsymbol{\Sigma}_2$, we have that*

$$\begin{aligned}&\mathrm{KL}(\mathcal{N}(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1) \parallel \mathcal{N}(\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2)) \\ &\quad = \frac{1}{2}\left((\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^\top \boldsymbol{\Sigma}_2^{-1} (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) + \log \frac{\det(\boldsymbol{\Sigma}_2)}{\det(\boldsymbol{\Sigma}_1)} + \mathrm{Tr}(\boldsymbol{\Sigma}_2^{-1}\boldsymbol{\Sigma}_1) - d \right).\end{aligned}$$

Since $\boldsymbol{m}''_{k+1} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_k)$, applying Proposition 3.10, we have that

$$\begin{aligned}&\mathbb{E}_{\boldsymbol{V}, \boldsymbol{Z}_k} \mathrm{KL}(\boldsymbol{h}''_{k+1} \mid \boldsymbol{Z}_k \parallel \boldsymbol{m}''_{k+1}) \\ &\quad = \frac{1}{2} \mathbb{E}_{\boldsymbol{V}, \boldsymbol{Z}_k} [\delta_{p,d}^2 v_{k+1} \boldsymbol{1}^\top \boldsymbol{D}_k^2 \boldsymbol{V}_k \boldsymbol{1} - \log\det(\boldsymbol{\Sigma}_k) + \mathrm{Tr}(\boldsymbol{\Sigma}_k) - k] \\ &\quad = \frac{1}{2} \mathbb{E}_{\boldsymbol{Z}_k, \boldsymbol{V}_k, v_{k+1}} [\delta_{p,d}^2 v_{k+1} \mathrm{Tr}(\boldsymbol{D}_k \boldsymbol{V}_k) - \log\det(\boldsymbol{\Sigma}_k) + \mathrm{Tr}(\boldsymbol{\Sigma}_k) - k].\end{aligned}$$

Since $v_k d$ has a $\chi^2(d)$ distribution, $\mathbb{E}[v_k] = \mathbb{E}[v_k d]/d = 1$. Further, since $\boldsymbol{V}$ and $\boldsymbol{D}$ are independent, by linearity of expectation we have that

$$\mathbb{E}_{\boldsymbol{V},\boldsymbol{Z}_k} \mathrm{KL}(\boldsymbol{h}''_{k+1} \mid \boldsymbol{Z}_k \parallel \boldsymbol{m}''_{k+1}) = \frac{1}{2}(\delta_{p,d}^2 \operatorname{Tr}(\boldsymbol{D}_k) + \mathbb{E}_{\boldsymbol{Z}_k}[-\log\det(\boldsymbol{\Sigma}_k)] + \mathbb{E}_{\boldsymbol{Z}_k} \operatorname{Tr}(\boldsymbol{\Sigma}_k) - k).$$

Additionally,

$$\begin{aligned}\mathbb{E}_{\boldsymbol{Z}_k} \operatorname{Tr}(\boldsymbol{\Sigma}_k) &= \operatorname{Tr}\left(\boldsymbol{D}_k \,\mathbb{E}_{\boldsymbol{Z}_k}\left[\frac{\boldsymbol{Z}_k\boldsymbol{Z}_k^\top}{d}\right]\boldsymbol{D}_k\right) + k - \operatorname{Tr}(\boldsymbol{D}_k) \\ &= \operatorname{Tr}(\boldsymbol{D}_k) + k - \operatorname{Tr}(\boldsymbol{D}_k) = k.\end{aligned}$$

Therefore, we obtain that

$$\mathbb{E}_{\boldsymbol{V},\boldsymbol{Z}_k} \mathrm{KL}(\boldsymbol{h}''_{k+1} \mid \boldsymbol{Z}_k \parallel \boldsymbol{m}''_{k+1}) = \frac{1}{2}(\delta_{p,d}^2 \operatorname{Tr}(\boldsymbol{D}_k) + \mathbb{E}_{\boldsymbol{Z}_k}[-\log\det(\boldsymbol{\Sigma}_k)]). \quad (12)$$

Next, we derive upper bounds for the two terms in the above display.

An upper bound on $|\delta_{p,d}| = |t_{p,d}\sqrt{d} - t_p|$ is shown in [16], which is stated as the following lemma.

Lemma 3.11 ([16, Lemma 1])**.** *Assume* $0 < p \leq 1/2$ *and* $d \geq \max\{(2/p)^2, 27\}$. *Then*

$$|t_{p,d}\sqrt{d} - t_p| \leq U_{p,d},$$

where

$$U_{p,d} = \kappa_p\sqrt{\log d/d} + \kappa'_p/\sqrt{d}$$

with $\kappa_p = 2\sqrt{2}\Phi^{-1}(1-p)$ *and* $\kappa'_p = 2\sqrt{2\pi}\exp((\Phi^{-1}(1-p/2))^2/2)$.

However, the $\log d$ factor above is an artifact due to the use of concentration inequalities; if we were to apply the lemma directly, it would show up in our final bounds. To address this, we present an improved result, which not only removes the $\log d$ factor in the upper bound but also tightens the upper bound such that it decays as $1/d$ instead of $1/\sqrt{d}$.

Lemma 3.12. *Assume* $0 < p \leq 1/2$. *There exists a constant* C_p *such that*

$$|t_{p,d}\sqrt{d} - t_p| \leq \frac{C_p}{d}.$$

Proof. By rotational invariance, fixing $\boldsymbol{x}_1 = \boldsymbol{e}_1$ and letting $\boldsymbol{x}_2 = \hat{\boldsymbol{z}} \coloneqq \boldsymbol{z}/\|\boldsymbol{z}\|$ with $\boldsymbol{z} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_d)$, we have that

$$\mathbb{P}(\langle \boldsymbol{x}_1, \boldsymbol{x}_2\rangle \geq t_{p,d}) = \mathbb{P}\left(\frac{z_1}{\|\boldsymbol{z}\|} \geq t_{p,d}\right) = p.$$

Then, by symmetry of the distribution, we have

$$\mathbb{P}\left(\frac{z_1}{\|\boldsymbol{z}\|} \leq -t_{p,d}\right) = p.$$

Therefore,

$$\mathbb{P}\left(\frac{z_1^2}{\sum_{i=1}^d z_i^2} \geq t_{p,d}^2\right) = 2p.$$

Since $z_1^2 \sim \chi^2(1)$ and $\sum_{i=2}^d z_i^2 \sim \chi^2(d-1)$ are independent, $z_1^2/\sum_{i=1}^d z_i^2$ has a Beta$(\frac{1}{2}, \frac{d-1}{2})$ distribution. For ease of presentation, we switch from dimension d to considering dimension $d+3$. By the probability density function of the beta distribution we have that

$$2p = \mathbb{P}\left(\frac{z_1^2}{\sum_{i=1}^{d+3} z_i^2} \geq t_{p,d+3}^2\right) = \frac{\Gamma\left(\frac{d+3}{2}\right)}{\Gamma\left(\frac{1}{2}\right)\Gamma\left(\frac{d+2}{2}\right)} \int_{t_{p,d+3}^2}^1 x^{-1/2}(1-x)^{d/2}\,dx.$$

The change of variables $x = z/d$, together with some rearranging, yields

$$\frac{\Gamma\left(\frac{d+2}{2}\right)\sqrt{d}}{\sqrt{2}\Gamma\left(\frac{d+3}{2}\right)} 2p = \frac{1}{\sqrt{2}\Gamma\left(\frac{1}{2}\right)} \int_{t_{p,d+3}^2 d}^d z^{-1/2}\left(1 - \frac{z}{d}\right)^{d/2} dz. \tag{13}$$

Wendel's double inequality (see [44, equation (7)]) states that for $0 < s < 1$,

$$\left(\frac{z}{z+s}\right)^{1-s} \leq \frac{\Gamma(z+s)}{z^s\Gamma(z)} \leq 1.$$

Then, by setting $s = 1/2$ and $z = d/2$, we have that

$$\frac{1}{\sqrt{d/2}} \leq \frac{\Gamma\left(\frac{d}{2}\right)}{\Gamma\left(\frac{d+1}{2}\right)} \leq \frac{\sqrt{2(d+1)}}{d}. \tag{14}$$

Thus,

$$\frac{\Gamma\left(\frac{d+2}{2}\right)\sqrt{d}}{\sqrt{2}\Gamma\left(\frac{d+3}{2}\right)} \geq \sqrt{\frac{d}{d+2}} \geq 1 - \frac{2}{d+2},$$

where the last inequality is due to the fact that $(1-x)^{1/2} \geq 1-x$ for $0 \leq x \leq 1$.

Since

$$\log\left(1 - \frac{z}{d}\right)^{d/2} = \frac{d}{2}\log\left(1 - \frac{z}{d}\right) \leq \frac{d}{2}\left(-\frac{z}{d}\right) = -\frac{z}{2}, \tag{15}$$

we have

$$\begin{aligned}
\frac{1}{\sqrt{2}\Gamma\left(\frac{1}{2}\right)} \int_{t_{p,d+3}^2 d}^d z^{-1/2}\left(1 - \frac{z}{d}\right)^{d/2} dz &\leq \frac{1}{\sqrt{2}\Gamma\left(\frac{1}{2}\right)} \int_{t_{p,d+3}^2 d}^d z^{-1/2} e^{-z/2}\,dz \\
&\leq \frac{1}{\sqrt{2}\Gamma\left(\frac{1}{2}\right)} \int_{t_{p,d+3}^2 d}^{+\infty} z^{-1/2} e^{-z/2}\,dz \\
&= 1 - F\left(t_{p,d+3}^2 d; \frac{1}{2}, \frac{1}{2}\right),
\end{aligned}$$

where $F(x; a, b)$ is the cumulative distribution function of the gamma distribution $\mathrm{Gamma}(a, b)$. Therefore, putting these inequalities back into (13), we obtain that

$$2\left(1 - \frac{2}{d+2}\right)p \le 1 - F\left(dt_{p,d+3}^2; \frac{1}{2}, \frac{1}{2}\right).$$

Since $\mathrm{Gamma}(\frac{1}{2}, \frac{1}{2})$ is also the distribution of a squared standard normal random variable,

$$\left(1 - \frac{2}{d+2}\right)p \le 1 - \Phi(t_{p,d+3}\sqrt{d}).$$

By the monotonicity of the cumulative distribution function, we have that

$$t_{p,d+3}\sqrt{d} \le \Phi^{-1}\left(1 - p + \frac{2p}{d+2}\right).$$

Since $\Phi^{-1}(x)$ is convex for $1/2 \le x < 1$, we have that for $0 < y < 1 - x$,

$$\Phi^{-1}(x+y) \le \Phi^{-1}(x) + y(\Phi^{-1})'(x+y).$$

Let φ be the probability density function of the standard normal distribution. Then,

$$(\Phi^{-1})'(x+y) = \frac{1}{\varphi(\Phi^{-1}(x+y))} = \sqrt{2\pi}\exp\left(\frac{1}{2}(\Phi^{-1}(x+y))^2\right).$$

Additionally, for $d \ge 2$,

$$\Phi^{-1}\left(1 - p + \frac{2p}{d+2}\right) \le \Phi^{-1}\left(1 - \frac{p}{2}\right) = t_{p/2}.$$

Therefore, for $d \ge 2$ we have that

$$t_{p,d+3}\sqrt{d} \le t_p + \frac{2p}{d+2}\sqrt{2\pi}\exp\left(\frac{1}{2}t_{p/2}^2\right).$$

Then,

$$\begin{aligned} t_{p,d+3}\sqrt{d+3} &\le \sqrt{1 + \frac{3}{d}}\left(t_p + \frac{2\sqrt{2\pi}}{d+2}\exp\left(\frac{1}{2}t_{p/2}^2\right)\right) \\ &\le \left(1 + \frac{3}{2d}\right)\left(t_p + \frac{2\sqrt{2\pi}}{d+2}\exp\left(\frac{1}{2}t_{p/2}^2\right)\right). \end{aligned}$$

By assuming $d \ge 6$,

$$\begin{aligned} t_{p,d}\sqrt{d} &\le \left(1 + \frac{3}{2(d-3)}\right)\left(t_p + \frac{2\sqrt{2\pi}}{d-1}\exp\left(\frac{1}{2}t_{p/2}^2\right)\right) \\ &\le \left(1 + \frac{3}{d}\right)\left(t_p + \frac{4\sqrt{2\pi}}{d}\exp\left(\frac{1}{2}t_{p/2}^2\right)\right) \le t_p + \frac{C_p}{d}, \end{aligned}$$

where $C_p = 3(t_p + 2\sqrt{2\pi}\exp(\frac{1}{2}t_{p/2}^2))$.

Similarly, we also have

$$\mathbb{P}\left(\frac{z_1^2}{\sum_{i=1}^{d+3} z_i^2} \le t_{p,d+3}^2\right) = 1 - 2p,$$

which gives

$$\frac{\Gamma\left(\frac{d+2}{2}\right)\sqrt{d}}{\sqrt{2}\Gamma\left(\frac{d+3}{2}\right)}(1-2p) = \frac{1}{\sqrt{2}\Gamma\left(\frac{1}{2}\right)}\int_0^{dt_{p,d+3}^2} x^{-1/2}\left(1-\frac{x}{d}\right)^{d/2} dx.$$

Employing (14) and (15) again, we have

$$\left(1-\frac{2}{d+2}\right)(1-2p) \le F\left(dt_{p,d+3}^2; \frac{1}{2}, \frac{1}{2}\right).$$

Then,

$$\Phi(t_{p,d+3}\sqrt{d}) \ge 1 - \frac{1}{2}\left(1-\left(1-\frac{2}{d+2}\right)(1-2p)\right) = 1 - p - \frac{1-2p}{d+2}.$$

By convexity of $\Phi^{-1}(x)$ in $(1/2, 1)$, for $1/2 < x + y < 1$,

$$\Phi^{-1}(x+y) \ge \Phi^{-1}(x) + y\sqrt{2\pi}\exp\left(\frac{1}{2}(\Phi^{-1}(x))^2\right).$$

Therefore, we have that

$$t_{p,d+3}\sqrt{d+3} \ge t_{p,d+3}\sqrt{d} \ge t_p - \frac{(1-2p)}{d+2}\sqrt{2\pi}\exp\left(\frac{1}{2}t_p^2\right).$$

Hence, by assuming $d \ge 6$,

$$t_{p,d}\sqrt{d} \ge t_p - \frac{C_p}{d},$$

where $C_p = 2(1-2p)\sqrt{2\pi}\exp(\frac{1}{2}t_p^2)$. □

Remark 3.13. From the proof of Lemma 3.12, we see that the lemma actually specifies a convergence rate for the quantile function of a scaled beta distribution to that of a gamma distribution. More general claims and a Berry–Esseen type result can be derived with the same techniques.

Corollary 3.14. *For $p \in (0, 1)$, there exists a constant C_p such that*

$$\delta_{p,d}^2 \le \frac{C_p}{d^2}.$$

Proof. Applying Lemma 3.12 for $0 < p \le 1/2$, we have that there exists a constant C_p' such that

$$\delta_{p,d}^2 \le \frac{C_p'}{d^2}.$$

As before, by fixing $\boldsymbol{x}_1 = \boldsymbol{e}_1$ and letting $\boldsymbol{x}_2 = \hat{\boldsymbol{z}} \coloneqq \boldsymbol{z}/\|\boldsymbol{z}\|$ with $\boldsymbol{z} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_d)$, we have

$$\mathbb{P}(\langle \boldsymbol{x}_1, \boldsymbol{x}_2 \rangle \geq t_{p,d}) = \mathbb{P}\left(\frac{z_1}{\|\boldsymbol{z}\|} \geq t_{p,d} \right) = p.$$

Since $z_1/\|\boldsymbol{z}\|$ has a symmetric distribution, we have $t_{p,d} \geq 0$ for $0 < p \leq 1/2$ and $t_{p,d} \leq 0$ for $1/2 \leq p < 1$. When $1/2 \leq p < 1$,

$$\mathbb{P}\left(\frac{z_1}{\|\boldsymbol{z}\|} \geq -t_{p,d} \right) = \mathbb{P}\left(-\frac{z_1}{\|\boldsymbol{z}\|} \leq t_{p,d} \right) = \mathbb{P}\left(\frac{z_1}{\|\boldsymbol{z}\|} \leq t_{p,d} \right) = 1 - p.$$

Applying Lemma 3.12 again, we obtain that there exists a constant C_p'' such that

$$(-t_{p,d}\sqrt{d} - \Phi^{-1}(p))^2 = (t_{p,d}\sqrt{d} + \Phi^{-1}(p))^2 = (t_{p,d}\sqrt{d} - \Phi^{-1}(1-p))^2 \leq \frac{C_p''}{d^2}.$$

By taking $C_p = C_p' + C_p''$, the claim directly follows. □

We now return to bounding the two terms in (12), starting with the first one. Let $\ell \coloneqq \operatorname{Tr}(\boldsymbol{D}_k)$ be the number of nonzero entries in $\boldsymbol{b}_{k+1}$. Then, ℓ is a function of $\boldsymbol{B}$, is independent of everything else, and has a binomial distribution. Corollary 3.14 gives

$$\delta_{p,d}^2 \operatorname{Tr}(\boldsymbol{D}_k) \leq \frac{C_p \ell}{d^2}. \tag{16}$$

Next, we turn to the upper bound for the second term in (12). Let $\boldsymbol{\Pi}$ be a permutation matrix such that $\boldsymbol{b}_{k+1}\boldsymbol{\Pi} = (1, \ldots, 1, 0, \ldots, 0)$ becomes a vector with its first ℓ entries equal to 1 and the remaining $k - \ell$ entries equal to 0. Then,

$$\boldsymbol{\Pi}^\top \boldsymbol{\Sigma}_k \boldsymbol{\Pi} = \begin{pmatrix} \boldsymbol{S} & \boldsymbol{O} \\ \boldsymbol{O} & \boldsymbol{I}_{k-\ell} \end{pmatrix},$$

where $\boldsymbol{O}$ is the all-zero matrix and $\boldsymbol{S}$ is distributed the same as $\boldsymbol{Z}_\ell \boldsymbol{Z}_\ell^\top / d$. Recall that $\boldsymbol{Z}_\ell \in \mathbb{R}^{\ell \times d}$ is the matrix of the first ℓ rows of $\boldsymbol{Z}$, which has independent standard normal entries. Since $\boldsymbol{\Pi}$ is a permutation matrix, its determinant is either 1 or -1. Therefore, we have that

$$\begin{aligned} \mathbb{E}_{\boldsymbol{Z}_k}[-\log\det(\boldsymbol{\Sigma}_k)] &= \mathbb{E}_{\boldsymbol{Z}_k}[-\log\det(\boldsymbol{\Pi}^\top \boldsymbol{\Sigma}_k \boldsymbol{\Pi})] = \mathbb{E}_{\boldsymbol{Z}_k}[-\log\det(\boldsymbol{S})] \\ &= \mathbb{E}_{\boldsymbol{Z}_\ell}\left[-\log\det\left(\frac{\boldsymbol{Z}_\ell {\boldsymbol{Z}_\ell}^\top}{d} \right) \right]. \end{aligned} \tag{17}$$

So our main focus is bounding (17) from above.

The study of covariance matrices has attracted broad interests in probability and statistics communities. As an example, Cai, Liang, and Zhou [13] showed a central limit theorem (CLT) for the log-determinant of such matrices. An upper bound of the expected negative log-determinant is given in [11] under a general log-concave measure assumption, serving as the major step towards an entropic CLT. We first state their result as the following lemma.

Lemma 3.15 ([11, Lemma 2]). *Let $\boldsymbol{Z}$ be an $n \times d$ random matrix with i.i.d. entries from a log-concave probability measure μ with zero mean and unit variance. There exists an absolute constant $C > 0$ such that for $d \geq Cn^2$,*

$$\mathbb{E}\left[-\log\det\left(\frac{\boldsymbol{Z}\boldsymbol{Z}^\top}{d}\right)\right] \leq C\left(\sqrt{\frac{n}{d}} + \frac{n^2}{d}\right).$$

A direct application of Lemma 3.15 results in an upper bound that is loose for our analysis. It is possible to leverage the normal distribution assumption to obtain an improved estimate, which we implement in Lemma 3.16 below. Applying the lemma results in a better upper bound on the distance.

