Electronic Journal of Probability

Vol. 16 (2011), Paper no. 90, pages 2481–2508.

Journal URL
http://www.math.washington.edu/~ejpecp/

High-dimensional random geometric graphs and their clique number

Luc Devroye* András György† Gábor Lugosi‡ Frederic Udina‡

Abstract

We study the behavior of random geometric graphs in high dimensions. We show that as the dimension grows, the graph becomes similar to an Erdős-Rényi random graph. We pay particular attention to the clique number of such graphs and show that it is very close to that of the corresponding Erdős-Rényi graph when the dimension is larger than $\log^3 n$ where n is the number of vertices. The problem is motivated by a statistical problem of testing dependencies..

Key words: Clique number; dependency testing; geometric graphs; random graphs.

AMS 2010 Subject Classification: Primary 05C80; 62H15.

Submitted to EJP on October 1, 2010, final version accepted November 29, 2011.

*School of Computer Science, McGill University, 3480 University Street Montreal, Canada H3A 2A7 (email: lucdevroye@gmail.com). Supported by NSERC.

†Machine Learning Research Group, Computer and Automation Research Institute of the Hungarian Academy of Sciences, Kende u. 13-17, 1111 Budapest, Hungary (email: gya@szit.bme.hu). Partially supported by the National Development Agency of Hungary from the Research and Technological Innovation Fund (KTIA-OTKA CNK 77782), and by the PASCAL2 Network of Excellence (EC grant no. 216886).

‡Department of Economics, Pompeu Fabra University, Ramon Trias Fargas 25-27, 08005, Barcelona, Spain (email: gabor.lugosi@gmail.com and frederic.udina@gmail.com). G. Lugosi is also with ICREA. Supported by the Spanish Ministry of Science and Technology grant MTM2009-09063 and by the PASCAL Network of Excellence under EC grant no. 216886.

1 Introduction

A *random geometric graph* is defined by n independent random points taking values in $\mathbb{R}^d$, drawn from the same distribution. These points correspond to the vertices of the graph and two of them are joined by an edge if and only if their Euclidean distance is less than a certain threshold. Such random geometric graphs have been studied extensively and many of their basic properties are now well understood. We refer to Penrose [16] for an extensive treatment. These graphs are usually studied in an asymptotic framework when the number n of vertices is very large (it grows to infinity) while the dimension d is held fixed. However, in some applications it is of interest to consider situations when the dimension is large. In such cases the graph is expected to behave differently. In this paper we consider random geometric graphs defined by n independent vectors uniformly distributed on the surface of the unit ball in $\mathbb{R}^d$. We show that if $d \to \infty$ while n is held fixed, the random graph becomes, in a very strong sense, similar to an Erdős-Rényi random graph. Motivated by a hypothesis testing problem, we pay particular attention to the clique number of such random geometric graphs. We show that if d is at least of the order of $\log^3 n$, then the clique number is essentially the same as that of the corresponding Erdős-Rényi random graph. This is in sharp contrast to the behavior of the clique number when the dimension is fixed.

The paper is organized as follows. In Section 2 the basic model is described and the asymptotic equivalence of the random geometric graph and the Erdős-Rényi random graph is presented (Theorem 2). In Section 3 the main results of the paper are stated and proved on the behavior of the clique number of high-dimensional random geometric graphs. In Section 4 some numerical experiments are reported in which the behavior of the clique number is illustrated. In Section 5 we show a statistical application that motivated our research. We describe a hypothesis testing problem arising in applications of remote sensing and finance and propose a test based on computing the clique number of random geometric graphs. Finally, the Appendix contains some of the proofs of results announced in Sections 3 and 5.

2 Notation, set-up

Denote the unit sphere in $\mathbb{R}^d$ by $S_{d-1} = \{\boldsymbol{x} \in \mathbb{R}^d : \|\boldsymbol{x}\| = 1\}$ where $\|\cdot\|$ stands for the Euclidean norm. Let $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ be independent vector-valued random variables, uniformly distributed in S_{d-1}. We denote the components of the random vector $\boldsymbol{X}_i$ by $\boldsymbol{X}_i = (X_{i,1}, \ldots, X_{i,d})$. For a given value of $p \in (0,1)$ (possibly depending on n and d) we define the *random geometric graph* $\overline{G}(n,d,p)$ as follows: the graph has n vertexes labeled by $1, \ldots, n$ and vertex i and vertex j are connected by an edge if an only if

$$\left(\boldsymbol{X}_i, \boldsymbol{X}_j\right) \geq t_{p,d} \ ,$$

where $(\boldsymbol{x}, \boldsymbol{y})$ denotes the inner product of the vectors $\boldsymbol{x}$ and $\boldsymbol{y}$ and $t_{p,d}$ is determined such that the probability of each edge equals p, that is,

$$\mathbb{P}\left\{\left(\boldsymbol{X}_i, \boldsymbol{X}_j\right) \geq t_{p,d}\right\} = p \ .$$

Equivalently, vertex i and vertex j are connected if and only if $\|\boldsymbol{X}_i - \boldsymbol{X}_j\| \leq \sqrt{2(1 - t_{p,d})}$.

For example, for $p = 1/2$, $t_{p,d} = 0$. To understand the behavior of $t_{p,d}$ as a function of p, we introduce some notation. Let μ_{d-1} denote the uniform probability measure over S_{d-1}. For a unit

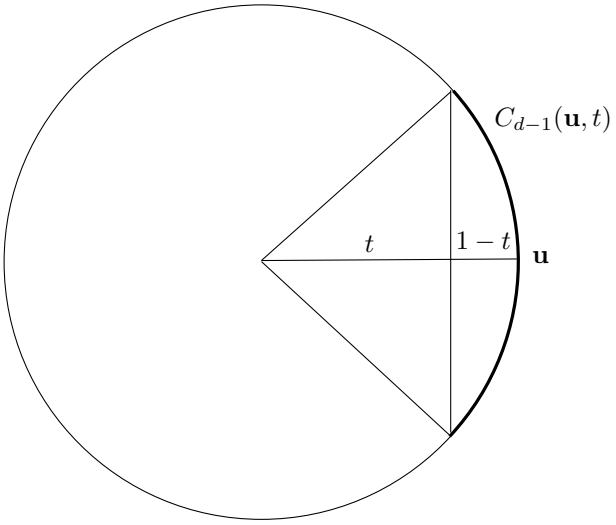

Figure 1: A spherical cap of height $1-t$.

vector $\boldsymbol{u} \in S_{d-1}$ and real number $0 \le t \le 1$, let $C_{d-1}(\boldsymbol{u},t) = \{\boldsymbol{x} \in \mathbb{R}^d : \boldsymbol{x} \in S_{d-1}, (\boldsymbol{x},\boldsymbol{u}) \ge t\}$ denote a spherical cap of height $1-t$ around $\boldsymbol{u}$ (see Figure 1). The *angle* of a spherical cap $C_{d-1}(\boldsymbol{u},t)$ is defined by $\arccos(t)$.

Then $p = \mu_{d-1}(C_{d-1}(\boldsymbol{e}, t_{p,d}))$ is the normalized surface area of a spherical cap of height $1-t_{p,d}$ centered at (say) the first standard basis vector $\boldsymbol{e} = (1,0,0,\ldots,0)$. The following estimates for the measure of a spherical cap will be used (see Brieden et al. [6]): for $\sqrt{2/d} \le t_{p,d} \le 1$,

$$\frac{1}{6t_{p,d}\sqrt{d}}(1-t_{p,d}^2)^{\frac{d-1}{2}} \le p \le \frac{1}{2t_{p,d}\sqrt{d}}(1-t_{p,d}^2)^{\frac{d-1}{2}} \ . \tag{1}$$

These bounds show that if p is fixed and d is large, $t_{p,d}$ is of the order of $1/\sqrt{d}$.

Sometimes it is useful to think about random points on S_{d-1} as projections of Gaussian vectors on the unit sphere. In particular, let $\boldsymbol{Z}_1,\ldots,\boldsymbol{Z}_n$ be independent standard normal vectors (i.e., $\boldsymbol{Z}_i$ has mean $\boldsymbol{0} = (0,\ldots,0)$ and unit covariance matrix). Then the vectors

$$\boldsymbol{X}_1 = \frac{\boldsymbol{Z}_1}{\|\boldsymbol{Z}_1\|}, \ldots, \boldsymbol{X}_n = \frac{\boldsymbol{Z}_n}{\|\boldsymbol{Z}_n\|}$$

are independent and uniformly distributed on S_{d-1}. This representation will be used in some proofs. For example, this representation may be used to determine the asymptotic value of $t_{p,d}$. Let $\boldsymbol{Z} = (Z_1,\ldots,Z_d)$ be a standard Gaussian vector and let $\boldsymbol{X} = \boldsymbol{Z}/\|\boldsymbol{Z}\| = (X_1,\ldots,X_d)$. Observe that $\mathbb{E}\|\boldsymbol{Z}\|^2 = d$. Also, by the law of large numbers, $\|\boldsymbol{Z}\|/\sqrt{d} \to 1$ in probability. This implies that $X_1\sqrt{d}$ converges, in distribution, to a standard normal random variable. In fact, for any fixed k, the joint distribution of $\sqrt{d}(X_1,\ldots,X_k)$ is asymptotically standard normal. One consequence of this is that for any $s > 0$,

$$\mu_{d-1}(C_{d-1}(\boldsymbol{e}, s/\sqrt{d})) = \mathbb{P}\{X_1 > s/\sqrt{d}\} = \mathbb{P}\{Z_1/\|\boldsymbol{Z}\| > s/\sqrt{d}\} \to 1-\Phi(s)$$

as $d \to \infty$ where $\Phi(x) = (2\pi)^{-1/2} \int_{-\infty}^{x} e^{-t^2/2} dt$ is the standard normal distribution function. This implies that $t_{p,d}$ satisfies, for any fixed $p \in (0,1)$,

$$\lim_{d\to\infty} t_{p,d}\sqrt{d} = \Phi^{-1}(1-p)\ . \tag{2}$$

Later we will need a quantitative estimate of the rate of convergence. Such a bound is given by the next lemma, proved in Appendix C.