Lemma 3.16. *Consider an $n \times d$ matrix $\boldsymbol{Z}$ with independent standard normal entries. For $d \geq 2n$,*

$$\mathbb{E}\left[-\log\det\left(\frac{\boldsymbol{Z}\boldsymbol{Z}^\top}{d}\right)\right] \leq \frac{4n}{d} + \frac{n^2}{d}.$$

Remark 3.17. Compared to Lemma 3.15, the improvement thanks to Lemma 3.16 is twofold. First, the upper bound removes the $\sqrt{n/d}$ term, which would be the leading term in our analysis, replacing it with an n/d term. Second, the inequality holds for $d \geq 2n$ rather than $d \geq Cn^2$. We shall see how this improvement is reflected in the upper bound on the distance in the final remarks.

Proof of Lemma 3.16. For a random matrix $\boldsymbol{W}$ following a Wishart distribution $\mathcal{W}_n(\boldsymbol{\Sigma}, d)$, the expectation of its log-determinant has an explicit formula (see, e.g., [6, (B.81)]):

$$\mathbb{E}[\log\det(\boldsymbol{W})] = \sum_{i=1}^{n} \psi\left(\frac{d-i+1}{2}\right) + n\log 2 + \log\det(\boldsymbol{\Sigma}), \tag{18}$$

where ψ is the digamma function. Applying (18) to $\boldsymbol{Z}\boldsymbol{Z}^\top$, we obtain

$$\begin{aligned}\mathbb{E}\left[-\log\det\left(\frac{\boldsymbol{Z}\boldsymbol{Z}^\top}{d}\right)\right] &= -\mathbb{E}[\log\det(\boldsymbol{Z}\boldsymbol{Z}^\top)] + n\log d \\ &= \sum_{i=1}^{n} -\psi\left(\frac{d-i+1}{2}\right) + n\log\frac{d}{2}.\end{aligned}$$

By the lower bound on the digamma function in (10),

$$\begin{aligned}\mathbb{E}\left[-\log\det\left(\frac{\boldsymbol{Z}\boldsymbol{Z}^\top}{d}\right)\right] &\leq \sum_{i=1}^{n}\left(\frac{2}{d-i+1} - \log\frac{d-i+1}{2}\right) + n\log\frac{d}{2} \\ &= \underbrace{\sum_{i=1}^{n}\frac{2}{d-i+1}}_{S_1} - \underbrace{\sum_{i=1}^{n}\log(d-i+1)}_{S_2} + n\log d.\end{aligned} \tag{19}$$

The rest of the proof is devoted to bounding the two sums S_1 and S_2 from above separately.

By the elementary inequality $x \le -\log(1-x)$, S_1 can be bounded from above by

$$S_1 \le 2\sum_{i=1}^{n} -\log\left(1 - \frac{1}{d-i+1}\right) = -2\log\left(\prod_{i=1}^{n} \frac{d-i}{d-i+1}\right) = -2\log\left(1-\frac{n}{d}\right).$$

Further, by $-\log(1-x) \le 2x$ for $0 \le x \le 1/2$, we have that for $d \ge 2n$,

$$S_1 \le -2\log\left(1-\frac{n}{d}\right) \le \frac{4n}{d}. \tag{20}$$

For S_2, we show a lower bound by constructing a continuous integral. Since $\log(d-i+1) \ge \log(d-x)$ for all $x \in [i-1, i)$,

$$S_2 \ge \int_0^n \log(d-x)\,dx = (x\log x - x)\Big|_{d-n}^{d} = -n + d\log d - (d-n)\log(d-n). \tag{21}$$

Bringing the inequalities (20) and (21) into (19), we conclude that for $d \ge 2n$,

$$\begin{aligned}
\mathbb{E}\left[-\log\det\left(\frac{\boldsymbol{Z}\boldsymbol{Z}^\top}{d}\right)\right] &\le \frac{4n}{d} + n - d\log d + (d-n)\log(d-n) + n\log d \\
&= \frac{4n}{d} + n + (d-n)\log\left(1-\frac{n}{d}\right) \\
&\le \frac{4n}{d} + n + (d-n)\left(-\frac{n}{d}\right) \\
&= \frac{4n}{d} + \frac{n^2}{d}.
\end{aligned}$$

The lemma is hence established. □

As a consequence of Lemma 3.16, there is an absolute constant $C > 0$ (we can take $C = 5$), such that for $d \ge 2n \ge 2k \ge 2\ell$,

$$\mathbb{E}\left[-\log\det\left(\frac{\boldsymbol{Z}_\ell\boldsymbol{Z}_\ell^\top}{d}\right)\right] \le \frac{C\ell^2}{d}. \tag{22}$$

Plugging the estimates in (16) and (22) into (12), we get that for $d \ge 2n$,

$$\mathbb{E}_{\boldsymbol{V},\boldsymbol{H}_k''}\,\mathrm{KL}(\boldsymbol{h}_{k+1}'' \mid \boldsymbol{H}_k'' \,\|\, \boldsymbol{m}_{k+1}'') \le \mathbb{E}_{\boldsymbol{V},\boldsymbol{Z}_k}\,\mathrm{KL}(\boldsymbol{h}_{k+1}'' \mid \boldsymbol{Z}_k \,\|\, \boldsymbol{m}_{k+1}'') \le C_p\frac{\ell}{d^2} + C\frac{\ell^2}{d} \tag{23}$$

for constants C, C_p.

Since $b_{i,j}$ has an independent Bernoulli distribution, $\ell = \sum_{j=1}^{k} b_{k+1,j}$ follows a binomial distribution $\mathrm{Bin}(k,q)$. Hence, we have $\mathbb{E}[\ell] = kq$, and $\mathbb{E}[\ell^2]$ can be bounded from above by

$$\mathbb{E}[\ell^2] = \mathbb{V}\mathrm{ar}[\ell] + \mathbb{E}[\ell]^2 = kq(1-q) + (kq)^2 \le kq + k^2q^2.$$

Therefore, by taking the expectation over $\boldsymbol{B}$ in (23), we obtain that for $d \geq 2n$,

$$\begin{aligned}\mathbb{E}_{\boldsymbol{B},\boldsymbol{V},\boldsymbol{H}_k''} \mathrm{KL}(\boldsymbol{h}_{k+1}'' \mid \boldsymbol{H}_k'' \parallel \boldsymbol{m}_{k+1}'') &\leq C_p \frac{\mathbb{E}[\ell]}{d^2} + C \frac{\mathbb{E}[\ell^2]}{d} \\ &\leq C_p \frac{kq}{d^2} + C\left(\frac{kq}{d} + \frac{k^2q^2}{d}\right).\end{aligned}$$

An iterative application of the chain rule yields

$$\begin{aligned}\mathbb{E}_{\boldsymbol{B},\boldsymbol{V}} \mathrm{KL}(\boldsymbol{H}'' \parallel \boldsymbol{M}'') &= \mathbb{E}_{\boldsymbol{B},\boldsymbol{V}}\left[\sum_{k=0}^{n-1} \mathbb{E}_{\boldsymbol{H}_k''} \mathrm{KL}(\boldsymbol{h}_{k+1}'' \mid \boldsymbol{H}_k'' \parallel \boldsymbol{m}_{k+1}'')\right] \\ &= \sum_{k=0}^{n-1} \mathbb{E}_{\boldsymbol{B},\boldsymbol{V},\boldsymbol{H}_k''} \mathrm{KL}(\boldsymbol{h}_{k+1}'' \mid \boldsymbol{H}_k'' \parallel \boldsymbol{m}_{k+1}'').\end{aligned}$$

Therefore, for $d \geq 2n$,

$$\begin{aligned}\mathbb{E}_{\boldsymbol{B},\boldsymbol{V}} \mathrm{KL}(\boldsymbol{H}'' \parallel \boldsymbol{M}'') &\leq \sum_{k=0}^{n-1}\left(C_p \frac{kq}{d^2} + C\left(\frac{kq}{d} + \frac{k^2q^2}{d}\right)\right) \\ &\leq C_p \frac{n^2q}{d^2} + C\left(\frac{n^2q}{d} + \frac{n^3q^2}{d}\right),\end{aligned} \tag{24}$$

for some $C, C_p < \infty$.

By Pinsker's inequality (Proposition 2.4) and Jensen's inequality, we have that

$$\begin{aligned}E_2 \coloneqq \mathbb{E}_{\boldsymbol{B},\boldsymbol{V}} \mathrm{TV}(\boldsymbol{M}'', \boldsymbol{H}'') &\leq \mathbb{E}_{\boldsymbol{B},\boldsymbol{V}} \sqrt{\frac{1}{2} \mathrm{KL}(\boldsymbol{H}'' \parallel \boldsymbol{M}'')} \\ &\leq \sqrt{\frac{1}{2} \mathbb{E}_{\boldsymbol{B},\boldsymbol{V}} \mathrm{KL}(\boldsymbol{H}'' \parallel \boldsymbol{M}'')}.\end{aligned}$$

Hence, using (24) we conclude that there exist constants $C, C_p < \infty$ such that for $d \geq 2n$,

$$E_2 \leq C_p \sqrt{\frac{n^2q}{d^2}} + C\left(\sqrt{\frac{n^2q}{d}} + \sqrt{\frac{n^3q^2}{d}}\right). \tag{25}$$

3.2.3. *Concluding the proof*

Plugging the estimates in (11) and (25) into (9), we have proven that there exist constants C, C_p such that for $d \geq 2n$,

$$\mathrm{TV}(\mathcal{G}(n,p), \mathcal{G}(n,p,d,q)) \leq C_p \sqrt{\frac{n^2q}{d^2}} + C\left(\sqrt{\frac{n^2q}{d}} + \sqrt{\frac{n^3q^2}{d}}\right). \tag{26}$$

We now explain why the $n^3q^2/d \to 0$ regime of Theorem 1.1(a) follows. First, note that $n^3q^2/d = (nq)^2 n/d$. Thus, if we were to have $d < 2n$, then $n^3q^2/d \to 0$

implies that $nq \to 0$, and under this assumption we have already shown in Section 3.1 that the conclusion of Theorem 1.1(a) holds. So we may assume that $d \geq 2n$, in which case the bound in (26) holds. Then, $n^3q^2/d \to 0$ implies that the last term in (26) goes to 0. For the second term, note that $n^2q/d = (n^4q^2/d^2)^{1/2} \leq (n^3q^2/d)^{1/2}$, where we used that $d \geq n$, and so this term also vanishes. This implies that the first term vanishes as well.

Remark 3.18. Using Lemma 3.15 in the place of Lemma 3.16 and following the same derivations, we would similarly obtain that for constants C', C'_p, when $d \geq C'n^2$,

$$\mathrm{TV}(\mathcal{G}(n,p), \mathcal{G}(n,p,d,q)) \leq C'_p\sqrt{\frac{n^2q}{d^2}} + C'\left(\sqrt[4]{\frac{n^3q}{d}} + \sqrt{\frac{n^3q^2}{d}}\right).$$

In this case, the dominating term becomes $\sqrt[4]{n^3q/d}$, resulting in a worse bound than (26), and hence the conclusion follows only in a smaller parameter regime.

Remark 3.19. Utilizing Lemma 3.16, an upper bound of the total variation distance between Wishart and GOE is readily available. Applying chain rule directly to the divergence between $\boldsymbol{Y}$ and $\boldsymbol{M}$, we conclude that for an absolute constant $C > 0$,

$$\mathrm{TV}(\boldsymbol{Y}, \boldsymbol{M}) \leq C\sqrt{\frac{n^3}{d}}.$$

This result removes the first term and log factors of Theorem 2 in [11] in this special case, and coincides with the exact formula given by Rácz and Richey [37] up to a multiplicative constant.

4. Detecting geometry using signed triangles

In this section, we show when detecting geometry in $\mathcal{G}(n,p,d,q)$ is possible and how to detect it. In particular, we demonstrate that the signed triangle statistic, proposed by Bubeck et al. [12], can be used to detect latent geometric structure whenever $n^3q^6/d \to \infty$, thus proving Theorem 1.1(b). Our strategy is to bound the expectation and variance of signed triangle statistic in $\mathcal{G}(n,p)$ and $\mathcal{G}(n,p,d,q)$ respectively, and then apply Chebyshev's inequality.

Consider a simple graph $G = (V, E)$, where V is the set of vertices and $E \subset \binom{V}{2}$ is the set of edges. For a set S, we use $\binom{S}{k}$ to denote the collection of all subsets of S with cardinality k. Let $\boldsymbol{A} = [a_{i,j}]$ be the adjacency matrix of G and write $a_e \coloneqq a_{i,j}$ for any edge $e = \{i,j\} \in \binom{V}{2}$. Let $H = (S, F)$ be another graph with $S \subset V$ and $F \subset \binom{S}{2}$. Define $I_{H|G}$ to be the indicator of H being a subgraph of G. When the graph G is clear from the context, we simply write I_H instead of $I_{H|G}$. Then,

$$I_H = \mathbb{1}\{F \subset E\} = \prod_{e \in F} a_e.$$

Further, for a constant $p \in [0, 1]$, let

$$\lambda_H \coloneqq \prod_{e \in F} (a_e - p)$$

be the signed indicator of the subgraph H.

We first state a lemma that connects the expected signed indicator in $\mathcal{G}(n, p, d, q)$ to that in $\mathcal{G}(n, p, d)$.

Lemma 4.1. *Let $H = (S, F)$ be a fixed graph. The signed indicator satisfies*

$$\mathbb{E}_{\mathcal{G}(n,p,d,q)}[\lambda_H] = q^{|F|}\, \mathbb{E}_{\mathcal{G}(n,p,d)}[\lambda_H].$$

Proof. By conditioning on $\boldsymbol{X}$, we have

$$\mathbb{E}_{\mathcal{G}(n,p,d,q)}[\lambda_H] = \mathbb{E}_{\boldsymbol{X}}\left[\mathbb{E}\left[\prod_{e \in F} (a_e - p) \,\middle|\, \boldsymbol{X}\right]\right] = \mathbb{E}_{\boldsymbol{X}}\left[\prod_{e \in F} \mathbb{E}[a_e - p \mid \boldsymbol{X}]\right],$$

where the last equality is by conditional independence of edges.

Given $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$, $a_{i,j}$ is a Bernoulli random variable with parameter

$$k_{i,j} = (1 - q)p + q s_{t_{p,d}}(\langle \boldsymbol{x}_i, \boldsymbol{x}_j \rangle).$$

Hence,

$$\mathbb{E}[a_{i,j} - p \mid \boldsymbol{X}] = (1-p)k_{i,j} + (-p)(1-k_{i,j}) = k_{i,j} - p = q(s_{t_{p,d}}(\langle \boldsymbol{x}_i, \boldsymbol{x}_j \rangle) - p). \quad (27)$$

Therefore,

$$\begin{aligned} \mathbb{E}_{\mathcal{G}(n,p,d,q)}[\lambda_H] &= \mathbb{E}_{\boldsymbol{X}}\left[\prod_{\{i,j\} \in F} q(s_{t_{p,d}}(\langle \boldsymbol{x}_i, \boldsymbol{x}_j \rangle) - p)\right] \\ &= q^{|F|}\, \mathbb{E}_{\boldsymbol{X}}\left[\prod_{\{i,j\} \in F} (s_{t_{p,d}}(\langle \boldsymbol{x}_i, \boldsymbol{x}_j \rangle) - p)\right]. \end{aligned} \quad (28)$$

On the other hand, for a hard random geometric graph $\mathcal{G}(n, p, d)$,

$$\mathbb{E}_{\mathcal{G}(n,p,d)}[\lambda_H] = \mathbb{E}\left[\prod_{e \in F} (a_e - p)\right] = \mathbb{E}_{\boldsymbol{X}}\left[\prod_{\{i,j\} \in F} (s_{t_{p,d}}(\langle \boldsymbol{x}_i, \boldsymbol{x}_j \rangle) - p)\right].$$

The claim directly follows. □

We first consider the graph $H = (S, F)$ being a complete graph on three vertices, namely a *triangle*. Since in this case the subgraph H is fully determined by its vertex set S, we denote $T_S \coloneqq I_H$ to emphasize the dependency. Given the adjacency matrix $\boldsymbol{A}$ of G, T_S can be expressed as

$$T_S = \prod_{\{i,j\} \subset S} a_{i,j}.$$

Then, the total number triangles in G, denoted by $T_3(G)$, can be written as

$$T_3(G) \coloneqq \sum_{S \in \binom{V}{3}} T_S.$$

The signed triangle and its count in G, following the proposal by Bubeck et al. [12], are defined as

$$\tau_{\{i,j,k\}} \coloneqq \prod_{e \subset \{i,j,k\}} (a_e - p) \quad \text{and} \quad \tau_3(G) \coloneqq \sum_{\{i,j,k\} \subset V} \tau_{\{i,j,k\}}.$$

For a sample random graph G with edge density p, $\tau_3(G)$ is called the *signed triangle statistic*.

To simplify our presentation, for random graphs with edge density p, we let

$$\overline{a}_{i,j} \coloneqq a_{i,j} - \mathbb{E}[a_{i,j}] = a_{i,j} - p \quad \text{and} \quad \overline{s}_{i,j} \coloneqq s_{t_{p,d}}(\langle \boldsymbol{x}_i, \boldsymbol{x}_j \rangle) - p.$$

For $\mathcal{G}(n,p)$, by the analyses in [12, Section 3.1], we have that

$$\mathbb{E}[\tau_3(\mathcal{G}(n,p))] = 0 \quad \text{and} \quad \mathbb{V}\text{ar}[\tau_3(\mathcal{G}(n,p))] = \binom{n}{3} p^3 (1-p)^3.$$

We analyze the expectation and variance of the signed triangle statistic in $\mathcal{G}(n,p,d,q)$ in the following two subsections. Various estimates in the previous work [12] largely simplify our calculations. In addition, for expository purposes, we show that the estimates on the expectation and variance are tight up to constants, and that natural generalizations of signed triangles—signed cycles and signed cliques—are unlikely to improve the detection boundary, both in the special case when $p = 1/2$. Details are provided in Supplementary Material.

4.1. Estimating the expectation

Consider the events

$$E^{\Lambda} \coloneqq \{\langle \boldsymbol{x}_1, \boldsymbol{x}_2 \rangle \geq t_{p,d}, \langle \boldsymbol{x}_1, \boldsymbol{x}_3 \rangle \geq t_{p,d}\} \tag{29}$$

and

$$E^{\Delta} \coloneqq \{\langle \boldsymbol{x}_1, \boldsymbol{x}_2 \rangle \geq t_{p,d}, \langle \boldsymbol{x}_2, \boldsymbol{x}_3 \rangle \geq t_{p,d}, \langle \boldsymbol{x}_3, \boldsymbol{x}_1 \rangle \geq t_{p,d}\}. \tag{30}$$

By rotation invariance on the sphere, we can fix $\boldsymbol{x}_1 = \boldsymbol{e}_1$. Then,

$$\begin{aligned}\mathbb{P}(E^{\Lambda}) &= \mathbb{P}(\langle \boldsymbol{x}_1, \boldsymbol{x}_2 \rangle \geq t_{p,d}, \langle \boldsymbol{x}_1, \boldsymbol{x}_3 \rangle \geq t_{p,d}) = \mathbb{P}(\langle \boldsymbol{e}_1, \boldsymbol{x}_2 \rangle \geq t_{p,d}, \langle \boldsymbol{e}_1, \boldsymbol{x}_3 \rangle \geq t_{p,d}) \\ &= \mathbb{P}(\langle \boldsymbol{e}_1, \boldsymbol{x}_2 \rangle \geq t_{p,d})\, \mathbb{P}(\langle \boldsymbol{e}_1, \boldsymbol{x}_3 \rangle \geq t_{p,d}) = p^2.\end{aligned}$$

The following technical lemma from [12] provides a lower bound on the probability of E^{Δ}.

Lemma 4.2 ([12, Lemma 1]). *For a fixed $p \in (0,1)$, there exists a $C_p > 0$ such that for all $d \geq 1/C_p$,*

$$\mathbb{P}(E^{\Delta}) \geq p^3\left(1 + \frac{C_p}{\sqrt{d}}\right).$$

By (28), the expectation of a signed triangle can be written as

$$\begin{aligned}
&\mathbb{E}_{\mathcal{G}(n,p,d,q)}[\tau_{\{1,2,3\}}] \\
&\quad = q^3\,\mathbb{E}_{\boldsymbol{X}}[(s_{t_{p,d}}(\langle \boldsymbol{x}_1, \boldsymbol{x}_2\rangle) - p)(s_{t_{p,d}}(\langle \boldsymbol{x}_2, \boldsymbol{x}_3\rangle) - p)(s_{t_{p,d}}(\langle \boldsymbol{x}_3, \boldsymbol{x}_1\rangle) - p)] \\
&\quad = q^3(\mathbb{P}(E^{\Delta}) - 3p\,\mathbb{P}(E^{\Lambda}) + 2p^3) = q^3(\mathbb{P}(E^{\Delta}) - p^3).
\end{aligned}$$

Using Lemma 4.2, we have that

$$\mathbb{E}_{\mathcal{G}(n,p,d,q)}[\tau_{\{1,2,3\}}] \geq \frac{C_p q^3}{\sqrt{d}}.$$

Therefore, we conclude that there exists a $C_p > 0$ depending only on p such that for $d \geq 1/C_p$,

$$\mathbb{E}_{\mathcal{G}(n,p,d,q)}[\tau_3(G)] \geq \frac{C_p n^3 q^3}{\sqrt{d}}. \tag{31}$$

4.2. Estimating the variance

The variance of $\tau_3(G)$ for $G \sim \mathcal{G}(n,p,d,q)$ satisfies

$$\begin{aligned}
\mathbb{V}\mathrm{ar}[\tau_3(\mathcal{G}(n,p,d,q))] &= \mathbb{E}[\tau_3(G)^2] - \mathbb{E}[\tau_3(G)]^2 \\
&= \mathbb{E}\left[\left(\sum_{\{i,j,k\}\in V} \tau_{\{i,j,k\}}\right)^2\right] - \left(\sum_{\{i,j,k\}\in V} \mathbb{E}[\tau_{\{i,j,k\}}]\right)^2.
\end{aligned}$$

Expanding the squares of sums and by linearity of expectation, we can decompose the variance into one summation of variances and three summations of covariances; they are grouped by the number of shared vertices. Since the variances or the covariances are identically distributed within each group, we can rewrite them as

$$\begin{aligned}
\mathbb{V}\mathrm{ar}[\tau_3(\mathcal{G}(n,p,d,q))] &= \binom{n}{3} V_{\{1,2,3\},\{1,2,3\}} + \binom{n}{4}\binom{4}{2} V_{\{1,2,3\},\{1,2,4\}} \\
&\quad + \binom{n}{5}\binom{5}{3}\binom{3}{1} V_{\{1,2,3\},\{1,4,5\}} \\
&\quad + \binom{n}{6}\binom{6}{3} V_{\{1,2,3\},\{4,5,6\}},
\end{aligned} \tag{32}$$

where

$$V_{\{i,j,k\},\{i',j',k'\}} \coloneqq \mathbb{E}[\tau_{\{i,j,k\}}\tau_{\{i',j',k'\}}] - \mathbb{E}[\tau_{\{i,j,k\}}]\,\mathbb{E}[\tau_{\{i',j',k'\}}]$$

and the coefficients arise from simple combinatorial computations. We bound the variance and covariances in the following parts respectively.

The signed triangles with vertex sets $\{1, 2, 3\}$ and $\{4, 5, 6\}$ are independent. Hence,

$$V_{\{1,2,3\},\{4,5,6\}} = 0.$$

For two triangles sharing a single vertex, by rotation invariance, we have

$$\begin{aligned}\mathbb{E}[\tau_{\{1,2,3\}}\tau_{\{1,4,5\}}] &= \mathbb{E}[\mathbb{E}[\tau_{\{1,2,3\}}\tau_{\{1,4,5\}} \mid \boldsymbol{x}_1]] = \mathbb{E}[\mathbb{E}[\tau_{\{1,2,3\}} \mid \boldsymbol{x}_1]\,\mathbb{E}[\tau_{\{1,4,5\}} \mid \boldsymbol{x}_1]]\\ &= \mathbb{E}[\mathbb{E}[\tau_{\{1,2,3\}}]\,\mathbb{E}[\tau_{\{1,4,5\}}]] = \mathbb{E}[\tau_{\{1,2,3\}}]\,\mathbb{E}[\tau_{\{1,4,5\}}].\end{aligned}$$

Thus, we have

$$V_{\{1,2,3\},\{1,4,5\}} = 0.$$

For two triangles with exactly the same vertices,

$$\mathbb{E}[(\tau_{\{1,2,3\}})^2] = \mathbb{E}[\overline{a}_{1,2}^2\overline{a}_{2,3}^2\overline{a}_{3,1}^2] = \mathbb{E}_{\boldsymbol{X}}[\mathbb{E}[\overline{a}_{1,2}^2 \mid \boldsymbol{X}]\,\mathbb{E}[\overline{a}_{2,3}^2 \mid \boldsymbol{X}]\,\mathbb{E}[\overline{a}_{3,1}^2 \mid \boldsymbol{X}]] \le 1.$$

Hence,

$$V_{\{1,2,3\},\{1,2,3\}} \le \mathbb{E}[(\tau_{\{1,2,3\}})^2] \le 1.$$

As a last step, for two triangles sharing two vertices,

$$\begin{aligned}&\mathbb{E}[\tau_{\{1,2,3\}}\tau_{\{1,2,4\}}]\\ &\quad= \mathbb{E}_{\boldsymbol{X}}[\mathbb{E}[\overline{a}_{1,2}^2 \mid \boldsymbol{X}]\,\mathbb{E}[\overline{a}_{2,3} \mid \boldsymbol{X}]\,\mathbb{E}[\overline{a}_{3,1} \mid \boldsymbol{X}]\,\mathbb{E}[\overline{a}_{2,4} \mid \boldsymbol{X}]\,\mathbb{E}[\overline{a}_{4,1} \mid \boldsymbol{X}]]\\ &\quad\le \mathbb{E}_{\boldsymbol{X}}[\mathbb{E}[\overline{a}_{2,3} \mid \boldsymbol{X}]\,\mathbb{E}[\overline{a}_{3,1} \mid \boldsymbol{X}]\,\mathbb{E}[\overline{a}_{2,4} \mid \boldsymbol{X}]\,\mathbb{E}[\overline{a}_{4,1} \mid \boldsymbol{X}]]\\ &\quad= q^4\,\mathbb{E}_{\boldsymbol{X}}[\overline{s}_{2,3}\overline{s}_{3,1}\overline{s}_{2,4}\overline{s}_{4,1}].\end{aligned}$$

Further by conditioning on $\boldsymbol{x}_1, \boldsymbol{x}_2$, we have

$$\begin{aligned}\mathbb{E}_{\boldsymbol{X}}[\overline{s}_{2,3}\overline{s}_{3,1}\overline{s}_{2,4}\overline{s}_{4,1}] &= \mathbb{E}_{\boldsymbol{x}_1,\boldsymbol{x}_2}[\mathbb{E}_{\boldsymbol{x}_3}[\overline{s}_{2,3}\overline{s}_{3,1} \mid \boldsymbol{x}_1, \boldsymbol{x}_2]\,\mathbb{E}_{\boldsymbol{x}_4}[\overline{s}_{2,4}\overline{s}_{4,1} \mid \boldsymbol{x}_1, \boldsymbol{x}_2]]\\ &= \mathbb{E}_{\boldsymbol{x}_1,\boldsymbol{x}_2}[\mathbb{E}_{\boldsymbol{x}_3}[\overline{s}_{2,3}\overline{s}_{3,1} \mid \boldsymbol{x}_1, \boldsymbol{x}_2]^2].\end{aligned}$$

A bound in [12, Lemma 4] implies that

$$\mathbb{E}_{\boldsymbol{x}_1,\boldsymbol{x}_2}[\mathbb{E}_{\boldsymbol{x}_3}[\overline{s}_{2,3}\overline{s}_{3,1} \mid \boldsymbol{x}_1, \boldsymbol{x}_2]^2] \le \frac{\pi^2}{d}. \tag{33}$$

Thus, we have that

$$\mathbb{E}[\tau_{\{1,2,3\}}\tau_{\{1,2,4\}}] \le \frac{\pi^2 q^4}{d}.$$

Therefore, we establish that

$$V_{\{1,2,3\},\{1,2,4\}} \le \mathbb{E}[\tau_{\{1,2,3\}}\tau_{\{1,2,4\}}] \le \frac{\pi^2 q^4}{d}.$$

Inserting the above bounds into (32), we conclude that for an absolute constant $C > 0$,

$$\mathbb{V}\mathrm{ar}[\tau_3(\mathcal{G}(n, p, d, q))] \le C\left(n^3 + \frac{n^4 q^4}{d}\right).$$

4.3. Concluding the proof

From the previous analyses, for a fixed $p \in (0,1)$, there exist constants $C_p > 0$ and $C < \infty$ such that

$$\mathbb{E}_{\mathcal{G}(n,p)}[\tau_3(G)] = 0, \quad \mathbb{E}_{\mathcal{G}(n,p,d,q)}[\tau_3(G)] \geq \frac{C_p n^3 q^3}{\sqrt{d}}$$

and

$$V_m := \max\{\mathbb{V}\mathrm{ar}[\tau_3(\mathcal{G}(n,p))], \mathbb{V}\mathrm{ar}[\tau_3(\mathcal{G}(n,p,d,q))]\} \leq C\bigg(n^3 + \frac{n^4 q^4}{d}\bigg).$$

Let $\Delta := \mathbb{E}[\tau_3(\mathcal{G}(n,p,d,q))]$. Chebyshev's inequality implies that for a constant C_p,

$$\mathbb{P}\bigg(\tau_3(\mathcal{G}(n,p,d,q)) \leq \frac{1}{2}\Delta\bigg) \leq \frac{4V_m}{\Delta^2} \leq \frac{C_p}{2}\bigg(\frac{d}{n^3 q^6} + \frac{1}{n^2 q^2}\bigg)$$

and

$$\mathbb{P}\bigg(\tau_3(\mathcal{G}(n,p)) \geq \frac{1}{2}\Delta\bigg) \leq \frac{4V_m}{\Delta^2} \leq \frac{C_p}{2}\bigg(\frac{d}{n^3 q^6} + \frac{1}{n^2 q^2}\bigg).$$

Therefore, we conclude that

$$\begin{aligned} &\mathrm{TV}(\mathcal{G}(n,p), \mathcal{G}(n,p,d,q)) \\ &\quad \geq \mathbb{P}\bigg(\tau_3(\mathcal{G}(n,p,d,q)) \geq \frac{1}{2}\Delta\bigg) - \mathbb{P}\bigg(\tau_3(\mathcal{G}(n,p)) \geq \frac{1}{2}\Delta\bigg) \\ &\quad \geq 1 - C_p\bigg(\frac{d}{n^3 q^6} + \frac{1}{n^2 q^2}\bigg). \end{aligned}$$

Theorem 1.1(b) directly follows. Note that when $n^3 q^6/d \to \infty$, we also have that $n^2 q^2 = (n^3 q^3)^{2/3} \geq (n^3 q^3 \cdot q^3/d)^{2/3} \to \infty$, since $q^3/d \leq 1$.

5. Random dot product graphs

The starting point of this paper is the random geometric graph $\mathcal{G}(n,p,d)$, where the underlying points $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n \in \mathbb{R}^d$ are distributed uniformly on the unit sphere. Our main object of study, $\mathcal{G}(n,p,d,q)$, builds upon $\mathcal{G}(n,p,d)$. Note that in $\mathcal{G}(n,p,d)$ there is an edge between two nodes if and only if the dot product of the corresponding latent vectors is greater than some threshold (see (1)). Models with this property are known as *random dot product graphs* and have been widely studied [38, 4, 39].

A natural variant of $\mathcal{G}(n,p,d)$ is to take $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n \in \mathbb{R}^d$ to be i.i.d. standard normal vectors and to consider the corresponding random dot product graph. In this section we extend our results to this variant; the proofs are kept brief, highlighting only the differences.

At a high level, since the standard normal distribution in $\mathbb{R}^d$ is concentrated on the sphere of radius $\sqrt{d}$, there should also be similar phase transition phenomena under this setting. However, there are important technical differences, due to the errors caused by normalization and different thresholds, so we do not simply reduce it to the previous setting. Please see Remark 5.2 below for details on this point.

Let $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n \in \mathbb{R}^d$ be i.i.d. standard normal vectors and define the threshold $u_{p,d}$ by

$$\mathbb{P}(\langle \boldsymbol{x}_i, \boldsymbol{x}_j \rangle \geq u_{p,d}) = p.$$

Consider the connection function

$$\phi_q(x) = (1-q)p + q s_{u_{p,d}}(x).$$

We denote the random graph generated using $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n \in \mathbb{R}^d$ and this connection function by $\dot{\mathcal{G}}(n,p,d,q)$. The results of Theorem 1.1 also hold under this setting, which we state as the following theorem.

Theorem 5.1 (Detecting geometry)**.** *Let $p \in (0,1)$ be fixed.*

(a) (Impossibility) If $nq \to 0$ or $n^3q^2/d \to 0$, then

$$\mathrm{TV}(\mathcal{G}(n,p), \dot{\mathcal{G}}(n,p,d,q)) \to 0.$$

(b) (Possibility) If $n^3q^6/d \to \infty$, then

$$\mathrm{TV}(\mathcal{G}(n,p), \dot{\mathcal{G}}(n,p,d,q)) \to 1.$$

Remark 5.2. There are two main differences between $\mathcal{G}(n,p,d,q)$ and $\dot{\mathcal{G}}(n,p,d,q)$.

(1) First, the coordinates of $\boldsymbol{x}_i \in \mathbb{R}^d$ are independent in $\dot{\mathcal{G}}(n,p,d,q)$; this is not the case for $\mathcal{G}(n,p,d,q)$.
(2) On the other hand, while $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n \in \mathbb{R}^d$ have rotation symmetry in both models, in $\dot{\mathcal{G}}(n,p,d,q)$ these vectors no longer have unit norm, which must be accounted for.

The first property simplifies the proof of Theorem 5.1(a); however, the second one adds complexity to the proof of Theorem 5.1(b).

Since the proofs are quite similar to those for Theorem 1.1, we only sketch them, highlighting the important adaptations.

5.1. Proof of part (a)

The proof in the regime $nq \to 0$, which is presented in Section 3.1, only uses the property that $\mathbb{P}(\langle \boldsymbol{x}_i, \boldsymbol{x}_j \rangle \geq t_{p,d}) = p$. Hence, it holds directly for $\dot{\mathcal{G}}(n,p,d,q)$ as well.

Now we consider the other regime. If we define $\boldsymbol{R} \coloneqq (1-q)s_{t_p}(\boldsymbol{M}) + q s_{u_{p,d}}(\boldsymbol{Z}\boldsymbol{Z}^\top)$, following the same arguments, we have

$$\mathrm{TV}(\mathcal{G}(n,p), \dot{\mathcal{G}}(n,p,d,q)) \leq \mathrm{TV}(\boldsymbol{P}, \boldsymbol{R}).$$

Note that for $1 \le i < j \le n$,

$$\langle \boldsymbol{z}_i, \boldsymbol{z}_j \rangle = w_{i,j} = \sqrt{d} y_{i,j}.$$

Then we have $\boldsymbol{R} = (1-q) s_{t_p}(\boldsymbol{M}) + q s_{u_{p,d}/\sqrt{d}}(\boldsymbol{Y})$. Let $\boldsymbol{H} := (\mathbf{1}\mathbf{1}^\top - \boldsymbol{B}) \circ \boldsymbol{M} + \boldsymbol{B} \circ (\boldsymbol{Y} + (t_p - u_{p,d}/\sqrt{d})\mathbf{1}\mathbf{1}^\top)$. We can implement the same procedure and obtain the upper bound on the distance, which gives

$$\mathrm{TV}(\boldsymbol{P}, \boldsymbol{R}) \le \mathbb{E}_{\boldsymbol{B}} \, \mathrm{TV}(\boldsymbol{M}', \boldsymbol{H}).$$

The challenge is that we have to bound $|u_{p,d}/\sqrt{d} - t_p|$ from above. The following lemma can be derived as a corollary of Lemma 13.4 in [17].

Lemma 5.3. *There exists a constant C_p, depending only on p, such that*

$$|u_{p,d}/\sqrt{d} - t_p| \le \frac{C_p}{\sqrt{d}}.$$

Plugging this bound into the proof, we conclude that there exist constants C, C_p, such that for $d \ge 2n$,

$$\mathrm{TV}(\mathcal{G}(n,p), \dot{\mathcal{G}}(n,p,d,q)) \le C_p \sqrt{\frac{n^2 q}{d}} + C\left(\sqrt{\frac{n^2 q}{d}} + \sqrt{\frac{n^3 q^2}{d}}\right),$$

similarly to (26). The conclusion follows.

5.2. Proof of part (b)

Compared to the random geometric graph, the proofs are similar but require the estimation of several quantities under a different setting.

We first present technical lemmas for bounding the probabilities of E^Λ and E^Δ defined as counterparts of (29) and (30) respectively, which are derived as corollaries from [17].

Lemma 5.4 (Corollary of [17, Lemma 13.10])**.** *For a fixed $p \in (0,1)$, we have*

$$\mathbb{P}(E^\Lambda) - p^2 \le \frac{8}{d}.$$

Lemma 5.5 (Corollary of [17, Theorem 13.5])**.** *For a fixed $p \in (0,1)$, there exist some constants $C_p, C_p', C_p'' > 0$ depending only on p such that for $d \ge C_p$,*

$$\frac{C_p'}{\sqrt{d}} \le \mathbb{P}(E^\Delta) - p^3 \le \frac{C_p''}{\sqrt{d}}.$$

Plugging Lemma 5.4 and Lemma 5.5 into the estimate for $\mathbb{E}[\tau_{\{1,2,3\}}]$, we obtain the same lower bounds as for $\mathcal{G}(n,p,d,q)$.

For estimating the variance, we still have

$$V_{\{1,2,3\},\{4,5,6\}} = 0 \quad \text{and} \quad V_{\{1,2,3\},\{1,2,3\}} \le 1.$$

However, bounding $V_{\{1,2,3\},\{1,2,4\}}$ and $V_{\{1,2,3\},\{1,4,5\}}$ requires the following estimate, which follows from the proof of [17, Lemma 13.11]:

$$\mathbb{E}_{\boldsymbol{x}_1,\boldsymbol{x}_2}[\mathbb{E}_{\boldsymbol{x}_3}[\overline{s}_{2,3}\overline{s}_{3,1}]^2] \le \frac{80}{d}. \tag{34}$$

Replacing (33) with (34), we obtain the same result for $V_{\{1,2,3\},\{1,2,4\}}$.