Lemma 1. *Assume* $0 < p \le 1/2$ *and* $d \ge \max\left\{(2/p)^2, 27\right\}$. *Then*

$$|t_{p,d}\sqrt{d} - \Phi^{-1}(1-p)| \le U_{p,d}\ ,$$

where

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

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

One of the main messages of this paper is that the random geometric graph $\overline{G}(n,d,p)$ defined above behaves like an Erdős-Rényi random graph when d is large. An Erdős-Rényi random graph $G(n,p)$ is defined as a graph on n vertices such that any pair of vertices is connected by an edge with probability p and all edges are present independently. The $G(n,p)$ random graph model was introduced by Erdős and Rényi [10] and most of its properties are well understood – see Bollobás [5], Palmer [15], Janson, Łuczak, and Ruciński [11] for monographs dedicated to the subject.

First we point out that asymptotically (i.e., as $d \to \infty$), the random geometric graph $\overline{G}(n,d,p)$ converges to $G(n,p)$ in total variation distance. However, our proof only implies a small total variation distance for astronomically large values of d. Certain characteristics of $\overline{G}(n,d,p)$ resemble those of $G(n,p)$ for moderate values of d. In Section 3 we show that when $(\log^3 n)/d = o(1)$, the clique numbers of the two random graphs behave quite similarly.

The next theorem states that the distribution of the random geometric graph $\overline{G}(n,d,p)$ converges to that of the Erdős-Rényi random graph $G(n,p)$ in total variation distance. The total variation distance between two random graphs G and G' defined on the same set of vertices (say $\{1,\dots,n\}$) is defined by

$$d_{TV}(G,G') = \max_{\mathscr{G}} |\mathbb{P}\{G \in \mathscr{G}\} - \mathbb{P}\{G' \in \mathscr{G}\}| = \frac{1}{2}\sum_{g} |\mathbb{P}\{G = g\} - \mathbb{P}\{G' = g\}|\ ,$$

where the maximum is taken over all $2^{2^{\binom{n}{2}}}$ sets $\mathscr{G}$ of graphs over n vertices and the sum is taken over all such graphs.

Theorem 2. *Fix a positive integer* n *and* $0 \le p \le 1$. *Then*

$$\lim_{d\to\infty} d_{TV}(\overline{G}(n,d,p), G(n,p)) = 0\ .$$

The proof, given in Appendix A, is based on a relatively straightforward application of the multivariate central limit theorem.

Theorem 2 shows that, asymptotically, the random geometric graph behaves like an ordinary random graph. However, by the bounds provided by the proof, astronomical values of d are required to make

the total variation distance small. (Just note that the total variation distance is the sum over all $2^{\binom{n}{2}}$ possible graphs g and therefore d needs to be much bigger than 2^{n^2} in order to make the obtained bound meaningful.) For this reason, the interest in Theorem 2 is purely theoretical.

On the other hand, the notion by which Theorem 2 relates the random geometric graph to an ordinary random graph is very strong. If one is interested in simple characteristics of $\overline{G}(n,d,p)$, it may behave as that of $G(n,p)$ for much smaller values of d. The main result of the paper, presented in the next section, shows that if d is poly-logarithmic in n, then the clique number of $\overline{G}(n,d,p)$ already behaves very similarly to that of $G(n,p)$. At the same time, for values of d significantly smaller than $\log n$, the clique number behaves very differently.

In this paper we study the (random) clique number $\omega(n,d,p)$ of $\overline{G}(n,d,p)$, that is, the number of vertices in the largest clique contained in $\overline{G}(n,d,p)$. It is well-known (see, e.g., Bollobás [5]) that the clique number of the Erdős-Rényi random graph $G(n,p)$ is, with probability converging to one, within a constant of $2\log_{1/p} n - 2\log_{1/p}\log_{1/p} n$ when p is held fixed as n grows. This is in sharp contrast with the behavior of $\omega(n,d,p)$ for small values of d. It is easy to see that for any fixed d, the clique number grows *linearly* with n and even for $d = \epsilon \log n$, for sufficiently small values of $\epsilon > 0$, $\omega(n,d,p)$ grows as n^α where $\alpha \to 1$ as $\epsilon \to 0$ (see Proposition 4 below).

Theorem 2 implies that, for very large values of d, $\omega(n,d,p)$ behaves similarly to the clique number of $G(n,p)$. The more interesting question is how large d needs to be. The main result of the paper (Theorem 3) establishes that when d is about $\log^3 n$, the behavior of the clique number is already similar to that of $G(n,p)$ (for fixed p). This result is complemented by Theorem 5 which implies that for $d \sim (3\log n)^2$, we have $\omega(n,d,p) = O_p(\log^3 n)$.

3 The clique number of $\overline{G}(n,d,p)$

The following result describes the behavior of the clique number of the random geometric graph $\overline{G}(n,d,p)$ for large values of d.

Theorem 3. *Fix $p \le 1/2$ and define the positive constant*

$$p' = p'(p) = \begin{cases} 1/2 & \text{if } p = 1/2 \\ 1 - \Phi(2\Phi^{-1}(1-p) + 2.5) & \text{if } p < 1/2. \end{cases}$$

Let $\delta_n \in (0,p)$ and suppose

$$d = d_n \ge \frac{\widehat{\kappa}_p}{\delta_n^2} \log^3_{1/(p-\delta_n)} n$$

where $\widehat{\kappa}_p = 65\ln^2(1/p')$. If either $\delta_n \to 0$ or $\delta_n \equiv \delta$ for some constant $0 < \delta < p$, then, with probability converging to 1 (as $n \to \infty$),

$$\omega(n,d,p) \le 2\log_{1/(p+\delta_n)} n - 2\log_{1/(p+\delta_n)}\log_{1/(p+\delta_n)} n + O(1)\,.$$

Also, if $\limsup_{n\to\infty} \delta_n \log^2 n < \infty$, then with probability converging to 1,

$$\omega(n,d,p) \ge 2\log_{1/(p-\delta_n)} n - 2\log_{1/(p-\delta_n)}\log_{1/(p-\delta_n)} n + \Omega(1)\,.$$

Observe that the theorem implies that if d is about $\log^3 n$ then $\omega(n,d,p)$ is already of the order of $\log n$. This is obtained by choosing δ_n as a constant. By letting δ_n go to zero slowly, we see that if $(\log^3 n)/d = o(1)$ then $\omega(n,d,p) \le (2+o(1))\log_{1/p} n$. Finally, by taking $\delta_n \sim 1/\log n$, we obtain that when $d \sim \log^5 n$ then $\omega(n,d,p) \le 2\log_{1/p} n - 2\log_{1/p}\log_{1/p} n + O(1)$ and therefore the clique number is at most as large as in an Erdős-Rényi random graph, up to an additive constant. For the lower bound, we need the extra condition that $\delta_n = O(1/\log^2 n)$ and therefore the lower bound is only meaningful for d at least of the order of $\log^7 n$. We believe that this condition is not necessary. In fact, we conjecture that for fixed d and p, the clique number is non-increasing in d (in the stochastic sense that $\mathbb{P}\{\omega(n,d,p) \ge k\}$ is non-increasing for each k). If this conjecture was true that the lower bound would hold without any condition for d simply because, as $d \to \infty$, by Theorem 2, $\omega(n,d,p)$ converges, in distribution, to the clique number of the Erdős-Rényi random graph.

The theorem follows from Theorems 8 and 9 (together with the observation that $p' \le \min(\widehat{p},\widetilde{p})$) which are shown in Section 3.2. Before turning to the proof we note that for small values of d, $\omega(n,d,p)$ behaves in an entirely different way, as the next simple proposition shows.