For two signed triangles sharing only one vertex, we have

$$\begin{aligned}\mathbb{E}[\tau_{\{1,2,3\}}\tau_{\{1,4,5\}}] &= \mathbb{E}[\mathbb{E}[\tau_{\{1,2,3\}}\tau_{\{1,4,5\}} \mid \boldsymbol{x}_1]] = \mathbb{E}[\mathbb{E}[\tau_{\{1,2,3\}} \mid \boldsymbol{x}_1]^2] \\ &= \mathbb{E}_{\boldsymbol{x}_1}[\mathbb{E}_{\boldsymbol{x}_2,\boldsymbol{x}_3}[\mathbb{E}[\overline{a}_{1,2}\overline{a}_{2,3}\overline{a}_{3,1} \mid \boldsymbol{x}_1,\boldsymbol{x}_2,\boldsymbol{x}_3]]^2] \\ &= q^6\,\mathbb{E}_{\boldsymbol{x}_1}[\mathbb{E}_{\boldsymbol{x}_2,\boldsymbol{x}_3}[\overline{s}_{1,2}\overline{s}_{2,3}\overline{s}_{3,1}]^2].\end{aligned}$$

By Jensen's inequality,

$$\begin{aligned}\mathbb{E}_{\boldsymbol{x}_1}[\mathbb{E}_{\boldsymbol{x}_2,\boldsymbol{x}_3}[\overline{s}_{1,2}\overline{s}_{2,3}\overline{s}_{3,1}]^2] &\le \mathbb{E}_{\boldsymbol{x}_1,\boldsymbol{x}_2}[\mathbb{E}_{\boldsymbol{x}_3}[\overline{s}_{1,2}\overline{s}_{2,3}\overline{s}_{3,1}]^2] \\ &= \mathbb{E}_{\boldsymbol{x}_1,\boldsymbol{x}_2}[\overline{s}_{1,2}^2\,\mathbb{E}_{\boldsymbol{x}_3}[\overline{s}_{2,3}\overline{s}_{3,1}]^2] \\ &\le \mathbb{E}_{\boldsymbol{x}_1,\boldsymbol{x}_2}[\mathbb{E}_{\boldsymbol{x}_3}[\overline{s}_{2,3}\overline{s}_{3,1}]^2].\end{aligned}$$

Using (34) again, we obtain

$$\mathbb{E}[\tau_{\{1,2,3\}}\tau_{\{1,4,5\}}] \le \frac{80q^6}{d}.$$

Therefore, similarly

$$V_{\{1,2,3\},\{1,4,5\}} \le \mathbb{E}[\tau_{\{1,2,3\}}\tau_{\{1,4,5\}}] \le \frac{80q^6}{d}.$$

Inserting the above estimates into (32), we conclude that for an absolute constant C,

$$\mathbb{V}\mathrm{ar}[\tau_3(\dot{\mathcal{G}}(n,p,d,q))] \le C\left(n^3 + \frac{n^4q^4}{d} + \frac{n^5q^6}{d}\right).$$

From the previous analyses, for a fixed $p \in (0,1)$, there exist constants $C_p > 0$ and $C < \infty$ such that

$$\mathbb{E}_{\mathcal{G}(n,p)}[\tau_3(G)] = 0, \quad \mathbb{E}_{\dot{\mathcal{G}}(n,p,d,q)}[\tau_3(G)] \ge \frac{C_p n^3 q^3}{\sqrt{d}}$$

and

$$V_m := \max\{\mathbb{V}\mathrm{ar}[\tau_3(\mathcal{G}(n,p))], \mathbb{V}\mathrm{ar}[\tau_3(\dot{\mathcal{G}}(n,p,d,q))]\} \le C\left(n^3 + \frac{n^4q^4}{d} + \frac{n^5q^6}{d}\right).$$

Repeating the same arguments as before, we conclude that there exists a constant C_p such that for $d \ge C_p$,

$$\mathrm{TV}(\mathcal{G}(n,p), \dot{\mathcal{G}}(n,p,d,q)) \ge 1 - C_p\left(\frac{d}{n^3q^6} + \frac{1}{n^2q^2} + \frac{1}{n}\right).$$

Supplementary Material

In this supplement, we explore whether the bounds of detecting geometry can be improved using generalizations of signed triangles. In particular, we study two families of natural extensions, signed cliques and signed cycles. We provide evidence suggesting that the detection boundary cannot be improved with them.

A. Detecting geometry using signed triangles in the case $p = \frac{1}{2}$

In this case, due to the symmetry of the distribution, the threshold satisfies $t_{1/2,d} = 0$, which no longer depends on the dimension. As a result, the vectors that have an inner product greater than or equal to the threshold with a fixed vector lie in a half space instead of a cone, thus allowing a projection argument. Utilizing the explicit distribution function in the projected space, we are able to obtain asymptotically tight bounds for both the expectation and the variance of the signed triangle statistic. Some results can also be derived as corollaries from more general statements in [17]. However, we include our much simplified proofs for completeness and as preparations for further claims.

Recall that the threshold $t_{p,d}$ is determined by $\mathbb{E}[s_{t_{p,d}}(\langle \boldsymbol{x}_i, \boldsymbol{x}_j \rangle)] = \mathbb{P}(\langle \boldsymbol{x}_i, \boldsymbol{x}_j \rangle \geq t_{p,d}) = p$. When p is set to $1/2$, by symmetry of the distribution, we have $t_{p,d} = 0$. Then, the connection probability becomes

$$k_{i,j} = \frac{1}{2}(1-q) + q s_0(\langle \boldsymbol{x}_i, \boldsymbol{x}_j \rangle).$$

Let $\boldsymbol{x}_1, \boldsymbol{x}_2, \boldsymbol{x}_3$ be independent random vectors uniformly in $\mathbb{S}^{d-1}$. Consider the event

$$E^{\Delta} := \{\langle \boldsymbol{x}_1, \boldsymbol{x}_2 \rangle \geq 0, \langle \boldsymbol{x}_2, \boldsymbol{x}_3 \rangle \geq 0, \langle \boldsymbol{x}_3, \boldsymbol{x}_1 \rangle \geq 0\}. \tag{35}$$

We first give asymptotically tight bounds for $\mathbb{P}(E^{\Delta})$ via a geometric argument.

A.1. Estimating the expectation

Before starting our main discussion, we present a proposition which gives an explicit probability density function for the angle between two uniform random vectors in $\mathbb{S}^{d-1}$. Note that the probability density expressed by sin and gamma functions was also derived in [21, 12] using different approaches.

For two vectors $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^d$, let $\theta(\boldsymbol{x}, \boldsymbol{y}) \in [0, \pi]$ stand for the angle between them. Then,

$$\theta(\boldsymbol{x}, \boldsymbol{y}) = \arccos \frac{\langle \boldsymbol{x}, \boldsymbol{y} \rangle}{\|\boldsymbol{x}\| \|\boldsymbol{y}\|}.$$

If we further assume that $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{S}^{d-1}$, then

$$\theta(\boldsymbol{x}, \boldsymbol{y}) = \arccos \langle \boldsymbol{x}, \boldsymbol{y} \rangle.$$

Proposition A.1. *The angle between two uniformly random vectors in* $\mathbb{S}^{d-1}$ *has the probability density function*

$$h(\theta) = \frac{1}{\zeta} \sin^{d-2} \theta, \quad \theta \in [0, \pi],$$

where

$$\zeta := \int_0^\pi \sin^{d-2} \theta = \frac{\sqrt{\pi}\Gamma\left(\frac{d-1}{2}\right)}{\Gamma\left(\frac{d}{2}\right)}$$

is the normalization factor.

Proof. Let $\boldsymbol{z} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I}_d)$. Then, $\hat{\boldsymbol{z}} := \boldsymbol{z}/\|\boldsymbol{z}\|$ is a uniform random point in $\mathbb{S}^{d-1}$ (see [29], also [30, 27]). By rotation invariance on the sphere, we can fix one vector to be $\boldsymbol{e}_1$, the first vector of the standard basis in $\mathbb{R}^d$. Then, the cumulative distribution function of the angle satisfies

$$F(\theta) = \mathbb{P}(\arccos \langle \boldsymbol{e}_1, \hat{\boldsymbol{z}} \rangle \le \theta) = \mathbb{P}\left(\frac{z_1}{\|\boldsymbol{z}\|} \ge \cos \theta \right).$$

For $\theta \in [0, \pi/2]$,

$$\mathbb{P}\left(\frac{z_1}{\|\boldsymbol{z}\|} \ge \cos \theta \right) = \frac{1}{2} \mathbb{P}\left(\frac{z_1^2}{\sum_{i=1}^d z_i^2} \ge \cos^2 \theta \right).$$

Since the z_i's are standard normal random variables, $z_1^2 \sim \chi^2(1)$ and $\sum_{i=2}^d z_i^2 \sim \chi^2(d-1)$ are independent. Therefore, $z_1^2/\sum_{i=1}^d z_i^2$ is distributed as $\mathrm{Beta}(\frac{1}{2}, \frac{d-1}{2})$. By the definition of the beta distribution,

$$\mathbb{P}\left(\frac{z_1^2}{\sum_{i=1}^d z_i^2} \ge \cos^2 \theta \right) = \frac{1}{\mathrm{B}\left(\frac{1}{2}, \frac{d-1}{2}\right)} \int_{\cos^2 \theta}^1 x^{-1/2}(1-x)^{d/2-3/2}\, dx,$$

where $\mathrm{B}(x, y) = \Gamma(x)\Gamma(y)/\Gamma(x+y)$ is the beta function. Taking the derivative with respect to θ, we have that

$$\begin{aligned} h(\theta) &= \frac{1}{2\mathrm{B}\left(\frac{1}{2}, \frac{d-1}{2}\right)} (-(\cos\theta)^{-1}(1-\cos^2\theta)^{d/2-3/2})(-2\cos\theta\sin\theta) \\ &= \frac{\Gamma\left(\frac{d}{2}\right)}{\sqrt{\pi}\Gamma\left(\frac{d-1}{2}\right)} \sin^{d-2}\theta. \end{aligned}$$

For $\theta \in [\pi/2, \pi]$,

$$\mathbb{P}\left(\frac{z_1}{\|\boldsymbol{z}\|} \ge \cos \theta \right) = \mathbb{P}\left(-\frac{z_1}{\|\boldsymbol{z}\|} \le \cos(\pi - \theta) \right) = \mathbb{P}\left(\frac{z_1}{\|\boldsymbol{z}\|} \le \cos(\pi - \theta) \right),$$

where the last equality is by symmetry of the distribution. Hence,

$$h(\theta) = \frac{\Gamma\left(\frac{d}{2}\right)}{\sqrt{\pi}\Gamma\left(\frac{d-1}{2}\right)} \sin^{d-2}(\pi - \theta) = \frac{\Gamma\left(\frac{d}{2}\right)}{\sqrt{\pi}\Gamma\left(\frac{d-1}{2}\right)} \sin^{d-2}\theta. \qquad \square$$

Lemma A.2. *For E^{Δ} defined in (35), we have*

$$\frac{1}{2\pi\sqrt{2\pi}} \cdot \frac{1}{\sqrt{d}} \le \mathbb{P}(E^{\Delta}) - \frac{1}{8} \le \frac{1}{4\sqrt{\pi}} \cdot \frac{1}{\sqrt{d}}.$$

Proof. We fix the plane determined by $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$, and then project $\boldsymbol{x}_3$ onto this plane. Since no direction in this plane is unique, the projected direction of $\boldsymbol{x}_3$ is uniform on the circle centered at the origin. Conditioning on $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$, $\{\langle \boldsymbol{x}_1, \boldsymbol{x}_3\rangle \ge 0, \langle \boldsymbol{x}_2, \boldsymbol{x}_3\rangle \ge 0\}$ happens if and only if the projection of $\boldsymbol{x}_3$ falls into the intersection of two half spaces with normal vectors $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$. Therefore,

$$\mathbb{P}(\langle \boldsymbol{x}_1, \boldsymbol{x}_3\rangle \ge 0, \langle \boldsymbol{x}_2, \boldsymbol{x}_3\rangle \ge 0 \mid \boldsymbol{x}_1, \boldsymbol{x}_2) = \frac{\pi - \theta(\boldsymbol{x}_1, \boldsymbol{x}_2)}{2\pi}.$$

Hence, the joint probability

$$\begin{aligned}\mathbb{P}(E^{\Delta}) &= \mathbb{E}[\mathbb{P}(\langle \boldsymbol{x}_1, \boldsymbol{x}_3\rangle \ge 0, \langle \boldsymbol{x}_2, \boldsymbol{x}_3\rangle \ge 0 \mid \boldsymbol{x}_1, \boldsymbol{x}_2) \mid \langle \boldsymbol{x}_1, \boldsymbol{x}_2\rangle \ge 0] \\ &= \mathbb{E}\left[\frac{\pi - \theta(\boldsymbol{x}_1, \boldsymbol{x}_2)}{2\pi} \mathbb{1}\left\{\theta(\boldsymbol{x}_1, \boldsymbol{x}_2) \le \frac{\pi}{2}\right\}\right].\end{aligned}$$

The density of $\theta(\boldsymbol{x}_1, \boldsymbol{x}_2)$ is given by Proposition A.1. Then,

$$\begin{aligned}\mathbb{P}(E^{\Delta}) &= \int_0^{\pi/2} \frac{\pi - \theta}{2\pi} \cdot \frac{1}{\zeta} \sin^{d-2}\theta \, d\theta = \int_0^{\pi/2} \frac{\pi/2 + \theta}{2\pi} \cdot \frac{1}{\zeta} \cos^{d-2}\theta \, d\theta \\ &= \frac{1}{2\zeta} \int_0^{\pi/2} \left(\frac{1}{2} + \frac{\theta}{\pi}\right) \cos^{d-2}\theta \, d\theta = \frac{1}{8} + \frac{1}{2\pi\zeta} \int_0^{\pi/2} \theta \cos^{d-2}\theta \, d\theta.\end{aligned}$$

The elementary bounds $2\theta/\pi \le \sin\theta \le \theta$, which hold for $\theta \in [0, \pi/2]$, give

$$\sin\theta \le \theta \le \frac{\pi}{2} \sin\theta. \tag{36}$$

Thus, multiplying by $\cos^{d-2}\theta$ and taking the integral, we have

$$\int_0^{\pi/2} \sin\theta \cos^{d-2}\theta \, d\theta \le \int_0^{\pi/2} \theta \cos^{d-2}\theta \, d\theta \le \frac{\pi}{2} \int_0^{\pi/2} \sin\theta \cos^{d-2}\theta \, d\theta.$$

A simple calculation gives

$$\begin{aligned}\int_0^{\pi/2} \sin\theta \cos^{d-2}\theta \, d\theta &= -\int_0^{\pi/2} \cos^{d-2}\theta \, d\cos\theta = \int_0^1 t^{d-2} \, dt = \left.\frac{t^{d-1}}{d-1}\right|_0^1 \\ &= \frac{1}{d-1}.\end{aligned}$$

The above display together with the definition of ζ yields

$$\frac{1}{\zeta} \int_0^{\pi/2} \sin\theta \cos^{d-2}\theta \, d\theta = \frac{\Gamma\left(\frac{d}{2}\right)}{\sqrt{\pi}(d-1)\Gamma\left(\frac{d-1}{2}\right)} = \frac{\Gamma\left(\frac{d}{2}\right)}{2\sqrt{\pi}\Gamma\left(\frac{d+1}{2}\right)}.$$

By (14),

$$\frac{\sqrt{2}}{\sqrt{d}} = \frac{1}{\sqrt{d/2}} \leq \frac{\Gamma\left(\frac{d}{2}\right)}{\Gamma\left(\frac{d+1}{2}\right)} \leq \sqrt{\frac{d+1}{d}} \cdot \frac{1}{\sqrt{d/2}} \leq \frac{2}{\sqrt{d}}.$$

Putting all the above together, we obtain

$$\frac{1}{8} + \frac{1}{2\pi\sqrt{2\pi}} \cdot \frac{1}{\sqrt{d}} \leq \mathbb{P}(E^{\Delta}) \leq \frac{1}{8} + \frac{1}{4\sqrt{\pi}} \cdot \frac{1}{\sqrt{d}}. \qquad \square$$

With Lemma A.2, we are able to estimate the expectation of the signed triangle statistic in $\mathcal{G}(n, 1/2, d, q)$.

Lemma A.3. *There exist absolute constants $C, C' > 0$ such that*

$$\frac{Cn^3q^3}{\sqrt{d}} \leq \mathbb{E}[\tau_3(\mathcal{G}(n, 1/2, d, q))] \leq \frac{C'n^3q^3}{\sqrt{d}}.$$

Proof. By (28), the expectation of the signed triangle $\tau_{\{1,2,3\}}$ satisfies

$$\begin{aligned}&\mathbb{E}_{\mathcal{G}(n,1/2,d,q)}[\tau_{\{1,2,3\}}]\\ &\quad = q^3\,\mathbb{E}\left[\left(s_0(\langle \boldsymbol{x}_1, \boldsymbol{x}_2\rangle) - \frac{1}{2}\right)\left(s_0(\langle \boldsymbol{x}_2, \boldsymbol{x}_3\rangle) - \frac{1}{2}\right)\left(s_0(\langle \boldsymbol{x}_3, \boldsymbol{x}_1\rangle) - \frac{1}{2}\right)\right].\end{aligned}$$

By rotation invariance on the sphere, we may fix the direction of $\boldsymbol{x}_1$ to be $\boldsymbol{e}_1$. Then,

$$\begin{aligned}\mathbb{E}[s_0(\langle \boldsymbol{x}_1, \boldsymbol{x}_2\rangle)s_0(\langle \boldsymbol{x}_1, \boldsymbol{x}_3\rangle)] &= \mathbb{P}(\langle \boldsymbol{x}_1, \boldsymbol{x}_2\rangle \geq 0, \langle \boldsymbol{x}_1, \boldsymbol{x}_3\rangle \geq 0)\\ &= \mathbb{P}(\langle \boldsymbol{e}_1, \boldsymbol{x}_2\rangle \geq 0, \langle \boldsymbol{e}_1, \boldsymbol{x}_3\rangle \geq 0)\\ &= \mathbb{P}(\langle \boldsymbol{e}_1, \boldsymbol{x}_2\rangle \geq 0)\,\mathbb{P}(\langle \boldsymbol{e}_1, \boldsymbol{x}_3\rangle \geq 0) = \frac{1}{2}\cdot\frac{1}{2} = \frac{1}{4}.\end{aligned}$$

Thus, by expanding the product in (A.1) and using the linearity of expectation, we obtain

$$\mathbb{E}_{\mathcal{G}(n,1/2,d,q)}[\tau_{\{1,2,3\}}] = q^3\left(\mathbb{P}(E^{\Delta}) - \frac{1}{8}\right).$$

Inserting Lemma A.2 yields

$$\frac{1}{2\pi\sqrt{2\pi}} \cdot \frac{q^3}{\sqrt{d}} \leq \mathbb{E}_{\mathcal{G}(n,1/2,d,q)}[\tau_{\{1,2,3\}}] \leq \frac{1}{4\sqrt{\pi}} \cdot \frac{q^3}{\sqrt{d}}. \tag{37}$$

Since all signed triangle indicators are identically distributed,

$$\mathbb{E}_{\mathcal{G}(n,1/2,d,q)}[\tau_3(G)] = \sum_{\{i,j,k\}\subset V} \mathbb{E}_{\mathcal{G}(n,1/2,d,q)}[\tau_{\{i,j,k\}}] = \binom{n}{3}\mathbb{E}_{\mathcal{G}(n,1/2,d,q)}[\tau_{\{1,2,3\}}].$$

The claim directly follows. $\square$

A.2. Estimating the variance

We perform similar analysis for each term of (32) in the special case when $p = 1/2$. The benefit is that by utilizing the symmetry, we obtain matching bounds for the variance, hence showing that the estimates are tight up to constants.

Two signed triangles that do not share any vertices are independent, which implies that

$$\mathbb{E}[\tau_{\{1,2,3\}}\tau_{\{4,5,6\}}] = \mathbb{E}[\tau_{\{1,2,3\}}]\,\mathbb{E}[\tau_{\{4,5,6\}}].$$

Thus, we have that

$$V_{\{1,2,3\},\{4,5,6\}} = 0.$$

For two signed triangles sharing a single vertex, by rotation invariance on the sphere, if we fix the direction of the shared vertex to be $\boldsymbol{e}_1$, they are also independent, which gives

$$\begin{aligned}\mathbb{E}[\tau_{\{1,2,3\}}\tau_{\{1,4,5\}}] &= \mathbb{E}[\mathbb{E}[\tau_{\{1,2,3\}}\tau_{\{1,4,5\}} \mid \boldsymbol{x}_1]] = \mathbb{E}[\mathbb{E}[\tau_{\{1,2,3\}} \mid \boldsymbol{x}_1]\,\mathbb{E}[\tau_{\{1,4,5\}} \mid \boldsymbol{x}_1]] \\ &= \mathbb{E}[\mathbb{E}[\tau_{\{1,2,3\}}]\,\mathbb{E}[\tau_{\{1,4,5\}}]] = \mathbb{E}[\tau_{\{1,2,3\}}]\,\mathbb{E}[\tau_{\{1,4,5\}}].\end{aligned}$$

Therefore,

$$V_{\{1,2,3\},\{1,4,5\}} = 0.$$

For two signed triangles on exactly the same vertices,

$$\mathbb{E}[(\tau_{\{1,2,3\}})^2] = \mathbb{E}[\overline{a}_{1,2}^2\overline{a}_{2,3}^2\overline{a}_{3,1}^2] = \mathbb{E}_{\boldsymbol{X}}[\mathbb{E}[\overline{a}_{1,2}^2 \mid \boldsymbol{X}]\,\mathbb{E}[\overline{a}_{2,3}^2 \mid \boldsymbol{X}]\,\mathbb{E}[\overline{a}_{3,1}^2 \mid \boldsymbol{X}]].$$

When $p = 1/2$, $\overline{a}_{1,2}$ is either $1/2$ or $-1/2$. Hence, $\overline{a}_{1,2}^2 = 1/4$ regardless of $\boldsymbol{X}$ and other randomness of $a_{1,2}$. Therefore,

$$\mathbb{E}[(\tau_{\{1,2,3\}})^2] = \frac{1}{64},$$

which combined with (37) gives that for $d \geq 8/\pi$,

$$\frac{1}{128} \leq \frac{1}{64} - \frac{1}{16\pi}\cdot\frac{q^6}{d} \leq V_{\{1,2,3\},\{1,2,3\}} \leq \frac{1}{64} - \frac{1}{8\pi^3}\cdot\frac{q^6}{d} \leq \frac{1}{64}.$$

That is, $V_{\{1,2,3\},\{1,2,4\}}$ is bounded between absolute constants.