Proposition 4. *If $d \ge 8$,*

$$\mathbb{E}\omega(n,d,p) \ge \frac{n}{3\sqrt{d}(1+t_{p,d}^2)}\left(1-\frac{\left(1+t_{p,d}^2\right)^2}{4}\right)^{\frac{d-1}{2}}.$$

The proposition follows simply by observing that if k points fall in any spherical cap C of angle $\arccos(t_{p,d})/2$ that is, a spherical cap of height $1-(1+t_{p,d}^2)/2$, then they are mutually connected and therefore form a clique. The expected number of points that fall in any such fixed cap C is $n\mu_{d-1}(C)$ which, by (1) is at least

$$n\frac{1}{3\sqrt{d}(1+t_{p,d}^2)}\left(1-\frac{\left(1+t_{p,d}^2\right)^2}{4}\right)^{\frac{d-1}{2}}$$

provided $\sqrt{2/d} \le (1+t_{p,d}^2)/2$. This lower bound may be improved by packing as many non-overlapping spherical caps of height $1-(1+t_{p,d}^2)/2$ in S_{d-1} as possible and considering the one containing the largest number of points. Even though the number of such caps is exponentially large in d, the refined bound is not significantly better than the one obtained above. The negligible benefit does not justify the inclusion of the more technical details.

On the one hand, Proposition 4 above shows that if $d \ll \log n$, the clique number grows linearly, or almost linearly, with n while according to Theorem 3, if d is at least of the order of $\log^3 n$, the clique number is logarithmic in n. The next result, using very different techniques, shows that when $d \sim \log^2 n$, then the clique number is already poly-logarithmic in n, at least for $p < 1/2$. The proof is given in Appendix B.

Theorem 5. *For any $p < 1/2$ and $0 < \eta < 1$, the clique number $\omega(n,d,p)$ of the random geometric graph $\overline{G}(n,d,p)$ satisfies, with probability at least $1-\eta$,*

$$\omega(n,d,p) \le n\sqrt{\frac{d+1}{d(dt_{p,d}+1)}}\exp\left(\frac{-(d-1)(dt_{p,d}+1)}{2(d+1)}\right) + 4(d+1)\ln\frac{2ne}{d+1} + 4\ln\frac{4}{\eta}.$$

To interpret this bound, recall that for $p < 1/2$ fixed and d large, $t_{p,d} \approx d^{-1/2}\Phi^{-1}(1-p)$. Thus, the upper bound is of the order $nd^{-1/2}\exp(-d^{1/2}/2) + d\log(n/d)$. Thus, when $d \sim (3\log n)^2$, we have $\omega(n,d,p) = O_p(\log^3 n)$. Notice also that as soon as $d \to \infty$, $\omega(n,d,p) = o_p(n)$.

3.1 The expected number of cliques

The proof of Theorem 3 is based on the first and second moment methods (see, e.g., Alon and Spencer [2]). To this end, first we need to study the expected number of cliques of size k in the random geometric graph $\overline{G}(n,d,p)$. In particular, we compare it to the expected number of cliques of the same size in $G(n,p)$ which is

$$\binom{n}{k} p^{\binom{k}{2}} .$$

Denote the (random) number of cliques of size k by $N_k = N_k(n,d,p)$. But

$$\mathbb{E}N_k = \binom{n}{k} \mathbb{P}\{\boldsymbol{X}_1, \ldots, \boldsymbol{X}_k \text{ form a clique}\}$$

and therefore it suffices to study the probability that k points are all connected with each other. Let $p_k = p_k(d,p) = \mathbb{P}\{\boldsymbol{X}_1, \ldots, \boldsymbol{X}_k \text{ form a clique}\}$ denote this probability.

The cases when $p = 1/2$ and $p < 1/2$ are slightly different and we treat them separately.

Theorem 6. (UPPER BOUND FOR THE EXPECTED NUMBER OF CLIQUES.)
Let $K \geq 2$ be a positive integer, let $\delta_n > 0$, and define

$$\widehat{p} = \widehat{p}(p) = 1 - \Phi(t_{p,d}\sqrt{d})$$

Assume

$$d \geq \frac{8(K+1)^2 \ln \frac{1}{\widehat{p}}}{\delta_n^2} \left(K \ln \frac{4}{\widehat{p}} + \ln \frac{K-1}{2} \right) .$$

Then, for any $1 \leq k \leq K$,

$$\mathbb{E}N_k(n,d,1/2) \leq e \binom{n}{k} \Phi(\delta_n)^{\binom{k}{2}} .$$

Furthermore, for $p < 1/2$, define $\beta = 2\sqrt{\ln(4/\widehat{p})}$ and for $\beta\sqrt{K/d} < 1$, let $\alpha = \sqrt{1 - \beta\sqrt{K/d}}$. Then for any $0 < \delta_n < \alpha t_{p,d}\sqrt{d}$ we have, for any $1 \leq k \leq K$,

$$\mathbb{E}N_k(n,d,p) \leq e^{1/\sqrt{2}} \binom{n}{k} \left(1 - \Phi\left(\alpha t_{p,d}\sqrt{d} - \delta_n \right) \right)^{\binom{k}{2}} .$$

Remark. Note that (2) implies that as $\alpha \to 1$ and $\delta_n \to 0$, $1 - \Phi\left(\alpha t_{p,d}\sqrt{d} - \delta_n\right) \to p$.

Proof. Fix a $k \leq K$. We use the Gaussian representation of the $\boldsymbol{X}_i$ described in Section 2. That is, we write $\boldsymbol{X}_i = \boldsymbol{Z}_i/\|\boldsymbol{Z}_i\|$ where $\boldsymbol{Z}_1, \ldots, \boldsymbol{Z}_n$ are independent standard normal vectors in $\mathbb{R}^d$. First we perform Gram-Schmidt orthogonalization for $\boldsymbol{Z}_1^{k-1} = \boldsymbol{Z}_1, \ldots, \boldsymbol{Z}_{k-1}$. In other words, let

$$\boldsymbol{\nu}_1 = \frac{\boldsymbol{Z}_1}{\|\boldsymbol{Z}_1\|}$$

and define $\boldsymbol{r}_1 = \mathbf{0}$ (the d-dimensional zero vector). For $j = 2, \ldots, k-1$, introduce, recursively,

$$\boldsymbol{r}_j = \sum_{i=1}^{j-1} (\boldsymbol{Z}_j, \boldsymbol{v}_i)\boldsymbol{v}_i \quad \text{and} \quad \boldsymbol{v}_j = \frac{\boldsymbol{Z}_j - \boldsymbol{r}_j}{\|\boldsymbol{Z}_j - \boldsymbol{r}_j\|} .$$

Then $\boldsymbol{v}_1, \ldots, \boldsymbol{v}_{k-1}$ are orthonormal vectors, depending on $\boldsymbol{Z}_1^{k-1}$ only.
First we treat the case $p < 1/2$. Introduce the "bad" event

$$B_{k-1} = \left\{ \exists j \le k-1 : \|\boldsymbol{r}_j\|^2 > 2(k+1)^2 \ln(1/\widehat{p}) \text{ or } \exists j \le k-1 : \|\boldsymbol{Z}_j\|^2 < \frac{d}{2} \right\}$$

and write

$$\begin{aligned} p_k &\le \mathbb{P}\{\boldsymbol{X}_1, \ldots, \boldsymbol{X}_k \text{ form a clique}, B_{k-1}^c\} + \mathbb{P}\{B_{k-1}\} \\ &= \mathbb{E}\left[\mathbb{P}\left\{ \left(\frac{\boldsymbol{Z}_k}{\|\boldsymbol{Z}_k\|}, \frac{\boldsymbol{Z}_j}{\|\boldsymbol{Z}_j\|} \right) \ge t_{p,d} \text{ for all } j \le k-1 | \boldsymbol{Z}_1^{k-1} \right\} \mathbb{I}_{\{\boldsymbol{X}_1, \ldots, \boldsymbol{X}_{k-1} \text{ form a clique}\}} \mathbb{I}_{\{B_{k-1}^c\}} \right] \\ &\quad + \mathbb{P}\{B_{k-1}\} . \end{aligned} \tag{3}$$

Now fix $\boldsymbol{Z}_1^{k-1}$ such that $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_{k-1}$ form a clique and B_{k-1} does not occur. Then, for any $\delta_n > 0$,

$$\begin{aligned} &\mathbb{P}\left\{ \left(\frac{\boldsymbol{Z}_k}{\|\boldsymbol{Z}_k\|}, \frac{\boldsymbol{Z}_j}{\|\boldsymbol{Z}_j\|} \right) \ge t_{p,d} \text{ for all } j \le k-1 | \boldsymbol{Z}_1^{k-1} \right\} \\ &= \mathbb{P}\left\{ \left(\boldsymbol{Z}_k, \frac{\boldsymbol{r}_j}{\|\boldsymbol{Z}_j\|} + \frac{\|\boldsymbol{Z}_j - \boldsymbol{r}_j\|}{\|\boldsymbol{Z}_j\|} \boldsymbol{v}_j \right) \ge t_{p,d}\|\boldsymbol{Z}_k\| \text{ for all } j \le k-1 | \boldsymbol{Z}_1^{k-1} \right\} \\ &\le \mathbb{P}\left\{ \left(\boldsymbol{Z}_k, \frac{\|\boldsymbol{Z}_j - \boldsymbol{r}_j\|}{\|\boldsymbol{Z}_j\|} \boldsymbol{v}_j \right) \ge t_{p,d}\|\boldsymbol{Z}_k\| - \delta_n \text{ for all } j \le k-1 | \boldsymbol{Z}_1^{k-1} \right\} \\ &\quad + \sum_{j=1}^{k-1} \mathbb{P}\left\{ \left(\boldsymbol{Z}_k, \frac{\boldsymbol{r}_j}{\|\boldsymbol{Z}_j\|} \right) \ge \delta_n | \boldsymbol{Z}_1^{k-1} \right\} . \end{aligned} \tag{4}$$