As a last step, for a pair of triangles sharing exactly two vertices, the following lemma provides asymptotically tight bounds for the expectation of their product.

Lemma A.4. *The expectation of two signed triangles sharing two vertices in $\mathcal{G}(n, 1/2, d, q)$ satisfies*

$$\frac{1}{16\pi^2}\cdot\frac{q^4}{d} \leq \mathbb{E}_{\mathcal{G}(n,1/2,d,q)}[\tau_{\{1,2,3\}}\tau_{\{1,2,4\}}] \leq \frac{1}{64}\cdot\frac{q^4}{d}.$$

Proof. By the definition of signed triangles and conditional independence of edges given $\boldsymbol{X}$,

$$\begin{aligned}\mathbb{E}[\tau_{\{1,2,3\}}\tau_{\{1,2,4\}}] &= \mathbb{E}[\overline{a}_{1,2}^2\overline{a}_{2,3}\overline{a}_{3,1}\overline{a}_{2,4}\overline{a}_{4,1}] = \mathbb{E}_{\boldsymbol{X}}[\mathbb{E}[\overline{a}_{1,2}^2\overline{a}_{2,3}\overline{a}_{3,1}\overline{a}_{2,4}\overline{a}_{4,1} \mid \boldsymbol{X}]] \\ &= \mathbb{E}_{\boldsymbol{X}}[\mathbb{E}[\overline{a}_{1,2}^2 \mid \boldsymbol{X}]\,\mathbb{E}[\overline{a}_{2,3} \mid \boldsymbol{X}]\,\mathbb{E}[\overline{a}_{3,1} \mid \boldsymbol{X}]\,\mathbb{E}[\overline{a}_{2,4} \mid \boldsymbol{X}]\,\mathbb{E}[\overline{a}_{4,1} \mid \boldsymbol{X}]] \\ &= \frac{1}{4}\,\mathbb{E}_{\boldsymbol{X}}[\mathbb{E}[\overline{a}_{2,3} \mid \boldsymbol{X}]\,\mathbb{E}[\overline{a}_{3,1} \mid \boldsymbol{X}]\,\mathbb{E}[\overline{a}_{2,4} \mid \boldsymbol{X}]\,\mathbb{E}[\overline{a}_{4,1} \mid \boldsymbol{X}]].\end{aligned}$$

Further by (27) and rotation invariance, we have

$$\begin{aligned}\mathbb{E}[\tau_{\{1,2,3\}}\tau_{\{1,2,4\}}] &= \frac{q^4}{4}\,\mathbb{E}_{\boldsymbol{X}}[\overline{s}_{2,3}\overline{s}_{3,1}\overline{s}_{2,4}\overline{s}_{4,1}] \\ &= \frac{q^4}{4}\,\mathbb{E}_{\boldsymbol{x}_1,\boldsymbol{x}_2}[\mathbb{E}_{\boldsymbol{x}_3,\boldsymbol{x}_4}[\overline{s}_{2,3}\overline{s}_{3,1}\overline{s}_{2,4}\overline{s}_{4,1} \mid \boldsymbol{x}_1,\boldsymbol{x}_2]] \\ &= \frac{q^4}{4}\,\mathbb{E}_{\boldsymbol{x}_1,\boldsymbol{x}_2}[\mathbb{E}_{\boldsymbol{x}_3}[\overline{s}_{2,3}\overline{s}_{3,1} \mid \boldsymbol{x}_1,\boldsymbol{x}_2]\,\mathbb{E}_{\boldsymbol{x}_4}[\overline{s}_{2,4}\overline{s}_{4,1} \mid \boldsymbol{x}_1,\boldsymbol{x}_2]] \\ &= \frac{q^4}{4}\,\mathbb{E}_{\boldsymbol{x}_1,\boldsymbol{x}_2}[\mathbb{E}_{\boldsymbol{x}_3}[\overline{s}_{2,3}\overline{s}_{3,1} \mid \boldsymbol{x}_1,\boldsymbol{x}_2]^2].\end{aligned} \tag{38}$$

The last equality holds since $\boldsymbol{x}_3$ and $\boldsymbol{x}_4$ are identically distributed. Recall that $\theta(\boldsymbol{x}_1,\boldsymbol{x}_2)$ denotes the angle between $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$. The conditional expectation can be written as

$$\begin{aligned}\mathbb{E}_{\boldsymbol{x}_3}[\overline{s}_{2,3}\overline{s}_{3,1} \mid \boldsymbol{x}_1,\boldsymbol{x}_2] &= \mathbb{E}_{\boldsymbol{x}_3}\left[\left(s_0(\langle\boldsymbol{x}_2,\boldsymbol{x}_3\rangle) - \frac{1}{2}\right)\left(s_0(\langle\boldsymbol{x}_3,\boldsymbol{x}_1\rangle) - \frac{1}{2}\right)\right] \\ &= \mathbb{P}(\langle\boldsymbol{x}_2,\boldsymbol{x}_3\rangle \geq 0, \langle\boldsymbol{x}_3,\boldsymbol{x}_1\rangle \geq 0 \mid \boldsymbol{x}_1,\boldsymbol{x}_2) - \frac{1}{4} \\ &= \frac{\pi - \theta(\boldsymbol{x}_1,\boldsymbol{x}_2)}{2\pi} - \frac{1}{4} = \frac{\pi/2 - \theta(\boldsymbol{x}_1,\boldsymbol{x}_2)}{2\pi}.\end{aligned}$$

Therefore, we have

$$\mathbb{E}[\tau_{\{1,2,3\}}\tau_{\{1,2,4\}}] = \frac{q^4}{4}\,\mathbb{E}\left[\left(\frac{\pi/2 - \theta(\boldsymbol{x}_1,\boldsymbol{x}_2)}{2\pi}\right)^2\right]. \tag{39}$$

Using Proposition A.1, we can write

$$\begin{aligned}\mathbb{E}\left[\left(\frac{\pi/2 - \theta(\boldsymbol{x}_1,\boldsymbol{x}_2)}{2\pi}\right)^2\right] &= \int_0^\pi \left(\frac{\pi/2-\theta}{2\pi}\right)^2 h(\theta)\,d\theta \\ &= \frac{1}{\zeta}\int_0^\pi \frac{(\pi/2-\theta)^2}{4\pi^2}\sin^{d-2}\theta\,d\theta \\ &= \frac{1}{\zeta}\int_{-\pi/2}^{\pi/2} \frac{\theta^2}{4\pi^2}\cos^{d-2}\theta\,d\theta \\ &= \frac{1}{2\pi^2\zeta}\int_0^{\pi/2} \theta^2\cos^{d-2}\theta\,d\theta,\end{aligned}$$

where the last line follows from a change of variables and the function inside the integral being even. Applying (36), we have

$$\frac{1}{2\pi^2\zeta}\int_0^{\pi/2}\sin^2\theta\cos^{d-2}\theta\,d\theta \le \mathbb{E}\left[\left(\frac{\pi/2-\theta(\boldsymbol{x}_1,\boldsymbol{x}_2)}{2\pi}\right)^2\right] \le \frac{1}{8\zeta}\int_0^{\pi/2}\sin^2\theta\cos^{d-2}\theta\,d\theta.$$

By the definition of ζ,

$$\int_0^{\pi/2}\cos^{d-2}\theta = \int_0^{\pi/2}\sin^{d-2}\theta = \frac{\zeta}{2} = \frac{\sqrt{\pi}\Gamma\left(\frac{d-1}{2}\right)}{2\Gamma\left(\frac{d}{2}\right)}. \tag{40}$$

Since

$$\int_0^{\pi/2}\sin^2\theta\cos^{d-2}\theta\,d\theta = \int_0^{\pi/2}(1-\cos^2\theta)\cos^{d-2}\theta\,d\theta = \int_0^{\pi/2}\cos^{d-2}\theta\,d\theta - \int_0^{\pi/2}\cos^{d}\theta\,d\theta,$$

by (40) and $\Gamma(z+1)=z\Gamma(z)$,

$$\frac{2}{\zeta}\int_0^{\pi/2}\sin^2\theta\cos^{d-2}\theta\,d\theta = 1-\frac{\int_0^{\pi/2}\cos^{d}\theta\,d\theta}{\int_0^{\pi/2}\cos^{d-2}\theta\,d\theta} = 1-\frac{\Gamma\left(\frac{d+1}{2}\right)}{\Gamma\left(\frac{d+2}{2}\right)}\cdot\frac{\Gamma\left(\frac{d}{2}\right)}{\Gamma\left(\frac{d-1}{2}\right)} = 1-\frac{d-1}{d} = \frac{1}{d}.$$

Putting them together, we obtain

$$\frac{1}{4\pi^2}\cdot\frac{1}{d} \le \mathbb{E}\left[\left(\frac{\pi/2-\theta(\boldsymbol{x}_1,\boldsymbol{x}_2)}{2\pi}\right)^2\right] \le \frac{1}{16}\cdot\frac{1}{d}. \tag{41}$$

The claim directly follows from combining (39) and (41). □

Lemma A.4 together with (37) shows that when $q \le 1/2$, for absolute constants $C, C' > 0$,

$$\frac{Cq^4}{d} \le V_{\{1,2,3\},\{1,2,4\}} \le \mathbb{E}_{\mathcal{G}(n,1/2,d,q)}[\tau_{\{1,2,3\}}\tau_{\{1,2,4\}}] \le \frac{C'q^4}{d}.$$

Putting the estimates together, we conclude with the following lemma.

Lemma A.5. *There exist absolute constants $C, C' > 0$ such that for $d \ge 8/\pi$ and $q \le 1/2$,*

$$C\left(n^3+\frac{n^4q^4}{d}\right) \le \mathbb{V}\mathrm{ar}[\tau_3(\mathcal{G}(n,1/2,d,q))] \le C'\left(n^3+\frac{n^4q^4}{d}\right).$$

A.3. Concluding the proof for $p = \frac{1}{2}$

Combining the estimates in Lemma A.3 and Lemma A.5, we establish that for absolute constants $C > 0$ and $C' < \infty$,

$$\mathbb{E}[\tau_3(\mathcal{G}(n, 1/2))] = 0, \quad \mathbb{E}[\tau_3(\mathcal{G}(n, 1/2, d, q))] \geq \frac{Cn^3q^3}{\sqrt{d}}$$

and

$$V_m := \max\{\mathbb{V}\mathrm{ar}[\tau_3(\mathcal{G}(n, 1/2))], \mathbb{V}\mathrm{ar}[\tau_3(\mathcal{G}(n, 1/2, d, q))]\} \leq C'\bigg(n^3 + \frac{n^4q^4}{d}\bigg).$$

Let $\Delta := \mathbb{E}[\tau_3(\mathcal{G}(n, 1/2, d, q))]$. Chebyshev's inequality implies that for an absolute constant C,

$$\mathbb{P}\bigg(\tau_3(\mathcal{G}(n, 1/2, d, q)) \leq \frac{1}{2}\Delta\bigg) \leq \frac{4V_m}{\Delta^2} \leq \frac{C}{2}\bigg(\frac{d}{n^3q^6} + \frac{1}{n^2q^2}\bigg)$$

and

$$\mathbb{P}\bigg(\tau_3(\mathcal{G}(n, 1/2)) \geq \frac{1}{2}\Delta\bigg) \leq \frac{4V_m}{\Delta^2} \leq \frac{C}{2}\bigg(\frac{d}{n^3q^6} + \frac{1}{n^2q^2}\bigg).$$

Therefore, we conclude that

$$\begin{aligned}
&\mathrm{TV}(\mathcal{G}(n, 1/2), \mathcal{G}(n, 1/2, d, q)) \\
&\quad \geq \mathbb{P}\bigg(\tau_3(\mathcal{G}(n, 1/2, d, q)) \geq \frac{1}{2}\Delta\bigg) - \mathbb{P}\bigg(\tau_3(\mathcal{G}(n, 1/2)) \geq \frac{1}{2}\Delta\bigg) \\
&\quad \geq 1 - C\bigg(\frac{d}{n^3q^6} + \frac{1}{n^2q^2}\bigg).
\end{aligned}$$

Theorem 1.1(b) in the case when $p = 1/2$ directly follows. Note that when $n^3q^6/d \to \infty$, we also have that $n^2q^2 = (n^3q^3)^{2/3} \geq (n^3q^3 \cdot q^3/d)^{2/3} \to \infty$, since $q^3/d \leq 1$.

B. Detecting geometry using signed cliques

The method introduced in Section 4 proves an upper bound for detecting the geometry in $\mathcal{G}(n, p, d, q)$, while providing an asymptotically powerful test that is computationally efficient. However, the upper bound for detection does not match the lower bound in Theorem 1.1(a). As a final remark, we explore whether the possibility results for detection can be improved via generalizations of the signed triangle statistic. Two families of extensions are studied: signed cliques and signed cycles. We show by special examples of subgraphs on four vertices, as well as those on a fixed number of vertices, that it is unlikely the detection boundary can be improved with them.

A first generalization of the signed triangle is by increasing the number of vertices in the set, resulting in the signed induced complete subgraphs of G, which we simply call *signed cliques*. Similar to the case of the signed triangle, let $S \subset V$ be a subset of vertices of G with cardinality $|S| = k$, where $k \leq n$ is fixed. T_S is again the indicator that the edges over the vertex set S form a clique; namely, the induced subgraph is complete. Given the adjacency matrix $\boldsymbol{A}$ of G, T_S can be expressed by

$$T_S = \prod_{\{i,j\} \subset S} a_{i,j}.$$

Then, the total number of cliques of size k in G, denoted by $T_k(G)$, can be written as

$$T_k(G) \coloneqq \sum_{S \in \binom{V}{k}} T_S.$$

For a constant $p \in [0, 1]$, define the signed indicator and its count in G by

$$\tau_S \coloneqq \prod_{\{i,j\} \subset S} (a_{i,j} - p) \quad \text{and} \quad \tau_k(G) \coloneqq \sum_{S \in \binom{V}{k}} \tau_S.$$

We first compute the expectation and variance of the signed clique statistic in $\mathcal{G}(n, p)$.

For $\mathcal{G}(n, p)$, since all edges are independent,

$$\mathbb{E}[\tau_{[k]}] = \prod_{\{i,j\} \subset [k]} \mathbb{E}[a_{i,j} - p] = 0.$$

Then, the expectation of the signed clique statistic satisfies

$$\mathbb{E}[\tau_k(\mathcal{G}(n, p))] = \binom{n}{k} \mathbb{E}[\tau_{[k]}] = 0. \tag{42}$$

Consider two sets of vertices S and S' of size k. If $S = S'$, we have

$$\mathbb{E}[\tau_S \tau_{S'}] = \mathbb{E}[(\tau_S)^2] = \prod_{\{i,j\} \in V_1} \mathbb{E}[(a_{i,j} - p)^2] = (p(1-p))^{\binom{k}{2}} = (p(1-p))^{k(k-1)/2}.$$

For $S \neq S'$, there is at least one signed edge that appears in τ_S but not in $\tau_{S'}$. Suppose this edge is e. By the independence of edges in $\mathcal{G}(n, p)$,

$$\mathbb{E}[\tau_S \tau_{S'}] = \mathbb{E}[a_e - p] \, \mathbb{E}\left[\tau_{S'} \prod_{e' \in \binom{S}{2} \setminus \{e\}} (a_{e'} - p)\right] = 0.$$

Therefore, the variance of the signed clique statistic in $\mathcal{G}(n, p)$ satisfies

$$\begin{aligned} \mathbb{V}\mathrm{ar}[\tau_k(\mathcal{G}(n, p))] &= \mathbb{E}\left[\left(\sum_{S \in \binom{n}{k}} \tau_S\right)^2\right] = \sum_{S, S' \in \binom{n}{k}} \mathbb{E}[\tau_S \tau_{S'}] = \binom{n}{k} \mathbb{E}[(\tau_{[k]})^2] \\ &= \binom{n}{k} (p(1-p))^{k(k-1)/2} \geq C_{k,p} n^k \end{aligned} \tag{43}$$

for some $C_{k,p} > 0$ depending only on k and p.

B.1. Signed quadruples

For $\mathcal{G}(n, p, d, q)$, we start with the special case when $p = 1/2$ and consider the signed clique on four vertices, called the *signed quadruple*.

Theorem B.1. *There exists an absolute constant C such that*

$$|\mathbb{E}[\tau_4(\mathcal{G}(n, 1/2, d, q))]| \leq \frac{Cn^4q^6}{d}.$$

Theorem B.1 together with (42) shows that

$$|\mathbb{E}[\tau_4(\mathcal{G}(n, 1/2, d, q))] - \mathbb{E}[\tau_4(\mathcal{G}(n, 1/2))]| \leq \frac{Cn^4q^6}{d}.$$

As we shall see, a lower bound on the variance of the signed quadruple statistic in $\mathcal{G}(n, 1/2, d, q)$ can be obtained from a more general argument in Lemma B.8, which combined with (43) gives

$$\min\{\mathbb{V}\mathrm{ar}[\tau_4(\mathcal{G}(n, 1/2))], \mathbb{V}\mathrm{ar}[\tau_4(\mathcal{G}(n, 1/2, d, q))]\} \geq C'n^4$$

for some constant $C' > 0$. Therefore, there exists a constant C such that

$$\frac{(\mathbb{E}[\tau_4(\mathcal{G}(n, 1/2, d, q))] - \mathbb{E}[\tau_4(\mathcal{G}(n, 1/2))])^2}{\min\{\mathbb{V}\mathrm{ar}[\tau_4(\mathcal{G}(n, 1/2))], \mathbb{V}\mathrm{ar}[\tau_4(\mathcal{G}(n, 1/2, d, q))]\}} \leq \frac{Cn^4q^{12}}{d^2}.$$

The above display implies that detecting geometry using the previous method with a signed quadruple statistic is only possible if $n^2q^6/d \to \infty$. We see that this is stronger than the condition $n^3q^6/d \to \infty$ given by the signed triangle. Note that we use the lower bound on the minimum of the variances instead of the maximum so that testing either hypothesis is not possible.

We prove Theorem B.1 in the following. A key estimation is that the expected signed quadruple in $\mathcal{G}(n, 1/2, d)$ is at most of the order $1/d$, formally stated as the following lemma.

Lemma B.2. *There exists an absolute constant C such that*

$$|\mathbb{E}_{\mathcal{G}(n,1/2,d)}[\tau_{[4]}]| \leq \frac{C}{d}.$$

The proof of Lemma B.2 is divided into estimating several quantities.

As computed before, for a sample from $\mathcal{G}(n, 1/2, d)$ with adjacency matrix $\boldsymbol{A} = [a_{i,j}]$, by conditioning on $\boldsymbol{x}_1$ and rotation invariance,

$$\mathbb{E}[a_{1,2}a_{1,3}] = \mathbb{E}[\mathbb{E}[a_{1,2}a_{1,3} \mid \boldsymbol{x}_1]] = \mathbb{E}[\mathbb{E}[a_{1,2} \mid \boldsymbol{x}_1] \, \mathbb{E}[a_{1,3} \mid \boldsymbol{x}_1]] = \left(\frac{1}{2}\right)^2.$$

We also have $\mathbb{E}[a_{1,2}] = \frac{1}{2}$ by definition.