For any fixed $1 \le j \le k-1$ and $\delta_n > 0$,

$$\begin{aligned} \mathbb{P}\left\{ \left(\boldsymbol{Z}_k, \frac{\boldsymbol{r}_j}{\|\boldsymbol{Z}_j\|} \right) \ge \delta_n | \boldsymbol{Z}_1^{k-1} \right\} &\le \mathbb{P}\left\{ \left(\boldsymbol{Z}_k, \boldsymbol{r}_j \right) > \delta_n \sqrt{d/2} | \boldsymbol{Z}_1^{k-1} \right\} \\ &\le \frac{1}{2} e^{-\frac{\delta_n^2 d}{4\|\boldsymbol{r}_j\|^2}} \le \frac{1}{2} e^{-\frac{\delta_n^2 d}{8(k+1)^2 \ln \frac{1}{\widehat{p}}}} , \end{aligned} \tag{5}$$

where we used the fact that, conditionally on $\boldsymbol{Z}_1^{k-1}$, $(\boldsymbol{Z}_k, \boldsymbol{r}_j)$ has centered normal distribution with variance $\|\boldsymbol{r}_j\|^2 \le 2(k+1)^2 \ln(1/\widehat{p})$. Furthermore, on B_{k-1}^c, for any $0 < \alpha < 1$, if $\alpha t_{p,d}\sqrt{d} > \delta_n$ then

$$\begin{aligned} &\mathbb{P}\left\{ \left(\boldsymbol{Z}_k, \frac{\|\boldsymbol{Z}_j - \boldsymbol{r}_j\|}{\|\boldsymbol{Z}_j\|} \boldsymbol{v}_j \right) \ge t_{p,d}\|\boldsymbol{Z}_k\| - \delta_n \text{ for all } j \le k-1 | \boldsymbol{Z}_1^{k-1} \right\} \\ &\le \mathbb{P}\left\{ \left(\boldsymbol{Z}_k, \boldsymbol{v}_j \right) \ge t_{p,d}\alpha\sqrt{d} - \delta_n \text{ for all } j \le k-1 | \boldsymbol{Z}_1^{k-1} \right\} + \mathbb{P}\{\|\boldsymbol{Z}_k\| < \alpha\sqrt{d}\} \qquad (6) \\ &\le \left(1 - \Phi\left(\alpha t_{p,d}\sqrt{d} - \delta_n \right) \right)^{k-1} + e^{-\frac{(1-\alpha)^2 d}{4}} , \qquad (7) \end{aligned}$$

where we used the fact that by rotational invariance of the multivariate standard normal distribution, the $(\boldsymbol{Z}_k, \boldsymbol{\nu}_1), \ldots, (\boldsymbol{Z}_k, \boldsymbol{\nu}_{k-1})$ are independent standard normal random variables, and the last term follows from the standard tail bound on the χ^2 distribution

$$\mathbb{P}\{\chi_d^2 < d - 2\sqrt{dt}\} \leq e^{-t} \tag{8}$$

with $t = (1-\alpha^2)^2 d/4$, where χ_d^2 denotes a random variable with χ^2 distribution with d degrees of freedom (see, e.g., Massart [13]). Therefore, the first term in (3) can be bounded as

$$\begin{aligned}
&\mathbb{E}\left[\mathbb{P}\left\{\left(\frac{\boldsymbol{Z}_k}{\|\boldsymbol{Z}_k\|}, \frac{\boldsymbol{Z}_j}{\|\boldsymbol{Z}_j\|}\right) \geq t_{p,d} \text{ for all } j \leq k-1 | \boldsymbol{Z}_1^{k-1}\right\} \mathbb{I}_{\{\boldsymbol{X}_1,\ldots,\boldsymbol{X}_{k-1} \text{ form a clique }\}} \mathbb{I}_{\{B_{k-1}^c\}}\right] \\
&\quad \leq \quad p_{k-1}\left(\left(1 - \Phi\left(\alpha t_{p,d}\sqrt{d} - \delta_n\right)\right)^{k-1} + e^{-\frac{(1-\alpha^2)^2 d}{4}} + \frac{k-1}{2} e^{-\frac{\delta_n^2 d}{8(k+1)^2 \ln(1/\widehat{p})}}\right).
\end{aligned} \tag{9}$$

Using the definition of α, the second term above may be bounded, by

$$e^{-\frac{(1-\alpha^2)^2 d}{4}} \leq \left(\frac{\widehat{p}}{4}\right)^K .$$

The last term in (9) can also be bounded by $(\widehat{p}/4)^K$ using

$$\delta_n^2 \geq \frac{8(k+1)^2 \ln\frac{1}{\widehat{p}}}{d}\left(K \ln\frac{4}{\widehat{p}} + \ln\frac{k-1}{2}\right) .$$

Thus, (9) is bounded from above as

$$\begin{aligned}
&p_{k-1}\left(\left(1 - \Phi\left(\alpha t_{p,d}\sqrt{d} - \delta_n\right)\right)^{k-1} + e^{-\frac{(1-\alpha^2)^2 d}{4}} + \frac{k-1}{2} e^{-\frac{\delta_n^2 d}{8(k+1)^2 \ln(1/\widehat{p})}}\right) \\
&\quad \leq \quad p_{k-1}\left(\left(1 - \Phi\left(\alpha t_{p,d}\sqrt{d} - \delta_n\right)\right)^{k-1} + 2\left(\frac{\widehat{p}}{4}\right)^K\right) \\
&\quad \leq \quad p_{k-1}\left(1 + 2^{-3K+k}\right)\left(1 - \Phi\left(\alpha t_{p,d}\sqrt{d} - \delta_n\right)\right)^{k-1} ,
\end{aligned} \tag{10}$$

where we used the fact that $\widehat{p} \leq 1 - \Phi\left(\alpha t_{p,d}\sqrt{d} - \delta_n\right) < 1/2$ (as $\alpha t_{p,d}\sqrt{d} > \delta_n$ by our assumptions). We may bound the probability of the "bad" event as follows.

$$\begin{aligned}
\mathbb{P}\{B_{k-1}\} \quad &\leq \quad \mathbb{P}\left\{\exists j \leq k-1 : \|\boldsymbol{r}_j\|^2 > 2(k+1)^2 \ln\frac{1}{\widehat{p}}\right\} + \mathbb{P}\left\{\exists j \leq k-1 : \|\boldsymbol{Z}_j\|^2 < \frac{d}{2}\right\} \\
&\leq \quad (k-1)\mathbb{P}\left\{\chi_{k-1}^2 > 2(k+1)^2 \ln\frac{1}{\widehat{p}}\right\} + (k-1)\mathbb{P}\left\{\chi_d^2 < \frac{d}{2}\right\} .
\end{aligned}$$

Here the second term can be bounded by using the tail inequality (8) with $t = d/16$, which yields $\mathbb{P}\{\chi_d^2 < d/2\} \leq e^{-d/16}$. The first term can be bounded using the standard tail bound

$$\mathbb{P}\{\chi_l^2 - l > 2t + 2\sqrt{lt}\} \leq e^{-t} \tag{11}$$

(see [13]) with

$$t = \frac{\left(\sqrt{4(k+1)^2 \ln \frac{1}{\widehat{p}}}) - \sqrt{k-1}\right)^2}{4} = (k+1)^2 \ln \frac{1}{\widehat{p}} - \frac{\sqrt{(k-1)\left(4(k+1)^2 \ln \frac{1}{\widehat{p}}) - k + 1\right)}}{2}$$

and $l = k-1$, which implies

$$\mathbb{P}\{\chi^2_{k-1} > 2(k+1)^2 \ln(1/\widehat{p})\} \le e^{-2(k+1)^2 \ln(1/\widehat{p})/4} = \widehat{p}^{(k+1)^2/2} .$$

Thus

$$\mathbb{P}\{B_{k-1}\} \le (k-1)\left(\widehat{p}^{(k+1)^2/2} + e^{-d/16}\right) . \tag{12}$$

If, in addition, $d \ge 8(k+1)^2 \ln(1/\widehat{p})$, we obtain

$$\mathbb{P}\{B_{k-1}\} \le 2(k-1)\widehat{p}^{(k+1)^2/2}, \tag{13}$$

and so, by (3), (9) and (10) we have

$$p_k \le p_{k-1}\left(1 + 2^{-3K+k}\right)\left(1 - \Phi\left(\alpha t_{p,d}\sqrt{d} - \delta_n\right)\right)^{k-1} + 2(k-1)\widehat{p}^{(k+1)^2/2}. \tag{14}$$

Next we show that

$$p_k \le \left(1 - \Phi\left(\alpha t_{p,d}\sqrt{d} - \delta_n\right)\right)^{\binom{k}{2}} \prod_{j=1}^{k-1}(1 + 2^{-j-1/2}) \tag{15}$$

which finishes the proof of the theorem for $p < 1/2$ since $\prod_{j=1}^{k}(1+2^{-j-1/2}) \le e^{\sum_{j=1}^{k} 2^{-j-1/2}} < e^{1/sqrt2}$. We proceed by induction. (15) trivially holds for $k = 1$. Assuming it holds for $k-1$ for some $k \ge 2$, from (14) we obtain

$$\begin{aligned}
p_k &\le \left(1 - \Phi\left(\alpha t_{p,d}\sqrt{d} - \delta_n\right)\right)^{\binom{k-1}{2}} \left(\prod_{j=1}^{k-2}(1 + 2^{-j-1/2})\right)\left(1 + 2^{-3K+k}\right)\left(1 - \Phi\left(\alpha t_{p,d}\sqrt{d} - \delta_n\right)\right)^{k-1} \\
&\quad + 2(k-1)\widehat{p}^{(k+1)^2/2} \\
&\le \left(1 - \Phi\left(\alpha t_{p,d}\sqrt{d} - \delta_n\right)\right)^{\binom{k}{2}} \left(\prod_{j=1}^{k-2}(1 + 2^{-j-1/2})\right)\left(1 + 2^{-3K+k} + 2(k-1)2^{-\frac{3k+1}{2}}\right) \\
&\le \left(1 - \Phi\left(\alpha t_{p,d}\sqrt{d} - \delta_n\right)\right)^{\binom{k}{2}} \prod_{j=1}^{k-1}(1 + 2^{-j-1/2})
\end{aligned}$$

where we used that $\widehat{p} \le 1 - \Phi\left(\alpha t_{p,d}\sqrt{d} - \delta_n\right) < 1/2$ (as $\alpha t_{p,d}\sqrt{d} > \delta_n$ by our assumptions) and that $2^{-3K+k} + 2(k-1)2^{-\frac{3k+1}{2}} < 2^{-k+1/2}$ for $K \ge 2$ since $2(k-1)2^{-k/2} \le 3/2$ for all k. This completes the proof of (15), and hence that of the theorem for $p < 1/2$.

For $p = 1/2$, we need the following modifications. B_{k-1} is now defined as

$$B_{k-1} = \left\{\exists j \le k-1 : \|\boldsymbol{r}_j\|^2 > 2(k+1)^2 \ln 2 \text{ or } \exists j \le k-1 : \|\boldsymbol{Z}_j - \boldsymbol{r}_j\|^2 < \frac{d}{2}\right\} .$$

Then (3) still holds, but instead of (4) we write

$$\begin{aligned}
&\mathbb{P}\left\{\left(\frac{\boldsymbol{Z}_k}{\|\boldsymbol{Z}_k\|},\frac{\boldsymbol{Z}_j}{\|\boldsymbol{Z}_j\|}\right)\geq 0 \text{ for all } j\leq k-1|\boldsymbol{Z}_1^{k-1}\right\}\\
&\quad\leq\quad \mathbb{P}\left\{\left(\boldsymbol{Z}_k,\boldsymbol{v}_j\right)\geq -\delta_n \text{ for all } j\leq k-1|\boldsymbol{Z}_1^{k-1}\right\}+\sum_{j=1}^{k-1}\mathbb{P}\left\{\left(\boldsymbol{Z}_k,\frac{\boldsymbol{r}_j}{\|\boldsymbol{Z}_j-\boldsymbol{r}_j\|}\right)>\delta_n|\boldsymbol{Z}_1^{k-1}\right\}\ .
\end{aligned}$$

From here, similarly to (5) and (7), the following analog of (9) can be obtained:

$$\begin{aligned}
&\mathbb{E}\left[\mathbb{P}\left\{\left(\boldsymbol{X}_k,\boldsymbol{X}_j\right)\geq t_{p,d} \text{ for all } j\leq k-1|\boldsymbol{Z}_1^{k-1}\right\}\mathbb{I}_{\{\boldsymbol{X}_1,\ldots,\boldsymbol{X}_{k-1}\text{ form a clique }\}}\mathbb{I}_{\{B_{k-1}^c\}}\right]\\
&\quad\leq\quad p_{k-1}\left(\left(1-\Phi\left(-\delta_n\right)\right)^{k-1}+\frac{k-1}{2}e^{-\frac{\delta_n^2 d}{8(k+1)^2\ln 2}}\right).
\end{aligned}$$

As the bound (12) remains valid for the redefined B_{k-1} (with $\widehat{p}=1/2$), the proof may be finished as before for $d\geq 8(k+1)^2\ln 2$ and

$$\delta_n^2\geq\frac{8(K+1)^2\ln 2}{d}\left(K\ln\frac{4}{\Phi(\delta_n)}+\ln\frac{K-1}{2}\right)\ .$$

□

Theorem 7. (LOWER BOUND FOR THE EXPECTED NUMBER OF CLIQUES.) *Introduce*

$$\widetilde{p}=\widetilde{p}(p)=\begin{cases}1/2 & \text{if } p=1/2;\\ 1-\Phi\left(2t_{p,d}\sqrt{d}+1\right) & \text{if } p<1/2;\end{cases}$$

and let $\delta_n\in(0,2/3]$ *and* $K\geq 3$. *Assume*

$$d>\frac{8(K+1)^2\ln\frac{1}{\widetilde{p}}}{\delta_n^2}\left(K\ln\frac{4}{\widetilde{p}}+\ln\frac{(K-1)}{2}\right)\ . \tag{16}$$

Then, for any $1\leq k\leq K$,

$$\mathbb{E}N_k(n,d,1/2)\geq\frac{4}{5}\binom{n}{k}\left(1-\Phi(\delta_n)\right)^{\binom{k}{2}}\ .$$

For $p<1/2$, *define* $\alpha>0$ *as*

$$\alpha^2=1+\sqrt{\frac{8K}{d}\ln\frac{4}{\widetilde{p}}}\ .$$

Then

$$\mathbb{E}N_k(n,d,p)\geq\frac{4}{5}\binom{n}{k}\left(1-\widetilde{\Phi}_K(d,p)\right)^{\binom{k}{2}}\ , \tag{17}$$

where $\widetilde{\Phi}_K(d,p)=\Phi\left(\frac{\alpha t_{p,d}\sqrt{d}+\delta_n}{\sqrt{1-\frac{2(K+1)^2\ln(1/\widetilde{p})}{d}}}\right)$.

Proof. The proof is a simple variant of the previous theorem, and we use the notation introduced there. Fix a $k \le K$. Define the "bad" event $\widetilde{B}_{k-1}$ as

$$\widetilde{B}_{k-1} = \left\{ \exists j \le k-1 : \|\boldsymbol{r}_j\|^2 > 2(k+1)^2 \ln(1/\widetilde{p}) \text{ or } \exists j \le k-1 : \|\boldsymbol{Z}_j - \boldsymbol{r}_j\|^2 < \frac{d}{2} \right\} . \tag{18}$$

Then

$$\begin{aligned} p_k &\ge \mathbb{P}\{\boldsymbol{X}_1, \ldots, \boldsymbol{X}_k \text{ form a clique }, \widetilde{B}_{k-1}^c\} \qquad (19) \\ &= \mathbb{E}\left[\mathbb{P}\left\{ \left(\frac{\boldsymbol{Z}_k}{\|\boldsymbol{Z}_k\|}, \frac{\boldsymbol{Z}_j}{\|\boldsymbol{Z}_j\|} \right) \ge t_{p,d} \text{ for all } \ j \le k-1 | \boldsymbol{Z}_1^{k-1} \right\} \mathbb{I}_{\{\boldsymbol{X}_1, \ldots, \boldsymbol{X}_{k-1} \text{ form a clique}\}} \mathbb{I}_{\{\widetilde{B}_{k-1}^c\}} \right] \end{aligned}$$

Fix $\boldsymbol{Z}_1, \ldots, \boldsymbol{Z}_{k-1}$ such that they form a clique and $\widetilde{B}_{k-1}$ does not occur. Then

$$\begin{aligned} &\mathbb{P}\left\{ \left(\frac{\boldsymbol{Z}_k}{\|\boldsymbol{Z}_k\|}, \frac{\boldsymbol{Z}_j}{\|\boldsymbol{Z}_j\|} \right) \ge t_{p,d} \text{ for all } j \le k-1 | \boldsymbol{Z}_1^{k-1} \right\} \\ &\quad \ge \mathbb{P}\left\{ \left(\boldsymbol{Z}_k, \frac{\|\boldsymbol{Z}_j - \boldsymbol{r}_j\|}{\|\boldsymbol{Z}_j\|} \boldsymbol{v}_j \right) \ge t_{p,d} \|\boldsymbol{Z}_k\| + \delta_n \text{ for all } j \le k-1 | \boldsymbol{Z}_1^{k-1} \right\} \\ &\qquad - \sum_{j=1}^{k-1} \mathbb{P}\left\{ \left(\boldsymbol{Z}_k, \frac{\boldsymbol{r}_j}{\|\boldsymbol{Z}_j\|} \le -\delta_n \right) | \boldsymbol{Z}_1^{k-1} \right\} . \end{aligned} \tag{20}$$