The expected signed quadruple in $\mathcal{G}(n, 1/2, d)$ can be written as

$$\begin{aligned}
\mathbb{E}_{\mathcal{G}(n,1/2,d)}[\tau_{[4]}] &= \mathbb{E}\Big[\prod_{i<j}^{4} \overline{a}_{i,j}\Big] = \mathbb{E}\Big[\prod_{i<j}^{4}\Big(a_{i,j} - \frac{1}{2}\Big)\Big] - \Big(\frac{1}{2} - \frac{1}{2}\Big)^6 \\
&= \underbrace{\binom{6}{6}\Big(\mathbb{E}[a_{1,2}a_{1,3}a_{1,4}a_{2,3}a_{2,4}a_{3,4}] - \Big(\frac{1}{2}\Big)^6\Big)}_{Q_1} \\
&\quad - \underbrace{\binom{6}{5}\Big(\frac{1}{2}\Big)\Big(\mathbb{E}[a_{1,2}a_{1,3}a_{2,3}a_{1,4}a_{2,4}] - \Big(\frac{1}{2}\Big)^5\Big)}_{Q_2} \\
&\quad + \underbrace{\binom{6}{4}\frac{1}{5}\Big(\frac{1}{2}\Big)^2\Big(\mathbb{E}[a_{1,3}a_{2,3}a_{1,4}a_{2,4}] - \Big(\frac{1}{2}\Big)^4\Big)}_{Q_3} \\
&\quad + \underbrace{\binom{6}{4}\frac{4}{5}\Big(\frac{1}{2}\Big)^2\Big(\mathbb{E}[a_{1,2}a_{2,3}a_{3,1}a_{1,4}] - \Big(\frac{1}{2}\Big)^4\Big)}_{Q_4} \\
&\quad - \underbrace{\binom{6}{3}\frac{1}{5}\Big(\frac{1}{2}\Big)^3\Big(\mathbb{E}[a_{1,2}a_{2,3}a_{3,1}] - \Big(\frac{1}{2}\Big)^3\Big)}_{Q_5},
\end{aligned} \tag{44}$$

where the fractions are from simple combinatorial calculations. In the following we compute and estimate $Q_1, \dots, Q_5$.

Following the definitions in the proof of Lemma A.2, we define the density

$$h(\theta) := \frac{1}{\zeta}\sin^{d-2}\theta,$$

where $\theta \in [0, \pi]$ and the normalization factor

$$\zeta := \int_0^{\pi} \sin^{d-2}\theta \, d\theta = \frac{\sqrt{\pi}\,\Gamma\big(\frac{d-1}{2}\big)}{\Gamma\big(\frac{d}{2}\big)}.$$

Let

$$\gamma := \int_0^{\pi/2} \frac{\pi/2 - \theta}{2\pi} h(\theta)\, d\theta. \tag{45}$$

By the computation in Lemma A.2, $\mathbb{E}[a_{1,2}a_{2,3}a_{3,1}] - 1/8 = \gamma$. Hence, we have

$$Q_5 = \frac{1}{2}\gamma.$$

The lemma also shows that

$$\frac{1}{2\pi\sqrt{2\pi}} \cdot \frac{1}{\sqrt{d}} \le \gamma \le \frac{1}{4\sqrt{\pi}} \cdot \frac{1}{\sqrt{d}}.$$

By conditional independence of $a_{1,4}$ and $a_{1,2}a_{2,3}a_{3,1}$ and rotation invariance,

$$\begin{aligned}\mathbb{E}[a_{1,2}a_{2,3}a_{3,1}a_{4,1}] &= \mathbb{E}_{\boldsymbol{x}_1}[\mathbb{E}[a_{1,2}a_{2,3}a_{3,1}a_{1,4}|\boldsymbol{x}_1]] \\ &= \mathbb{E}_{\boldsymbol{x}_1}[\mathbb{E}[a_{1,2}a_{2,3}a_{3,1}|\boldsymbol{x}_1]\,\mathbb{E}[a_{1,4}|\boldsymbol{x}_1]] = \frac{1}{2}\,\mathbb{E}[a_{1,2}a_{2,3}a_{3,1}].\end{aligned}$$

Hence, we have

$$Q_4 = \frac{3}{2}\gamma.$$

Let

$$\eta \coloneqq \int_0^{\pi/2} \left(\frac{\pi/2-\theta}{2\pi}\right)^2 h(\theta)\,d\theta. \tag{46}$$

By (41),

$$\frac{1}{4\pi^2}\cdot\frac{1}{d} \le \eta \le \frac{1}{16}\cdot\frac{1}{d}.$$

Lemma B.3. *Let* η *be defined in* (46)*. Then,*

$$\mathbb{E}[a_{1,3}a_{2,3}a_{1,4}a_{2,4}] = \frac{1}{16} + 2\eta.$$

Proof. By conditional independence of $a_{1,3}a_{2,3}$ and $a_{1,4}a_{2,4}$,

$$\begin{aligned}\mathbb{E}[a_{1,3}a_{2,3}a_{1,4}a_{2,4}] &= \mathbb{E}[\mathbb{E}[a_{1,3}a_{2,3}a_{1,4}a_{2,4} \mid \boldsymbol{x}_1,\boldsymbol{x}_2]] \\ &= \mathbb{E}[\mathbb{E}[a_{1,3}a_{2,3} \mid \boldsymbol{x}_1,\boldsymbol{x}_2]\,\mathbb{E}[a_{1,4}a_{2,4} \mid \boldsymbol{x}_1,\boldsymbol{x}_2]] \\ &= \mathbb{E}[\mathbb{E}[a_{1,3}a_{2,3} \mid \boldsymbol{x}_1,\boldsymbol{x}_2]^2].\end{aligned}$$

The last equality is because $\boldsymbol{x}_3$ and $\boldsymbol{x}_4$ are identically distributed.

Similar to the proof of Lemma A.2, we can fix the space spanned by $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$. The angle between them has the density $h(\theta)$ given in Proposition A.1. We have that $a_{1,3}a_{2,3} = 1$ if and only if the projection of $\boldsymbol{x}_3$ onto this plane lies in $[\theta - \pi/2, \pi/2]$. Therefore,

$$\begin{aligned}\mathbb{E}[a_{1,3}a_{2,3}a_{1,4}a_{2,4}] &= \int_0^{\pi}\left(\frac{\pi-\theta}{2\pi}\right)^2 h(\theta)\,d\theta \\ &= \int_0^{\pi/2}\left(\frac{\pi-\theta}{2\pi}\right)^2 h(\theta)\,d\theta + \int_{\pi/2}^{\pi}\left(\frac{\pi-\theta}{2\pi}\right)^2 h(\theta)\,d\theta.\end{aligned} \tag{47}$$

For the first integral in the above display,

$$\int_0^{\pi/2}\left(\frac{\pi-\theta}{2\pi}\right)^2 h(\theta)\,d\theta = \int_0^{\pi/2}\left(\frac{\pi/2+\pi/2-\theta}{2\pi}\right)^2 h(\theta)\,d\theta = \frac{1}{32}+\frac{1}{2}\gamma+\eta.$$

For the second integral, by the symmetry of the sin function,

$$\begin{aligned}\int_{\pi/2}^{\pi}\left(\frac{\pi-\theta}{2\pi}\right)^2 h(\theta)\,d\theta &= \int_0^{\pi/2}\left(\frac{\theta}{2\pi}\right)^2 h(\theta)\,d\theta \\ &= \int_0^{\pi/2}\left(\frac{\pi/2-(\pi/2-\theta)}{2\pi}\right)^2 h(\theta)\,d\theta = \frac{1}{32}-\frac{1}{2}\gamma+\eta.\end{aligned}$$

The claim directly follows by adding them. ☐

By Lemma B.3, we have that

$$Q_3 = \frac{3}{2}\eta.$$

Since

$$\mathbb{E}[a_{1,2}a_{1,3}a_{2,3}a_{1,4}a_{2,4}] = \int_0^{\pi/2} \left(\frac{\pi-\theta}{2\pi}\right)^2 h(\theta)\, d\theta,$$

which equals the first integral in (47), we directly have the following lemma.

Lemma B.4. *Let γ and η be defined in (45) and (46) respectively. Then,*

$$\mathbb{E}[a_{1,2}a_{1,3}a_{2,3}a_{1,4}a_{2,4}] = \frac{1}{32} + \frac{1}{2}\gamma + \eta.$$

Therefore, we have

$$Q_2 = \frac{3}{2}\gamma + 3\eta.$$

Plugging the previous estimates into (42), we obtain that

$$\mathbb{E}[\tau_{[4]}] = Q_1 - Q_2 + Q_3 + Q_4 - Q_5 = Q_1 - \frac{1}{2}\gamma - \frac{3}{2}\eta. \tag{48}$$

An estimation for Q_1 is provided in the following lemma.

Lemma B.5. *Let γ and η be defined before. Then,*

$$\frac{1}{2}\gamma + \frac{1}{2}\eta + \frac{1}{16\pi^2}\cdot\frac{1}{d} \le \mathbb{E}\left[\prod_{i<j}^{4} a_{i,j}\right] - \frac{1}{64} \le \frac{1}{2}\gamma + \frac{1}{2}\eta + \frac{1}{8\pi}\cdot\frac{1}{d}.$$

The proof of Lemma B.5 involves extending the argument in the proof of Lemma A.2 to a three-dimensional subspace. Before proving Lemma B.5, we show the following claim concerning the distribution of the angle between a uniform random vector in $\mathbb{R}^d$ and an arbitrary two-dimensional plane.

Proposition B.6. *Let $\boldsymbol{x}$ be a uniform random point in $\mathbb{S}^{d-1}$. Let $\varphi \in [0, \pi/2]$ be the angle between the vector $\boldsymbol{x}$ and any fixed 2-dimensional subspace. Then, the density $g(\varphi)$ satisfies*

$$g(\varphi) = (d-2)\sin^{d-3}\varphi\cos\varphi.$$

Proof. Let $\boldsymbol{z} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_d)$ be a d-dimensional random vector. Then, $\hat{\boldsymbol{z}} \coloneqq \boldsymbol{z}/\|\boldsymbol{z}\|$ is a uniform random point in $\mathbb{S}^{d-1}$. By rotation invariance, we can fix the plane to be that spanned by the first two vectors of the standard basis. The projection of $\hat{\boldsymbol{z}}$ onto this plane is $\tilde{\boldsymbol{z}} = (z_1, z_2, 0, \dots, 0)/\|\boldsymbol{z}\|$. Then, the angle between $\boldsymbol{x}$ and the plane is equal to the angle between $\hat{\boldsymbol{z}}$ and $\widetilde{\boldsymbol{z}}$. Hence, the cumulative distribution function satisfies

$$F(\varphi) = \mathbb{P}\left(\arccos\frac{\hat{\boldsymbol{z}}\cdot\widetilde{\boldsymbol{z}}}{\|\widetilde{\boldsymbol{z}}\|} \le \varphi\right) = \mathbb{P}\left(\frac{\hat{\boldsymbol{z}}\cdot\widetilde{\boldsymbol{z}}}{\|\widetilde{\boldsymbol{z}}\|} \ge \cos\varphi\right) = \mathbb{P}\left(\frac{z_1^2+z_2^2}{\sum_{i=1}^d z_i^2} \ge \cos^2\varphi\right).$$

Since the z_i's are standard normal random variables, we have $z_1^2+z_2^2 \sim \chi^2(2)$ and $\sum_{i=3}^d z_i^2 \sim \chi^2(d-2)$, and these are independent. Therefore, $(z_1^2+z_2^2)/\sum_{i=1}^d z_i^2$ has a Beta$(1, \frac{d-2}{2})$ distribution. Hence, by the definition of the beta distribution,

$$\mathbb{P}\left(\frac{z_1^2+z_2^2}{\sum_{i=1}^d z_i^2} \geq \cos^2\varphi\right) = \frac{\Gamma\left(\frac{d}{2}\right)}{\Gamma\left(\frac{d-2}{2}\right)} \int_{\cos^2\varphi}^1 (1-x)^{d/2-2}\,dx$$
$$= \frac{d-2}{2} \int_{\cos^2\varphi}^1 (1-x)^{d/2-2}\,dx.$$

Taking the derivative with respect to φ, we obtain

$$g(\varphi) = -\frac{d-2}{2}(1-\cos^2\varphi)^{d/2-2}(-2\cos\varphi\sin\varphi) = (d-2)\sin^{d-3}\varphi\cos\varphi. \quad \square$$

Proof of Lemma B.3. Consider the space spanned by $\boldsymbol{x}_1, \boldsymbol{x}_2, \boldsymbol{x}_3$. Without loss of generality, we can fix the coordinates as follows:

$$\boldsymbol{x}_1 = (1, 0, 0, 0, \ldots, 0),$$
$$\boldsymbol{x}_2 = (\cos\theta, \sin\theta, 0, \ldots, 0),$$
$$\boldsymbol{x}_3 = (\cos\varphi\cos\psi, \cos\varphi\sin\psi, \sin\varphi, 0, \ldots, 0).$$

By symmetry on the sphere, we can constrain the parameters in the following space:

$$\theta \in [0, \pi],$$
$$\psi \in [-\pi, \pi],$$
$$\varphi \in [0, \pi/2].$$

Let $f(\theta, \psi, \varphi)$ be the probability density function. Then, by independence of the vectors,

$$f(\theta, \psi, \varphi) = \frac{1}{2\pi} h(\theta) g(\varphi) = \frac{d-2}{2\pi\zeta} \sin^{d-2}\theta \sin^{d-3}\varphi\cos\varphi.$$

Denoting by $\theta(\boldsymbol{x}, \boldsymbol{y}) \in [0, \pi]$ the angle between two d-dimensional vectors, we also have

$$\cos\theta(\boldsymbol{x}_1, \boldsymbol{x}_2) = \langle \boldsymbol{x}_1, \boldsymbol{x}_2 \rangle = \cos\theta, \tag{49}$$
$$\cos\theta(\boldsymbol{x}_1, \boldsymbol{x}_3) = \langle \boldsymbol{x}_1, \boldsymbol{x}_3 \rangle = \cos\varphi\cos\psi, \tag{50}$$
$$\cos\theta(\boldsymbol{x}_2, \boldsymbol{x}_3) = \langle \boldsymbol{x}_2, \boldsymbol{x}_3 \rangle = \cos\varphi\cos\psi\cos\theta + \cos\varphi\sin\psi\sin\theta = \cos\varphi\cos(\psi - \theta). \tag{51}$$

The event $1 \sim 2$ happens if and only if $\theta \in [0, \pi]$. Vertex 3 is connected to both 1 and 2 if and only if the projection of $\boldsymbol{x}_3$ onto the plane determined by $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$ forms an angle no greater than $\pi/2$ with both $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$. Therefore, we have $\psi \in [\theta - \pi/2, \pi/2]$. The last vertex is connected to all of them if and only if the direction of $\boldsymbol{x}_4$ falls in the spherical triangle determined by the three half planes with normal vectors $\boldsymbol{x}_1, \boldsymbol{x}_2, \boldsymbol{x}_3$ respectively. As a well-known fact (see, e.g., [42, proposition 99]), the surface area of the spherical triangle equals the spherical excess defined by

$$\begin{aligned} S &\coloneqq 2\pi - \theta(\boldsymbol{x}_1, \boldsymbol{x}_2) - \theta(\boldsymbol{x}_1, \boldsymbol{x}_3) - \theta(\boldsymbol{x}_2, \boldsymbol{x}_3) \\ &= \Big(\pi - \theta\Big) + \Big(\frac{\pi}{2} - \theta(\boldsymbol{x}_1, \boldsymbol{x}_3)\Big) + \Big(\frac{\pi}{2} - \theta(\boldsymbol{x}_2, \boldsymbol{x}_3)\Big). \end{aligned}$$

Since the surface area of the sphere is 4π, the probability that the four vertices form a clique is

$$\begin{aligned} &\int_0^{\pi/2}\int_0^{\pi/2}\int_{\theta-\pi/2}^{\pi/2} \frac{S}{4\pi} f(\theta,\psi,\varphi)\, d\psi\, d\theta\, d\varphi \\ &\quad = \underbrace{\int_0^{\pi/2}\int_0^{\pi/2}\int_{\theta-\pi/2}^{\pi/2} \frac{\pi-\theta}{4\pi} f(\theta,\psi,\varphi)\, d\psi\, d\theta\, d\varphi}_{I_1} \\ &\qquad + \underbrace{\int_0^{\pi/2}\int_0^{\pi/2}\int_{\theta-\pi/2}^{\pi/2} \frac{\pi/2-\theta(\boldsymbol{x}_1,\boldsymbol{x}_3)}{4\pi} f(\theta,\psi,\varphi)\, d\psi\, d\theta\, d\varphi}_{I_2} \\ &\qquad + \underbrace{\int_0^{\pi/2}\int_0^{\pi/2}\int_{\theta-\pi/2}^{\pi/2} \frac{\pi/2-\theta(\boldsymbol{x}_2,\boldsymbol{x}_3)}{4\pi} f(\theta,\psi,\varphi)\, d\psi\, d\theta\, d\varphi}_{I_3}. \end{aligned} \tag{52}$$

We deal with the three integrals separately as follows.

By integrating over ψ and φ,

$$\begin{aligned} I_1 &= \frac{1}{2}\int_0^{\pi/2}\left(\frac{\pi-\theta}{2\pi}\right)^2 h(\theta)\, d\theta = \frac{1}{2}\int_0^{\pi/2}\left(\frac{\pi/2+(\pi/2-\theta)}{2\pi}\right)^2 h(\theta)\, d\theta \\ &= \frac{1}{2}\left(\frac{1}{32} + \frac{1}{2}\gamma + \eta\right). \end{aligned}$$

Plugging (50) into I_2, we have

$$\begin{aligned} I_2 &= \frac{1}{4\pi}\int_0^{\pi/2}\int_0^{\pi/2}\int_{\theta-\pi/2}^{\pi/2}\left(\frac{\pi}{2} - \arccos(\cos\varphi\cos\psi)\right) f(\theta,\psi,\varphi)\, d\psi\, d\theta\, d\varphi \\ &= \frac{1}{4\pi}\int_0^{\pi/2}\int_0^{\pi/2}\int_{\theta-\pi/2}^{\pi/2} \arcsin(\cos\varphi\cos\psi) f(\theta,\psi,\varphi)\, d\psi\, d\theta\, d\varphi. \end{aligned} \tag{53}$$

The Taylor expansion of arcsin gives

$$\arcsin x = \sum_{n=0}^{\infty} a_n x^{2n+1},$$

where

$$a_n := \frac{(2n)!}{4^n (n!)^2 (2n+1)}.$$

Hence, we get

$$\arcsin(\cos\varphi\cos\psi) = \sum_{n=0}^{\infty} a_n \cos^{2n+1}\varphi \cos^{2n+1}\psi. \tag{54}$$

Inserting the expansion (54) into (53) and interchanging the summation and integration, we have

$$\begin{aligned} I_2 &= \frac{1}{4\pi}\int_0^{\pi/2}\int_0^{\pi/2}\int_{\theta-\pi/2}^{\pi/2}\sum_{n=0}^{\infty} a_n (\cos^{2n+1}\varphi)(\cos^{2n+1}\psi) f(\theta,\psi,\varphi)\, d\psi\, d\theta\, d\varphi \\ &= \frac{1}{8\pi^2}\sum_{n=0}^{\infty} a_n \int_0^{\pi/2} g(\varphi)\cos^{2n+1}\varphi\, d\varphi \int_0^{\pi/2} h(\theta)\int_{\theta-\pi/2}^{\pi/2}\cos^{2n+1}\psi\, d\psi\, d\theta. \end{aligned}$$

Since

$$\int_{\theta-\pi/2}^{\pi/2}\cos^{2n+1}\psi\, d\psi = \int_0^{\pi/2}\cos^{2n+1}\psi\, d\psi + \int_{\theta-\pi/2}^{0}\cos^{2n+1}\psi\, d\psi, \tag{55}$$

the integral I_2 can also be split into two integrals accordingly:

$$\begin{aligned} I_2 = &\underbrace{\frac{1}{8\pi^2}\sum_{n=0}^{\infty} a_n \int_0^{\pi/2} g(\varphi)\cos^{2n+1}\varphi\, d\varphi \int_0^{\pi/2} h(\theta)\int_0^{\pi/2}\cos^{2n+1}\psi\, d\psi\, d\theta}_{I_{2(a)}} \\ &+ \underbrace{\frac{1}{8\pi^2}\sum_{n=0}^{\infty} a_n \int_0^{\pi/2} g(\varphi)\cos^{2n+1}\varphi\, d\varphi \int_0^{\pi/2} h(\theta)\int_{\theta-\pi/2}^{0}\cos^{2n+1}\psi\, d\psi\, d\theta}_{I_{2(b)}}. \end{aligned}$$

We deal with $I_{2(a)}$ and $I_{2(b)}$ separately.