Now for any $\alpha > 1$, the first term can be bounded as

$$\begin{aligned} &\mathbb{P}\left\{ \left(\boldsymbol{Z}_k, \frac{\|\boldsymbol{Z}_j - \boldsymbol{r}_j\|}{\|\boldsymbol{Z}_j\|} \boldsymbol{v}_j \right) \ge t_{p,d} \|\boldsymbol{Z}_k\| + \delta_n \text{ for all } j \le k-1 | \boldsymbol{Z}_1^{k-1} \right\} \\ &\ge \mathbb{P}\left\{ \left(\boldsymbol{Z}_k, \sqrt{1 - \frac{2(k+1)^2 \ln(1/\widetilde{p})}{d}} \boldsymbol{v}_j \right) \ge t_{p,d} \|\boldsymbol{Z}_k\| + \delta_n \text{ for all } j \le k-1 | \boldsymbol{Z}_1^{k-1} \right\} \\ &\ge \mathbb{P}\left\{ \left(\boldsymbol{Z}_k, \boldsymbol{v}_j \right) \ge \frac{\alpha t_{p,d} \sqrt{d} + \delta_n}{\sqrt{1 - \frac{2(k+1)^2 \ln(1/\widetilde{p})}{d}}} \text{ for all } j \le k-1 | \boldsymbol{Z}_1^{k-1} \right\} - \mathbb{P}\left\{ \|\boldsymbol{Z}_k\| > \alpha \sqrt{d} \right\} \\ &\ge \left(1 - \widetilde{\Phi}_k(d,p) \right)^{k-1} - e^{-\frac{d}{4} \cdot \frac{(\alpha^2-1)^2}{\alpha^2}} , \end{aligned} \tag{21}$$

where the first inequality holds since $\boldsymbol{r}_j$ and $\boldsymbol{Z}_j - \boldsymbol{r}_j$ are orthogonal and $\|\boldsymbol{r}_j\|^2 < 2(k+1)^2 \ln(1/\widetilde{p})$ on $\widetilde{B}_{k-1}$, implying $\|\boldsymbol{Z}_j - \boldsymbol{r}_j\| / \|\boldsymbol{Z}_j\| \ge \sqrt{1 - 2(k+1)^2 \ln(1/\widetilde{p})/d}$, and the last inequality follows again by (11) (with $t = \frac{d}{4} \cdot \frac{(\alpha^2-1)^2}{\alpha^2}$) and the fact that the $(\boldsymbol{Z}_k, \boldsymbol{v}_1), \ldots, (\boldsymbol{Z}_k, \boldsymbol{v}_{k-1})$ are independent standard normal random variables. The second term in (20) can be bounded similarly to (5). The conditions

of the theorem for α and d imply

$$\begin{aligned}
&\mathbb{P}\left\{\left(\frac{\boldsymbol{Z}_k}{\|\boldsymbol{Z}_k\|},\frac{\boldsymbol{Z}_j}{\|\boldsymbol{Z}_j\|}\right)\geq t_{p,d} \text{ for all } j\leq k-1|\boldsymbol{Z}_1^{k-1}\right\}\\
&\quad\geq \left(1-\widetilde{\Phi}_k(d,p)\right)^{k-1}-e^{-\frac{d}{4}\cdot\frac{(\alpha^2-1)^2}{\alpha^2}}-\frac{k-1}{2}e^{-\frac{\delta_n^2 d}{8(k+1)^2\ln\frac{1}{\widetilde{p}}}} \qquad (22)\\
&\quad\geq \left(1-\widetilde{\Phi}_k(d,p)\right)^{k-1}-2\left(\frac{\widetilde{p}}{4}\right)^K\\
&\quad\geq \left(1-\widetilde{\Phi}_k(d,p)\right)^{k-1}\left(1-2^{-2K+1}\right) \qquad (23)
\end{aligned}$$

where at the last step we used the fact that $\widetilde{p}<1-\widetilde{\Phi}_k(d,p)<1/2$,[1]

To finish the proof of (17), we proceed, again, by induction, to prove that

$$p_k\geq \eta_K^k\left(1-\sum_{i=2}^k 4^{-i}\right)\left(1-\widetilde{\Phi}_k(d,p)\right)^{\binom{k}{2}}$$

with $\eta_K=\left(1-2^{-2K+1}\right)$. This is sufficient to prove the theorem because $\eta_K^k\left(1-\sum_{i=2}^k 4^{-i}\right)>4/5$ for all $k\leq K$ when $K\geq 3$. This clearly holds for $k=1$. Assuming it holds for some $k-1$, $k\geq 2$, and taking into account that, similarly to (13),

$$\mathbb{P}\{\widetilde{B}_{k-1}\}\leq 2(k-1)\widetilde{p}^{(k+1)^2/2}\leq 2(k-1)\left(1-\widetilde{\Phi}_k(d,p)\right)^{(k+1)^2/2},$$

we obtain

$$\begin{aligned}
p_k &\geq \eta_K\left(1-\widetilde{\Phi}_k(d,p)\right)^{k-1}\left(p_{k-1}-\mathbb{P}\{\widetilde{B}_{k-1}\}\right)_+\\
&\geq \eta_K\left(1-\widetilde{\Phi}_k(d,p)\right)^{k-1}\\
&\quad\times\left(\eta_K^{k-1}\left(1-\sum_{i=2}^{k-1}4^{-i}\right)\left(1-\widetilde{\Phi}_k(d,p)\right)^{\binom{k-1}{2}}-2(k-1)\left(1-\widetilde{\Phi}_k(d,p)\right)^{(k+1)^2/2}\right)_+\\
&= \eta_K\left(1-\widetilde{\Phi}_k(d,p)\right)^{\binom{k}{2}}\left(\eta_K^{k-1}\left(1-\sum_{i=2}^{k-1}4^{-i}\right)-2(k-1)\left(1-\widetilde{\Phi}_k(d,p)\right)^{(5k+1)/2}\right)_+\\
&\geq \eta_K^k\left(1-\widetilde{\Phi}_k(d,p)\right)^{\binom{k}{2}}\left(1-\sum_{i=2}^{k-1}4^{-i}-4^{-k}\right)
\end{aligned}$$

where $x_+=\max(x,0)$ denotes the positive part of a real number x and we used that $1-\widetilde{\Phi}_k(d,p)<1/2$ and $2(k-1)2^{-k/2-1}<\eta_K^k<\eta_K^{k-1}$ for all $2\leq k\leq K$ when $K\geq 3$.

[1]The second inequality is trivial. The first one can be obtained by noting that (16) implies $\alpha\leq\sqrt{1+\delta_n}\leq 1+\delta_n/2$ and $2(K+1)^2\ln(1/\widetilde{p})/d<\delta_n^2/4$. From here, using $\delta_n\leq 2/3$, we have

$$\frac{\alpha t_{p,d}\sqrt{d}+\delta_n}{\sqrt{1-\frac{2(K+1)^2\ln(1/\widetilde{p})}{d}}}<\frac{(1+\delta_n/2)t_{p,d}\sqrt{d}+\delta_n}{1-\delta_n/2}<2t_{p,d}\sqrt{d}+1$$

which implies $\widetilde{p}<1-\widetilde{\Phi}_K(d,p)$.

For $p = 1/2$, we proceed similarly: (19) also holds in this case, but instead of (20)-(23) we have

$$\begin{aligned}
&\mathbb{P}\left\{\left(\frac{\boldsymbol{Z}_k}{\|\boldsymbol{Z}_k\|}, \frac{\boldsymbol{Z}_j}{\|\boldsymbol{Z}_j\|}\right) \geq 0 \text{ for all } j \leq k-1 | \boldsymbol{Z}_1^{k-1}\right\} \\
&\geq \quad \mathbb{P}\left\{\left(\boldsymbol{Z}_k, \boldsymbol{v}_j\right) \geq \delta_n \text{ for all } j \leq k-1 | \boldsymbol{Z}_1^{k-1}\right\} - \sum_{j=1}^{k-1} \mathbb{P}\left\{\left(\boldsymbol{Z}_k, \frac{\boldsymbol{r}_j}{\|\boldsymbol{Z}_j - \boldsymbol{r}_j\|}\right) \leq -\delta_n | \boldsymbol{Z}_1^{k-1}\right\} \\
&\geq \quad \left(1 - \Phi(\delta_n)\right)^{k-1} - \frac{k-1}{2} e^{-\frac{\delta_n^2 d}{8(k+1)^2 \ln 2}} \\
&\geq \quad \left(1 - \Phi(\delta_n)\right)^{k-1} \left(1 - 2^{-2K+1}\right) .
\end{aligned}$$

From here the proof can be finished easily as in the case of $p < 1/2$. □

3.2 Upper bound for the clique number

Theorem 8. *Let $p \in (0, 1/2]$, and let $\delta_n \in (0, p)$, be such that either $\delta_n \to 0$ or $\delta_n \equiv \delta$ for some $\delta \in (0, p)$. Define $\widehat{p}$ as in Theorem 6. If*

$$d \geq \frac{65 \ln^2(1/\widehat{p})}{\delta_n^2} \log^3_{1/(p+\delta_n)} n$$

then, with probability converging to 1 (as n tends to infinity),

$$\omega(n, d, p) \leq 2 \log_{1/(p+\delta_n)} n - 2 \log_{1/(p+\delta_n)} \log_{1/(p+\delta_n)} n + O(1) .$$

Proof. We use the first moment method. Denote the (random) number of cliques of size k by N_k. Since

$$\mathbb{P}\{\omega(n, d, 1/2) \geq k\} = \mathbb{P}\{N_k \geq 1\} \leq \mathbb{E} N_k ,$$

it suffices to show that $\mathbb{E} N_k \to 0$ for the clique sizes indicated in the statement of the theorem. We apply Theorem 6. Let $K = \lfloor 2 \log_{1/(p+\delta_n)} n \rfloor$. Note that, for large enough n, any d and δ_n satisfying the conditions in this theorem also satisfy the conditions in Theorem 6. We will show that the conditions of the theorem imply

$$1 - \Phi(\alpha t_{p,d} \sqrt{d} - \delta_n) < p + \delta_n \tag{24}$$

where α is defined as in Theorem 6. Once we show this, we have, for all $k \leq K$, $\mathbb{E} N_k \leq e \binom{n}{k} (p + \delta_n)^{\binom{k}{2}}$ which converges to zero for all possible feasible choices of δ_n if

$$k \geq 2 \log_b n - 2 \log_b \log_b n + 1 + 2 \log_b (e/2) + \epsilon'$$

for any fixed $\epsilon' > 0$ where $b = 1/(p + \delta_n)$ (see Palmer [15] or Bollobás [5]), proving the theorem, since $K \geq k$ if n is sufficiently large.

It remains to show (24). It clearly holds for $p = 1/2$, since $\Phi(-\delta_n) \geq 1/2 - \delta_n/\sqrt{2\pi} \geq 1/2 - \delta_n$. For $p < 1/2$, we have

$$\begin{aligned} 1 - \Phi(\alpha t_{p,d}\sqrt{d} - \delta_n) &= 1 - \Phi\left(\Phi^{-1}(1-p) + (t_{p,d}\sqrt{d} - \Phi^{-1}(1-p)) - (1-\alpha)t_{p,d}\sqrt{d} - \delta_n\right) \\ &\leq 1 - \Phi\left(\Phi^{-1}(1-p)\right) + \frac{|t_{p,d}\sqrt{d} - \Phi^{-1}(1-p)| + (1-\alpha)t_{p,d}\sqrt{d} + \delta_n}{\sqrt{2\pi}} \\ &\leq p + \frac{U_{p,d} + \beta\sqrt{K/d}(\Phi^{-1}(1-p) + U_{p,d}) + \delta_n}{\sqrt{2\pi}} \\ &\leq p + \delta_n \end{aligned}$$

for all sufficiently large n, where $U_{p,d}$ is defined in Lemma 1, and we used that $1 - \alpha \leq \beta\sqrt{K/d} < 1$, $U_{p,d} = O(\sqrt{\log d/d})$, and therefore $U_{p,d} = o(\delta_n)$ by the condition of the theorem. □

3.3 Lower bound for the clique number

Theorem 9. *Let $p \in (0, 1/2]$, and let $\delta_n \in (0, p)$, be such that $\limsup_{n\to\infty} \delta_n \log^2 n < \infty$, Define $\widetilde{p}$ as in Theorem 3. Suppose*

$$d = d_n \geq \frac{65\ln^2(1/\widetilde{p})}{\delta_n^2} \log^3_{1/(p-\delta_n)} n \,.$$

Then, with probability converging to 1 (as $n \to \infty$),

$$\omega(n, d, p) \geq 2\log_{1/(p-\delta_n)} n - 2\log_{1/(p-\delta_n)} \log_{1/(p-\delta_n)} n + O(1) \,.$$

Proof. In order to prove a lower bound by the second moment method, we need to show that $var(N_k)/(\mathbb{E}N_k)^2 \to 0$ for

$$k = 2\log_{1/(p-\delta_n)} n - 2\log_{1/(p-\delta_n)} \log_{1/(p-\delta_n)} n - C \tag{25}$$

with some appropriate constant $C > 0$, since then we have

$$\mathbb{P}\{N_k \geq 1\} \geq \frac{1}{1 + var(N_k)/(\mathbb{E}N_k)^2} \to 1.$$

Recall that if $p_k = \mathbb{P}\{\boldsymbol{X}_1, \ldots, \boldsymbol{X}_k \text{ form a clique}\}$ then

$$(\mathbb{E}N_k)^2 = \binom{n}{k}^2 p_k^2$$

On the other hand,

$$\begin{aligned} &var(N_k) \\ &\leq \Delta_k \\ &\stackrel{\text{def}}{=} \sum_{m=2}^{k} \binom{n}{k}\binom{k}{m}\binom{n-k}{k-m} \\ &\quad \times \mathbb{P}\{\boldsymbol{X}_1, \ldots, \boldsymbol{X}_k \text{ form a clique and } \boldsymbol{X}_{k-m+1}, \boldsymbol{X}_{k-m+2}, \ldots, \boldsymbol{X}_{2k-m} \text{ form a clique}\} \end{aligned}$$

and therefore it suffices to prove that $\Delta_k/(\mathbb{E}N_k)^2 \to 0$. To this end, we may write

$$\begin{aligned}
&\mathbb{P}\{\boldsymbol{X}_1,\ldots,\boldsymbol{X}_k \text{ form a clique and } \boldsymbol{X}_{k-m+1},\boldsymbol{X}_{k-m+2},\ldots,\boldsymbol{X}_{2k-m} \text{ form a clique}\} \\
&\quad = \mathbb{E}\left[\mathbb{I}_{\{\boldsymbol{X}_1,\ldots,\boldsymbol{X}_k \text{ form a clique}\}}\mathbb{P}\{\boldsymbol{X}_{k-m+1},\boldsymbol{X}_{k-m+2},\ldots,\boldsymbol{X}_{2k-m} \text{ form a clique}|\boldsymbol{X}_1,\ldots,\boldsymbol{X}_k\}\right] \\
&\quad = \mathbb{E}\left[\mathbb{I}_{\{\boldsymbol{X}_1,\ldots,\boldsymbol{X}_k \text{ form a clique}\}}\mathbb{P}\{\boldsymbol{X}_{k-m+1},\boldsymbol{X}_{k-m+2},\ldots,\boldsymbol{X}_{2k-m} \text{ form a clique}|\boldsymbol{X}_{k-m+1},\ldots,\boldsymbol{X}_k\}\right] .
\end{aligned}$$

Now the conditional probability $p_k^{(m)} = \mathbb{P}\{\boldsymbol{X}_{k-m+1},\boldsymbol{X}_{k-m+2},\ldots,\boldsymbol{X}_{2k-m} \text{ form a clique}|\boldsymbol{X}_{k-m+1},\ldots,\boldsymbol{X}_k\}$ may be bounded similarly to the last $k-m$ steps of the same recursive argument of the proof of Theorem 6, since (15) holds for $p_k^{(m)}$ in place of p_k. Under the same conditions as there, we obtain

$$\begin{aligned}
&\mathbb{P}\{\boldsymbol{X}_{k-m+1},\boldsymbol{X}_{k-m+2},\ldots,\boldsymbol{X}_{2k-m} \text{ form a clique}|\boldsymbol{X}_{k-m+1},\ldots,\boldsymbol{X}_k\} \\
&\quad \le e^{1/\sqrt{2}}\left(1-\Phi\left(\alpha t_{p,d}\sqrt{d}-\delta_n\right)\right)^{\binom{k}{2}-\binom{m}{2}} \le e^{1/\sqrt{2}}(p+\delta_n)^{\binom{k}{2}-\binom{m}{2}}
\end{aligned}$$

where the second inequality follows from (24). Thus, we have

$$\begin{aligned}
&\mathbb{P}\{\boldsymbol{X}_1,\ldots,\boldsymbol{X}_k \text{ form a clique and } \boldsymbol{X}_{k-m+1},\boldsymbol{X}_{k-m+2},\ldots,\boldsymbol{X}_{2k-m} \text{ form a clique}\} \\
&\qquad\qquad \le p_k e^{1/\sqrt{2}}(p+\delta_n)^{\binom{k}{2}-\binom{m}{2}} .
\end{aligned}$$

From here, we obtain

$$\frac{\Delta_k}{(\mathbb{E}N_k)^2} \le \frac{e^{1/\sqrt{2}}(p+\delta_n)^{\binom{k}{2}}}{p_k}\sum_{m=2}^{k}\frac{\binom{k}{m}\binom{n-k}{k-m}(p+\delta_n)^{-\binom{m}{2}}}{\binom{n}{k}} .$$

Now we may lower bound for p_k as

$$p_k \ge \frac{4}{5}(p-\delta_n)^{\binom{k}{2}} ,$$

where the last inequality holds for $k \le K = \lfloor 2\log_{1/(p-\delta_n)} n\rfloor$ by Theorem 7 since its conditions are satisfied by the assumptions of our theorem for the actual choice of K and

$$1-\widetilde{\Phi}_K(d,p) \ge p-\delta_n , \tag{26}$$

where $\widetilde{\Phi}(d,1/2) = \Phi(\delta_n)$, as it will be shown later.