Using the definition of $g(\varphi)$ in Proposition B.6 and by a change of variables $x = \sin^2\varphi$, we have

$$\begin{aligned} \int_0^{\pi/2} g(\varphi)\cos^{2n+1}\varphi\, d\varphi &= (d-2)\int_0^{\pi/2}\cos^{2n+1}\varphi\sin^{d-3}\varphi\cos\varphi\, d\varphi \\ &= \frac{d-2}{2}\int_0^1 (1-x)^{n+1/2} x^{d/2-2}\, dx \\ &= \frac{d-2}{2}\mathrm{B}\left(\frac{d}{2}-1, n+\frac{3}{2}\right) = \frac{(d-2)\Gamma\big(\frac{d}{2}-1\big)\Gamma\big(n+\frac{3}{2}\big)}{2\Gamma\big(n+\frac{d}{2}+\frac{1}{2}\big)} \\ &= \frac{\Gamma\big(\frac{d}{2}\big)\Gamma\big(n+\frac{3}{2}\big)}{\Gamma\big(n+\frac{d}{2}+\frac{1}{2}\big)}, \end{aligned}$$

where the last equality is due to the identity $x\Gamma(x) = \Gamma(x+1)$.

By (40),

$$\int_0^{\pi/2} \cos^{2n+1}\psi \, d\psi = \frac{\sqrt{\pi}\Gamma(n+1)}{2\Gamma\left(n+\frac{3}{2}\right)}.$$

Combining them and rearranging the terms, we have

$$\begin{aligned}\frac{\Gamma\left(\frac{d}{2}\right)\Gamma\left(n+\frac{3}{2}\right)}{\Gamma\left(n+\frac{d}{2}+\frac{1}{2}\right)} \cdot \frac{\sqrt{\pi}\Gamma(n+1)}{2\Gamma\left(n+\frac{3}{2}\right)} &= \frac{\sqrt{\pi}\Gamma\left(\frac{d}{2}\right)}{\Gamma\left(\frac{d-1}{2}\right)} \cdot \frac{\Gamma\left(\frac{d-1}{2}\right)\Gamma(n+1)}{2\Gamma\left(n+\frac{d}{2}+\frac{1}{2}\right)} \\ &= \frac{\pi}{\zeta} \cdot \frac{1}{2}\mathrm{B}\left(\frac{d-1}{2}, n+1\right).\end{aligned}$$

By the definition of the beta function,

$$\frac{1}{2}\mathrm{B}\left(\frac{d-1}{2}, n+1\right) = \frac{1}{2}\int_0^1 x^{d/2-3/2}(1-x)^n \, dx = \int_0^{\pi/2} \cos^{2n+1}\theta \sin^{d-2}\theta \, d\theta.$$

Interchanging the summation and integration,

$$\begin{aligned}\sum_{n=0}^{\infty} a_n \int_0^{\pi/2} \cos^{2n+1}\theta \sin^{d-2}\theta \, d\theta &= \int_0^{\pi/2} \sum_{n=0}^{\infty} a_n \cos^{2n+1}\theta \sin^{d-2}\theta \, d\theta \\ &= \int_0^{\pi/2} \arcsin(\cos\theta) \sin^{d-2}\theta \, d\theta \\ &= \int_0^{\pi/2} \left(\frac{\pi}{2} - \theta\right) \sin^{d-2}\theta \, d\theta = 2\pi\zeta\gamma.\end{aligned}$$

Further, by symmetry of $\sin\theta$,

$$\int_0^{\pi/2} h(\theta) \, d\theta = \frac{1}{2}.$$

Putting them together, we obtain

$$I_{2(a)} = \frac{1}{8\pi^2} \cdot \frac{1}{2} \cdot \frac{\pi}{\zeta} \cdot 2\pi\zeta\gamma = \frac{1}{8}\gamma.$$

Now we turn to $I_{2(b)}$ and show upper and lower bounds on it.

By the symmetry of $\cos\psi$ and a change of variables $x = \sin\varphi$,

$$\begin{aligned}\int_{\theta-\pi/2}^{0} \cos^{2n+1}\psi \, d\psi = \int_0^{\pi/2-\theta} \cos^{2n+1}\psi \, d\psi &= \int_0^{\pi/2-\theta} \cos^{2n}\psi \, d\sin\psi \\ &= \int_0^{\cos\theta} (1-x^2)^n \, dx.\end{aligned}$$

Since for $x \in [0, \cos\theta]$,

$$\sin^{2n}\theta \le (1-x^2)^n \le 1,$$

we have

$$\sin^{2n}\theta\cos\theta \leq \int_{\theta-\pi/2}^{0}\cos^{2n+1}\psi\,d\psi \leq \cos\theta.$$

Integrating over the density of θ gives

$$\begin{aligned}\int_0^{\pi/2} h(\theta)\sin^{2n}\theta\cos\theta\,d\theta &= \frac{1}{\zeta}\int_0^{\pi/2}\sin^{2n+d-2}\theta\cos\theta\,d\theta\\ &= \frac{1}{\zeta}\int_0^{\pi/2}\sin^{2n+d-2}\theta\,d\sin\theta = \frac{1}{(2n+d-1)\zeta}\end{aligned}$$

and

$$\begin{aligned}\int_0^{\pi/2} h(\theta)\cos\theta\,d\theta = \frac{1}{\zeta}\int_0^{\pi/2}\sin^{d-2}\theta\cos\theta\,d\theta = \frac{1}{\zeta}\int_0^{\pi/2}\sin^{d-2}\theta\,d\sin\theta\\ = \frac{1}{(d-1)\zeta}.\end{aligned}$$

We deal with the upper bound first. Interchanging the summation and integration yields

$$\begin{aligned}\sum_0^\infty a_n\int_0^{\pi/2} g(\varphi)\cos^{2n+1}\varphi\,d\varphi &= \int_0^{\pi/2}\sum_0^\infty a_n g(\varphi)\cos^{2n+1}\varphi\,d\varphi\\ &= \int_0^{\pi/2} g(\varphi)\arcsin(\cos\varphi)\,d\varphi\\ &= (d-2)\int_0^{\pi/2}\left(\frac{\pi}{2}-\varphi\right)\sin^{d-3}\varphi\cos\varphi\,d\varphi.\end{aligned}$$

We can derive an upper bound on the above display by the upper bound in (36). Since $\varphi\in[0,\pi/2]$,

$$\frac{\pi}{2}-\varphi \leq \frac{\pi}{2}\cos\varphi.$$

By a change of variables $x=\sin^2\varphi$ and the definition of the beta function,

$$\begin{aligned}\int_0^{\pi/2}\sin^{d-3}\varphi\cos^2\varphi\,d\varphi &= \frac{1}{2}\int_0^{\pi/2}\sin^{d-4}\varphi\cos\varphi\,d\sin^2\varphi\\ &= \frac{1}{2}\int_0^1(1-x)^{1/2}x^{d/2-2}\,dx = \frac{1}{2}\mathrm{B}\left(\frac{d}{2}-1,\frac{3}{2}\right)\\ &= \frac{\Gamma\left(\frac{d}{2}-1\right)\Gamma\left(\frac{3}{2}\right)}{2\Gamma\left(\frac{d+1}{2}\right)}.\end{aligned}$$

Therefore, $I_{2(b)}$ can be upper bounded by

$$\begin{aligned}I_{2(b)} &\leq \frac{1}{8\pi^2}\cdot\frac{\Gamma\left(\frac{d}{2}\right)}{(d-1)\sqrt{\pi}\,\Gamma\left(\frac{d-1}{2}\right)}\cdot\frac{\pi(d-2)\Gamma\left(\frac{d}{2}-1\right)\Gamma\left(\frac{3}{2}\right)}{4\Gamma\left(\frac{d+1}{2}\right)} = \frac{1}{64\pi}\left(\frac{\Gamma\left(\frac{d}{2}\right)}{\Gamma\left(\frac{d+1}{2}\right)}\right)^2\\ &\leq \frac{1}{16\pi}\cdot\frac{1}{d},\end{aligned}$$

where the last inequality is by (14).

For the lower bound on $I_{2(b)}$, we have

$$\begin{aligned}
&\int_0^{\pi/2} \frac{1}{2n+d-1} g(\varphi)\cos^{2n+1}\varphi\, d\varphi \\
&= \frac{d-2}{2n+d-1}\int_0^{\pi/2}\cos^{2n+1}\varphi \sin^{d-3}\varphi \cos\varphi\, d\varphi \\
&= \frac{d-2}{2n+d-1}\cdot\frac{1}{2}\mathrm{B}\left(\frac{d}{2}-1, n+\frac{3}{2}\right) = \frac{d-2}{2n+d-1}\cdot\frac{\Gamma\left(\frac{d}{2}-1\right)\Gamma\left(n+\frac{3}{2}\right)}{2\Gamma\left(n+\frac{d}{2}+\frac{1}{2}\right)} \\
&\geq \frac{d-2}{2n+d+1}\cdot\frac{\Gamma\left(\frac{d}{2}-1\right)\Gamma\left(n+\frac{3}{2}\right)}{2\Gamma\left(n+\frac{d}{2}+\frac{1}{2}\right)} = \frac{\Gamma\left(\frac{d}{2}\right)\Gamma\left(n+\frac{3}{2}\right)}{2\Gamma\left(n+\frac{d}{2}+\frac{3}{2}\right)} = \frac{1}{2}\mathrm{B}\left(\frac{d}{2}, n+\frac{3}{2}\right).
\end{aligned}$$

Hence, by the definition of beta function and a change of variables,

$$\begin{aligned}
I_{2(b)} &\geq \frac{1}{16\pi^2\zeta}\sum_{n=0}^{\infty} a_n \mathrm{B}\left(\frac{d}{2}, n+\frac{3}{2}\right) \\
&= \frac{1}{8\pi^2\zeta}\sum_{n=0}^{\infty} a_n \int_0^{\pi/2}\cos^{2n+1}\psi \sin^{d-1}\psi\cos\psi\, d\psi \\
&= \frac{1}{8\pi^2\zeta}\int_0^{\pi/2}\sum_{n=0}^{\infty} a_n \cos^{2n+1}\psi \sin^{d-1}\psi\cos\psi\, d\psi \\
&= \frac{1}{8\pi^2\zeta}\int_0^{\pi/2}\left(\frac{\pi}{2}-\psi\right)\sin^{d-1}\psi\cos\psi\, d\psi.
\end{aligned}$$

Further, by the lower bound in (36), we obtain

$$I_{2(b)} \geq \frac{1}{8\pi^2\zeta}\int_0^{\pi/2}\sin^{d-1}\psi\cos^2\psi\, d\psi = \frac{1}{16\pi^2\zeta}\mathrm{B}\left(\frac{d}{2},\frac{3}{2}\right).$$

Inserting the definitions of ζ and the beta function gives

$$\begin{aligned}
I_{2(b)} &\geq \frac{1}{16\pi^2}\cdot\frac{\Gamma\left(\frac{d}{2}\right)}{\sqrt{\pi}\Gamma\left(\frac{d-1}{2}\right)}\cdot\frac{\Gamma\left(\frac{d}{2}\right)\Gamma\left(\frac{3}{2}\right)}{\Gamma\left(\frac{d+3}{2}\right)} = \frac{1}{32\pi^2}\cdot\frac{d-1}{d+1}\left(\frac{\Gamma\left(\frac{d}{2}\right)}{\Gamma\left(\frac{d+1}{2}\right)}\right)^2 \\
&\geq \frac{1}{64\pi^2}\left(\frac{\Gamma\left(\frac{d}{2}\right)}{\Gamma\left(\frac{d+1}{2}\right)}\right)^2.
\end{aligned}$$

Using (14), we have

$$I_{2(b)} \geq \frac{1}{32\pi^2}\cdot\frac{1}{d}.$$

For the integral I_3 in (52), by a change of variables $\xi = \theta - \psi$,

$$\begin{aligned} I_3 &= \int_0^{\pi/2}\int_0^{\pi/2}\int_{\theta-\pi/2}^{\pi/2} \arcsin(\cos\varphi\cos(\theta-\psi))h(\theta,\psi,\varphi)\,d\psi\,d\theta\,d\varphi \\ &= \int_0^{\pi/2}\int_0^{\pi/2}\int_{\pi/2}^{\theta-\pi/2} -\arcsin(\cos\varphi\cos\xi)h(\theta,\psi,\varphi)\,d\xi\,d\theta\,d\varphi \\ &= \int_0^{\pi/2}\int_0^{\pi/2}\int_{\theta-\pi/2}^{\pi/2} \arcsin(\cos\varphi\cos\psi)h(\theta,\psi,\varphi)\,d\psi\,d\theta\,d\varphi = I_2. \end{aligned}$$

Combining the estimates of I_1, I_2, I_3 given above proves the claim. □

Plugging Lemma B.5 into (48), we arrive at the claims in Lemma B.2. Using Lemma 4.1 and Lemma B.2, we have

$$|\mathbb{E}_{\mathcal{G}(n,1/2,d,q)}[\tau_{[4]}]| = q^6|\mathbb{E}_{\mathcal{G}(n,1/2,d,q)}[\tau_{[4]}]| \le \frac{Cq^6}{d}.$$

Then,

$$|\mathbb{E}[\tau_4(\mathcal{G}(n,1/2,d,q))]| \le \binom{n}{4}|\mathbb{E}_{\mathcal{G}(n,1/2,d,q)}[\tau_{[4]}]| \le \frac{Cn^4q^6}{d}.$$

Theorem B.1 is hence proved.

B.2. General signed cliques

We next turn to general signed cliques in $\mathcal{G}(n,p,d,q)$. Similarly, we start with estimations in $\mathcal{G}(n,p,d)$. By the definition of the signed clique,

$$\begin{aligned} |\mathbb{E}_{\mathcal{G}(n,p,d)}[\tau_{[k]}]| &= \left|\mathbb{E}\left[\prod_{\{i,j\}\subset[k]}(a_{i,j}-p)\right]\right| \\ &= \left|\mathbb{E}\left[\prod_{\{i,j\}\subset[k]}(a_{i,j}-p)\right] - \prod_{\{i,j\}\subset[k]}(p-p)\right| \\ &= \left|\sum_{S\in 2^{\binom{[k]}{2}}}(\mathbb{P}(S)-p^{|S|})(-p)^{\binom{k}{2}-|S|}\right| \\ &\le \sum_{S\in 2^{\binom{[k]}{2}}}|\mathbb{P}(S)-p^{|S|}|p^{\binom{k}{2}-|S|}. \end{aligned}$$

The following corollary, derived from a result in Section 3.2 of the main article, facilitates our calculations.

Corollary B.7. *Let $V = [k]$ be a set of vertices and $E \subset V \times V$ be a set of edges. Denote by $|E|$ the cardinality of E. Then, we have that for a constant $C_{k,p}$,*

$$|\mathbb{P}_{\mathcal{G}(k,p,d)}(E) - p^{|E|}| \le \frac{C_{k,p}}{\sqrt{d}}.$$

Proof. In (26), by setting $q = 1$, we have that

$$\mathrm{TV}(\mathcal{G}(n,p),\mathcal{G}(n,p,d)) \le C_p \frac{n}{d} + C\left(\sqrt{\frac{n^2}{d}} + \sqrt{\frac{n^3}{d}}\right).$$

By the definition of the distance,

$$\begin{aligned}|\mathbb{P}_{\mathcal{G}(k,p,d)}(E) - p^{|E|}| &= \left|\mathbb{P}_{\mathcal{G}(k,p,d)}(E) - \mathbb{P}_{\mathcal{G}(k,p)}(E)\right| \le \mathrm{TV}(\mathcal{G}(k,p),\mathcal{G}(k,p,d)) \\ &\le C_p\sqrt{\frac{k^3}{d}}.\end{aligned}$$

The claim directly follows. □

By Corollary B.7,

$$|\mathbb{E}_{\mathcal{G}(n,p,d)}[\tau_{[k]}]| \le \frac{C_{k,p}}{\sqrt{d}}. \tag{56}$$

Therefore, for a constant $C_{k,p}$,

$$\begin{aligned}|\mathbb{E}[\tau_k(\mathcal{G}(n,p,d,q))]| &= q^{\binom{k}{2}}|\mathbb{E}[\tau_k(\mathcal{G}(n,p,d))]| \le q^{\binom{k}{2}}\binom{n}{k}|\mathbb{E}[\tau_{[k]}]| \\ &\le \frac{C_{k,p} n^k q^{k(k-1)/2}}{\sqrt{d}}.\end{aligned} \tag{57}$$

Lemma B.8. *There exists a constant $C_{k,p} > 0$, depending only on p and k, such that*

$$\mathbb{V}\mathrm{ar}[\tau_k(\mathcal{G}(n,p,d,q))] \ge C_{k,p} n^k.$$

Proof. Consider two sets of vertices S and S' of size k. Since $(a_{i,j} - p)^2$ equals $(1-p)^2$ or p^2, we have that $(a_{i,j} - p)^2 \ge p^2(1-p)^2$. Now if $S = S'$, then

$$\mathbb{E}[\tau_S \tau_{S'}] = \mathbb{E}[(\tau_S)^2] = \mathbb{E}\left[\prod_{\{i,j\}\in V_1} (a_{i,j} - p)^2\right] \ge (p^2(1-p)^2)^{\binom{k}{2}} = (p(1-p))^{k(k-1)}.$$

By (56), there exists a $C'_{k,p} > 0$, such that for $d \ge C'_{k,p}$,

$$\mathbb{E}[\tau_S]^2 \le \frac{1}{2}\,\mathbb{E}[(\tau_S)^2].$$

For $S \ne S'$, let $V' = S \cap S'$ be the set of overlapping vertices. Then, we have that

$$\mathbb{E}[\tau_S \tau_{S'}] = \mathbb{E}[\mathbb{E}[\tau_S \tau_{S'} \mid V']] = \mathbb{E}[\mathbb{E}[\tau_S \mid V']\,\mathbb{E}[\tau_{S'} \mid V']] = \mathbb{E}[\mathbb{E}[\tau_S \mid V']^2] \ge \mathbb{E}[\tau_S]^2,$$

where the inequality is by Jensen's.

Therefore, there exists a $C_{k,p} > 0$, such that when $d \ge C_{k,p}$,

$$\begin{aligned}\mathbb{V}\mathrm{ar}[\tau_k(\mathcal{G}(n,p,d,q))] &= \mathbb{E}\left[\left(\sum_{S\in\binom{n}{k}} \tau_S\right)^2\right] - \left(\sum_{S\in\binom{n}{k}} \mathbb{E}[\tau_S]\right)^2 \ge \binom{n}{k}\mathbb{V}\mathrm{ar}[\tau_{[k]}] \\ &\ge C_{k,p} n^k.\end{aligned}$$

□

Putting them together, we have that for some constant $C_{k,p}$,

$$\frac{(\mathbb{E}[\tau_k(\mathcal{G}(n,p,d,q))] - \mathbb{E}[\tau_k(\mathcal{G}(n,p))])^2}{\min\{\mathbb{V}\mathrm{ar}[\tau_k(\mathcal{G}(n,p))], \mathbb{V}\mathrm{ar}[\tau_k(\mathcal{G}(n,p,d,q))]\}} \leq \frac{C_{k,p} n^k q^{k(k-1)}}{d}.$$

The above display implies that the method used to derive the possibility of detection does not work when $n^k q^{k(k-1)}/d \to 0$, which suggests a certain boundary of detection using general signed clique statistics. Note that for $k \geq 4$ this does not rule out the whole region where signed triangles are not able to distinguish. However, based on the computation of the expected signed quadruple count, we see that the upper bound on the expectation in (57) is not precise; in particular, the dependence on d can be improved. In general, we do not expect the detection boundary to be improved by signed cliques. Towards this, we present the following conjecture.

Conjecture B.9. *There exists a constant $C_{k,p}$ such that*

$$|\mathbb{E}_{\mathcal{G}(n,p,d)}[\tau_{[k]}]| \leq \frac{C_{k,p}}{d^{k/6}}.$$

We briefly argue why this bound should hold. In the proof of Theorem 1.1(b), we see that the dominating term comes from the ratio between the variance of the signed triangle count in $\mathcal{G}(n,p)$ and the squared expectation of the signed triangle count in $\mathcal{G}(n,p,d,q)$. Suppose this still holds for general signed cliques. Then, the dominating term if we use a signed clique statistic becomes $C_{k,p} n^k/d^{2\alpha}$, where $C_{k,p}/d^{\alpha}$ is a lower bound for $|\mathbb{E}_{\mathcal{G}(n,p,d)}[\tau_{[k]}]|$. Since n^3/d is the precise order for the phase transition in $\mathcal{G}(n,p,d)$, we must have $k/(2\alpha) \leq 3$, which gives $\alpha \geq k/6$. Note that this argument does not give a tight bound on the power of d, as witnessed by the case of $k = 4$, when we know from Lemma B.2 that $|\mathbb{E}_{\mathcal{G}(n,1/2,d)}[\tau_{[4]}]|$ decays as C/d.

In any case, assuming Conjecture B.9 holds, and by the same arguments presented in this subsection, we obtain that detection is not possible with this method if $n^3 q^{3(k-1)}/d \to 0$, and this bound gets worse as k grows.

C. Detecting geometry using signed cycles

Let S be a subset of V and denote $k := |S|$. Consider a cycle $C \subset \binom{S}{2}$, which is a set of edges forming a closed chain. There are $(k-1)!/2$ possible Hamilton cycles (each vertex is visited exactly once) on S; they are distributed identically to $C^0 := \{\{1,2\},\{2,3\},\ldots,\{k-1,k\},\{k,1\}\}$. Denote by K_C the indicator that the pairs in C form a cycle of G. Given the adjacency matrix $\boldsymbol{A}$ of G, K_C can be expressed by

$$K_C = \prod_{e \in C} a_e.$$

Then, the total number of length k cycles in G, denoted by $K_k(G)$, can be written as

$$K_k(G) = \sum_{C \subset \binom{S}{2}, S \in \binom{V}{k}} K_C.$$

We similarly define the signed cycle and its count by

$$\kappa_C = \prod_{e\in C}(a_e - p) \quad \text{and} \quad \kappa_k(G) = \sum_{C\subset\binom{S}{2}, S\in\binom{V}{k}} \kappa_C.$$

We again start with estimating the expectation and variance of the signed cycle statistic in $\mathcal{G}(n,p)$.

For $\mathcal{G}(n,p)$, again by independence of edges, a signed cycle has expectation zero:

$$\mathbb{E}[\kappa_{C^0}] = \prod_{e\in C^0}\mathbb{E}[a_e - p] = 0.$$

Hence, the expectation of the signed length k cycle statistic in $\mathcal{G}(n,p)$ is also zero:

$$\mathbb{E}[\kappa_k(\mathcal{G}(n,p))] = \binom{n}{k}\frac{(k-1)!}{2}\,\mathbb{E}[\kappa_{C^0}] = 0. \tag{58}$$

Consider two cycles C and C' of length k. If $C = C'$, then

$$\mathbb{E}[\kappa_C\kappa_{C'}] = \mathbb{E}[(\kappa_C)^2] = \mathbb{E}\left[\prod_{e\in C}(a_e - p)^2\right] = \prod_{e\in C}\mathbb{E}[(a_e - p)^2] = (p(1-p))^k.$$

For $C \neq C'$, there exists at least one edge e that is in C but not in C'. Hence,

$$\mathbb{E}[\kappa_C\kappa_{C'}] = \mathbb{E}[a_e - p]\,\mathbb{E}\left[\kappa_{C'}\prod_{e'\in C\setminus\{e\}}(a_{e'} - p)\right] = 0.$$

Therefore, for some $C_{k,p} > 0$,

$$\begin{aligned}\mathbb{V}\mathrm{ar}[\kappa_k(\mathcal{G}(n,p))] &= \mathbb{E}\left[\left(\sum_{C\subset\binom{S}{2}, S\in\binom{V}{k}}\kappa_C\right)^2\right] = \frac{n!}{(n-k)!2k}\,\mathbb{E}[(\kappa_{C^0})^2] \\ &= \frac{n!}{(n-k)!2k}(p(1-p))^k \geq C_{k,p}n^k.\end{aligned} \tag{59}$$

In order to estimate the mean of the signed cycle statistic in $\mathcal{G}(n,p,d,q)$, we additionally need the following lemma concerning the probability of an open path (i.e., an open chain of edges) in $\mathcal{G}(n,p,d)$.

Lemma C.1. *In $\mathcal{G}(n,p,d)$, any open path of length k has probability p^k.*

Proof. We prove the claim by induction on the length. For each edge in $\mathcal{G}(n,p,d)$, by definition we have $\mathbb{P}(i\sim j) = \mathbb{E}[a_{i,j}] = p$. Suppose we have a path $P_{k+1} \coloneqq \{\{1,2\},\{2,3\},\ldots,\{k,k+1\}\}$ on vertices $[k+1]$. Then, the probability of the path is

$$\mathbb{P}(P_{k+1}) = \mathbb{E}\left[\prod_{i=1}^{k}a_{i,i+1}\right].$$

By conditional independence of the edges and rotation symmetry on sphere,

$$\mathbb{E}\left[\prod_{i=1}^{k} a_{i,i+1}\right] = \mathbb{E}\left[\mathbb{E}\left[\prod_{i=1}^{k} a_{i,i+1} \;\middle|\; \boldsymbol{x}_k\right]\right] = \mathbb{E}\left[\mathbb{E}\left[\prod_{i=1}^{k-1} a_{i,i+1} \;\middle|\; \boldsymbol{x}_k\right]\mathbb{E}[a_{k,k+1} \mid \boldsymbol{x}_k]\right]$$
$$= p\,\mathbb{E}\left[\prod_{i=1}^{k-1} a_{i,i+1}\right] = p\,\mathbb{P}(P_k).$$

□

Expanding the product of a signed cycle,

$$\mathbb{E}_{\mathcal{G}(n,p,d)}\left[\kappa_C\right] = \sum_{S\in 2^C} (-p)^{k-|S|}\left(\mathbb{E}\left[\prod_{e\in S} a_e\right] - p^{|S|}\right).$$

Since C is a cycle, all proper subsets of C are a union of independent paths. By Lemma C.1,

$$\mathbb{E}\left[\prod_{e\in S} a_e\right] = \mathbb{P}(S) = p^{|S|}$$

for all $S \subset C$ except for $S = C$. Hence, the expectation of a signed length k cycle in $\mathcal{G}(n,p,d)$ satisfies

$$\mathbb{E}_{\mathcal{G}(n,p,d)}[\kappa_C] = \mathbb{E}\left[\prod_{e\in C} a_e\right] - p^k. \tag{60}$$

C.1. Signed quadrilaterals

We start with the expected number of signed cycles of length four, which are called *signed quadrilaterals*. By (60), the expectation of a signed quadrilateral C^0 in $\mathcal{G}(n,1/2,d)$ is

$$\mathbb{E}[\kappa_{C^0}] = \mathbb{E}[\overline{a}_{1,2}\overline{a}_{2,3}\overline{a}_{3,4}\overline{a}_{4,1}] = \mathbb{E}[a_{1,2}a_{2,3}a_{3,4}a_{4,1}] - \left(\frac{1}{2}\right)^4.$$

Using Lemma B.3, we have that

$$\mathbb{E}[\kappa_{C^0}] = 2\eta.$$

Hence, by (41),

$$\frac{1}{2\pi^2}\cdot\frac{1}{d} \le \mathbb{E}_{\mathcal{G}(n,1/2,d)}[\kappa_{C^0}] \le \frac{1}{8}\cdot\frac{1}{d}.$$

Therefore, for absolute constants $C, C' > 0$, we have that

$$\frac{Cn^4q^4}{d} \le \mathbb{E}[\kappa_4(\mathcal{G}(n,1/2,d,q))] \le \frac{C'n^4q^4}{d}.$$

Together with (58), we have that

$$|\mathbb{E}[\kappa_4(\mathcal{G}(n,1/2,d,q))] - \mathbb{E}[\kappa_4(\mathcal{G}(n,1/2))]| \le \frac{Cn^4q^4}{d}.$$

By (59), we also have that

$$\max\{\mathbb{V}\text{ar}[\kappa_4(\mathcal{G}(n,1/2))], \mathbb{V}\text{ar}[\kappa_4(\mathcal{G}(n,1/2,d,q))]\} \geq \mathbb{V}\text{ar}[\kappa_4(\mathcal{G}(n,1/2))] \geq Cn^4$$

for an absolute constant $C > 0$.

Therefore, there is an absolute constant C such that

$$\frac{(\mathbb{E}[\kappa_4(\mathcal{G}(n,1/2,d,q))] - \mathbb{E}[\kappa_4(\mathcal{G}(n,1/2))])^2}{\max\{\mathbb{V}\text{ar}[\kappa_4(\mathcal{G}(n,1/2))], \mathbb{V}\text{ar}[\kappa_4(\mathcal{G}(n,1/2,d,q))]\}} \leq \frac{Cn^4q^8}{d^2}.$$

This implies that if detection is possible using this method, then we should have $n^2q^4/d \to \infty$. This is worse than the condition $n^3q^6/d \to \infty$ under which signed triangles can detect.

C.2. General signed cycle

Next we estimate the signed length k cycle count in $\mathcal{G}(n,p,d,q)$ with the help of Lemma C.1.

By Corollary B.7, the probability of a cycle satisfies

$$|\mathbb{P}(C^0) - p^k| \leq \frac{C_{k,p}}{\sqrt{d}}.$$

Hence, by (60),

$$|\mathbb{E}_{\mathcal{G}(n,p,d)}[\kappa_{C^0}]| \leq \frac{C_{k,p}}{\sqrt{d}}. \tag{61}$$

Thus,

$$\begin{aligned}|\mathbb{E}[\kappa_k(\mathcal{G}(n,p,d,q))]| = |q^k\,\mathbb{E}[\kappa(\mathcal{G}(n,p,d))]| &\leq \frac{q^k n!}{(n-k)!2k}|\mathbb{E}_{\mathcal{G}(n,p,d)}[\kappa_{C^0}]| \\ &\leq \frac{C_{k,p}n^kq^k}{\sqrt{d}}\end{aligned}$$

for a constant $C_{k,p}$.

We cannot find an easy derivation for a lower bound of the variance in $\mathcal{G}(n,p,d,q)$. However, $\mathbb{V}\text{ar}[\kappa_k(\mathcal{G}(n,p))]$ is already of order $\Omega(n^k)$ which we believe is also the correct order of $\mathbb{V}\text{ar}[\kappa_k(\mathcal{G}(n,p,d,q))]$. Using (59),

$$\max\{\mathbb{V}\text{ar}[\kappa_k(\mathcal{G}(n,p))], \mathbb{V}\text{ar}[\kappa_k(\mathcal{G}(n,p,d,q))]\} \geq \mathbb{V}\text{ar}[\kappa_k(\mathcal{G}(n,p))] \geq C_{k,p}n^k.$$

Therefore, for some $C_{k,p} > 0$,

$$\frac{(\mathbb{E}[\kappa_k(\mathcal{G}(n,p,d,q))] - \mathbb{E}[\kappa_k(\mathcal{G}(n,p))])^2}{\max\{\mathbb{V}\text{ar}(\kappa_k(\mathcal{G}(n,p))), \mathbb{V}\text{ar}(\kappa_k(\mathcal{G}(n,p,d,q)))\}} \leq \frac{C_{k,p}n^kq^{2k}}{d}.$$

The above display implies that detection is not possible using the previous method with a signed length k cycle statistic when $n^kq^{2k}/d \to 0$. Note that for $k \geq 4$ this does not rule out all regions not detectable by signed triangles. However, based on the computations for signed quadrilaterals, we believe that the dependence on d in the bound in (61) is not tight. Analogously to Conjecture B.9, we have the following conjecture.

Conjecture C.2. *There exists a constant $C_{k,p}$ such that*

$$|\mathbb{E}_{\mathcal{G}(n,p,d)}[\kappa_{C^0}]| \leq \frac{C_{k,p}}{d^{k/6}}.$$

Assuming Conjecture C.2 holds, detection is not possible with this method when $n^3q^6/d \to 0$. This would imply that all signed cycles have the same detecting power. However, as witnessed by signed quadrilaterals, the above conjecture is not tight, suggesting that signed triangles yield the best bound.

Acknowledgments

S. L. thanks Cong Ma, Kaizheng Wang, and Jiacheng Zhang for numerous insightful discussions throughout the years. We thank the editors and anonymous referees for their valuable comments and help to improve the manuscript. The authors acknowledge the generous support from NSF grant DMS-1811724.

Funding

Research supported in part by NSF grant DMS-1811724.

References

[1] Alzer, H. (1997). On some inequalities for the gamma and psi functions. *Mathematics of Computation* **66** 373–389. MR1388887

[2] Araya, E. and De Castro, Y. (2019). Latent distance estimation for random geometric graphs. In *Advances in Neural Information Processing Systems* 8724–8734.

[3] Araya Valdivia, E. (2018). Relative concentration bounds for the spectrum of kernel matrices. Preprint available at arXiv:1812.02108.

[4] Athreya, A., Fishkind, D. E., Tang, M., Priebe, C. E., Park, Y., Vogelstein, J. T., Levin, K., Lyzinski, V., Qin, Y. and Sussman, D. L. (2018). Statistical inference on random dot product graphs: A survey. *Journal of Machine Learning Research* **18** 1–92. MR3827114

[5] Bangachev, K. and Bresler, G. (2023). Random algebraic graphs and their convergence to Erdős-Rényi. Preprint available at arXiv:2305.04802.

[6] Bishop, C. M. (2006). *Pattern Recognition and Machine Learning*. Springer. MR2247587

[7] Bollobás, B. and Erdős, P. (1976). Cliques in random graphs. *Mathematical Proceedings of the Cambridge Philosophical Society* **80** 419–427. MR0498256

[8] Brennan, M., Bresler, G. and Nagaraj, D. (2020). Phase transitions for detecting latent geometry in random graphs. *Probability Theory and Related Fields* **178** 1215–1289. MR4168398

[9] Brennan, M., Bresler, G. and Huang, B. (2021). De Finetti-style results for Wishart matrices: Combinatorial structure and phase transitions. Preprint available at arXiv:2103.14011.

[10] Brennan, M., Bresler, G. and Huang, B. (2022). Threshold for detecting high dimensional geometry in anisotropic random geometric graphs. Preprint available at arXiv:2206.14896.

[11] Bubeck, S. and Ganguly, S. (2016). Entropic CLT and phase transition in high-dimensional Wishart matrices. *International Mathematics Research Notices* **2018** 588–606. MR3801440

[12] Bubeck, S., Ding, J., Eldan, R. and Rácz, M. Z. (2016). Testing for high-dimensional geometry in random graphs. *Random Structures & Algorithms* **49** 503–532. MR3545825

[13] Cai, T. T., Liang, T. and Zhou, H. H. (2015). Law of log determinant of sample covariance matrix and optimal estimation of differential entropy for high-dimensional Gaussian distributions. *Journal of Multivariate Analysis* **137** 161–172. MR3332804

[14] Chételat, D. and Wells, M. T. (2019). The middle-scale asymptotics of Wishart matrices. *Annals of Statistics* **47** 2639–2670. MR3988768

[15] Dettmann, C. P. and Georgiou, O. (2016). Random geometric graphs with general connection functions. *Physical Review E* **93** 032313. MR3657789

[16] Devroye, L., György, A., Lugosi, G. and Udina, F. (2011). High-dimensional random geometric graphs and their clique number. *Electronic Journal of Probability* **16** 2481–2508. MR2861682

[17] Eldan, R. and Mikulincer, D. (2020). Information and dimensionality of anisotropic random geometric graphs. In *Geometric Aspects of Functional Analysis* 273–324. Springer. MR4175752

[18] Eldan, R., Mikulincer, D. and Pieters, H. (2022). Community detection and percolation of information in a geometric setting. *Combinatorics, Probability and Computing* 1–22. https://doi.org/10.1017/S0963548322000098. MR4496021

[19] Erdős, P. and Rényi, A. (1959). On random graphs I. *Publicationes Mathematicae Debrecen* **6** 290–297. MR0120167

[20] Gupta, P. and Kumar, P. R. (2000). The capacity of wireless networks. *IEEE Transactions on Information Theory* **46** 388–404. MR1748976

[21] Hammersley, J. M. (1950). The distribution of distance in a hypersphere. *The Annals of Mathematical Statistics* 447–452. MR0037481

[22] Hoff, P. D., Raftery, A. E. and Handcock, M. S. (2002). Latent space approaches to social network analysis. *Journal of the American Statistical Association* **97** 1090–1098. MR1951262

[23] Jiang, T. and Li, D. (2015). Approximation of rectangular beta-laguerre ensembles and large deviations. *Journal of Theoretical Probability* **28** 804–847. MR3413957

[24] Levin, D. A. and Peres, Y. (2017). *Markov Chains and Mixing Times*. American Mathematical Society. MR3726904

[25] Liu, S. and Rácz, M. Z. (2023). A probabilistic view of latent space graphs

and phase transitions. *Bernoulli* **29** 2417–2441. MR4580922

[26] Liu, S., Mohanty, S., Schramm, T. and Yang, E. (2022). Testing thresholds for high-dimensional sparse random geometric graphs. In *Proceedings of the 54th Annual ACM SIGACT Symposium on Theory of Computing* 672–677. MR4490031

[27] Marsaglia, G. (1972). Choosing a point from the surface of a sphere. *Ann. Math. Statist.* **43** 645–646.

[28] Mikulincer, D. (2022). A CLT in Stein's distance for generalized Wishart matrices and higher-order tensors. *International Mathematics Research Notices* **2022** 7839–7872. MR4418721

[29] Muller, M. E. (1956). Some continuous Monte Carlo methods for the Dirichlet problem. *The Annals of Mathematical Statistics* 569–589. MR0088786

[30] Muller, M. E. (1959). A note on a method for generating points uniformly on n-dimensional spheres. *Communications of the ACM* **2** 19–20.

[31] Ostilli, M. and Bianconi, G. (2015). Statistical mechanics of random geometric graphs: Geometry-induced first-order phase transition. *Physical Review E* **91** 042136. MR3470991

[32] Parthasarathy, S., Sivakoff, D., Tian, M. and Wang, Y. (2017). A quest to unravel the metric structure behind perturbed networks. In *33rd International Symposium on Computational Geometry (SoCG 2017)* **77** 53. MR3685725

[33] Penrose, M. (2003). *Random Geometric Graphs* **5**. Oxford University Presss. MR1986198

[34] Penrose, M. D. (2016). Connectivity of soft random geometric graphs. *The Annals of Applied Probability* **26** 986–1028. MR3476631

[35] Polyanskiy, Y. and Wu, Y. (2012-2017). Lecture notes on Information Theory. *MIT (6.441), UIUC (ECE 563), Yale (STAT 664).*

[36] Rácz, M. Z. and Bubeck, S. (2017). Basic models and questions in statistical network analysis. *Statistics Surveys* **11** 1–47. MR3696007

[37] Rácz, M. Z. and Richey, J. (2019). A smooth transition from Wishart to GOE. *Journal of Theoretical Probability* **32** 898–906. MR3959632

[38] Scheinerman, E. R. and Tucker, K. (2010). Modeling graphs using dot product representations. *Computational Statistics* **25** 1–16. MR2586721

[39] Smith, A. L., Asta, D. M. and Calder, C. A. (2019). The geometry of continuous latent space models for network data. *Statistical Science* **34** 428–453. MR4017522

[40] Sussman, D. L., Tang, M. and Priebe, C. E. (2013). Consistent latent position estimation and vertex classification for random dot product graphs. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **36** 48–57.

[41] Tee, P. and Balmain, A. (2018). Phase transitions in spatial networks as a model of cellular symbiosis. In *International Conference on Complex Networks and their Applications* 586–597. Springer.

[42] Todhunter, I. (1863). *Spherical Trigonometry, for the Use of Colleges and Schools: With Numerous Examples.* Macmillan.

[43] Wainwright, M. J. (2019). *High-dimensional Statistics: A Non-asymptotic Viewpoint* **48**. Cambridge University Press. MR3967104

[44] Wendel, J. (1948). Note on the gamma function. *The American Mathematical Monthly* **55** 563–564. MR0029448

[45] Wilsher, M., Dettmann, C. P. and Ganesh, A. (2020). Connectivity in one-dimensional soft random geometric graphs. *Physical Review E* **102** 062312. MR4199518