Summarizing,

$$\frac{\Delta_k}{(\mathbb{E}N_k)^2} \le \frac{5}{4}e^{1/\sqrt{2}}\left(\frac{p+\delta_n}{p-\delta_n}\right)^{\binom{k}{2}}\sum_{m=2}^{k}\frac{\binom{k}{m}\binom{n-k}{k-m}(p+\delta_n)^{-\binom{m}{2}}}{\binom{n}{k}} .$$

Notice that by the condition of the theorem for δ_n, the factor $\big((p+\delta_n)/(p-\delta_n)\big)^{\binom{k}{2}}$ is bounded and therefore it suffices to show that

$$\sum_{m=2}^{k}\frac{\binom{k}{m}\binom{n-k}{k-m}(p+\delta_n)^{-\binom{m}{2}}}{\binom{n}{k}} \to 0$$

which follows from the same calculations as in Bollobás [5, Corollary 11.2] for the actual choice of k given in (25) – see also Palmer [15, Theorem 5.3.1]. Thus we have $\Delta_k/(\mathbb{E}N_k)^2 \to 0$.

To finish the proof, we need to show (26). The statement clearly holds for $p = 1/2$. For $p < 1/2$, using α as defined in Theorem 7, and defining $\eta = \sqrt{1 - \frac{2(K+1)^2 \ln(1/\widetilde{p})}{d}}$, we have

$$\begin{aligned}
1 - \widetilde{\Phi}_K(d,p) &= 1 - \Phi\left(\frac{\alpha t_{p,d}\sqrt{d} + \delta_n}{\eta}\right) \\
&= 1 - \Phi\left(\Phi^{-1}(1-p) + \left(t_{p,d}\sqrt{d} - \Phi^{-1}(1-p)\right) + \left(\frac{\alpha}{\eta} - 1\right) t_{p,d}\sqrt{d} + \frac{\delta_n}{\eta}\right) \\
&\geq p - \frac{|t_{p,d}\sqrt{d} - \Phi^{-1}(1-p)| + \left(\frac{\alpha}{\eta} - 1\right) t_{p,d}\sqrt{d} + \frac{\delta_n}{\eta}}{\sqrt{2\pi}} \\
&\geq p - \frac{U_{p,d} + \left(\frac{8(K+1)^2 \ln\frac{1}{\widetilde{p}}}{3d}\right)\left(\Phi^{-1}(1-p) + U_{p,d}\right) + \frac{4}{3}\delta_n}{\sqrt{2\pi}},
\end{aligned}$$

where $U_{p,d}$ is defined in Lemma 1, and in the last step we used that

$$\eta \geq \eta^2 = 1 - \frac{2(K+1)^2 \ln(1/\widetilde{p})}{d} \geq 3/4$$

by the conditions of the theorem, and

$$\frac{\alpha}{\eta} - 1 \leq \frac{\alpha^2}{\eta^2} - 1 \leq \frac{8}{3}\frac{(K+1)^2 \ln(1/\widetilde{p})}{d}.$$

From here (26) follows by the fact that

$$\frac{U_{p,d} + \frac{8(K+1)^2 \ln\frac{1}{\widetilde{p}}}{3d}\left(\Phi^{-1}(1-p) + U_{p,d}\right) + \frac{4}{3}\delta_n}{\sqrt{2\pi}} < \delta_n$$

for all sufficiently large n by the condition of the theorem on δ_n and d. □

4 Numerical experiments

In this section we report some numerical experiments in which we simulated random geometric graphs $\overline{G}(n,d,p)$ for various values of the parameters and computed their clique number. Since finding the clique number is a computationally difficult problem, we had to limit ourselves to moderate values of n and/or p.

In Figure 2 the expected clique number of $\overline{G}(n,d,p)$ is approximated by averaging $B = 2000$ simulated graphs for some values of n, p, d. The horizontal axis is the dimension d, while the vertical axis is the clique number. The value of n is color coded. Note that as n increases, the curve is higher in the graph.

To keep running time under control, our algorithm reports failure when a clique of size 20 or higher is found. Such occurrences do not appear in the figure.

On the right side of each graph, the expected clique number of the corresponding Erdős-Rényi random graph is shown: small disks plot the value obtained as the average of $B = 2000$ simulations, while small dashes plot the value as obtained by the asymptotic formula $2\log_{1/p} n - 2\log_{1/p}\log_{1/p} n$.

In Figure 3 we plot the approximated value of $\mathbb{E}\omega(n,d,p)$ for the entire range of values of $p \in [0,1]$ when $n = 15$, $n = 50$, and $n = 100$.

5 An application: a problem of testing dependency

Finally, we describe a statistical hypothesis testing problem from remote sensing and finance. Upon observing random vectors $\boldsymbol{Z}_1, \ldots, \boldsymbol{Z}_n$, each of d independent components, one wishes to test whether these vectors are independent or, alternatively, if there exists a small group of vectors that depend on each other. In remote sensing the n vectors represent the signal captured at n sensors in a noisy environment and one wishes to determine if there is a subset of the sensors that detect a common weak signal. In financial applications the n vectors represent the evolution of the price of n assets and one may be interested in the existence of a small subset that depend on each other in a certain way.

The simplest way to formalize such a hypothesis testing problem is the following. Under the *null hypothesis* H_0, all vectors $\boldsymbol{Z}_i$ are standard normal (i.e., with mean $\mathbf{0}$ and unit covariance matrix). Under the *alternative hypothesis* H_1 there is a small subset of vectors that are more correlated among themselves. This may be modeled as follows. Under H_1, there exists a subset $S \subset \{1,\ldots,n\}$ with a given size $|S| = m \ll n$ such that $\boldsymbol{Z}_i = (Z_{i,1},\ldots,Z_{i,d})$ where

$$Z_{i,t} = \begin{cases} N_{i,t} & \text{if } i \notin S \\ (N_{i,t} + Y_t)/\sqrt{1+\sigma^2} & \text{if } i \in S, \end{cases}$$

where the $N_{i,t}$ are i.i.d. standard normal random variables and $Y_1,\ldots,Y_t$ are independent normal $(0,\sigma^2)$ random variables, independent of the $N_{i,t}$. The Y_t represent a common "signal" present in the vectors $\boldsymbol{Z}_i$ for $i \in S$. Clearly, $Z_{i,t}$ is standard normal for all i and t and $\mathbb{E}Z_{i,t}Z_{j,t} = 0$ if either i or j are not in S. If $i,j \in S$, then $\mathbb{E}Z_{i,t}Z_{j,t} = \sigma^2/(1+\sigma^2)$. In the applications of interest, $\sigma \ll 1$ and it is so small that calculating simply the correlation of $\boldsymbol{Z}_i$ and $\boldsymbol{Z}_j$ one cannot easily tell whether both i and j are in S or not. In particular, the largest empirical correlations $(\boldsymbol{X}_i, \boldsymbol{X}_j)$ do not necessarily belong to vertices belonging to S unless $d\sigma^2 \gg \sqrt{d\log n}$. The interesting values of σ^2 are those for which the "signal" is covered by "noise". Clearly, if d is sufficiently large, the problem becomes easy but in the applications mentioned above it is important to keep the value of d as small as possible in order to make quick decisions.

A simple test is obtained by considering the random geometric graph defined by the normalized vectors $\boldsymbol{X}_i = \boldsymbol{Z}_i/\|\boldsymbol{Z}_i\|$. Fix some p (say $p = 1/2$ for concreteness and simplicity), and define the random geometric graph based on the points $\boldsymbol{X}_1,\ldots,\boldsymbol{X}_n$, connecting vertex i and vertex j if and only if $(\boldsymbol{X}_i, \boldsymbol{X}_j) \geq t_{p,d}$. (Recall that $t_{1/2,d} = 0$ for all d.) In other words, vertices i and j are connected if and only if the empirical correlation $(\boldsymbol{X}_i, \boldsymbol{X}_j)$ of the observed vectors $\boldsymbol{Z}_i$ and $\boldsymbol{Z}_j$ exceeds the threshold $t_{p,d}$. The test is based on computing the clique number of the obtained graph. Under the null hypothesis, the obtained random graph is just $\overline{G}(n,d,p)$ studied in this paper and Theorem 3 summarizes some properties of its clique number.

Under the alternative hypothesis, one expects that for sufficiently large values of σ, vertices belonging to S form a clique. This intuition may be quantified by the following lemma. For simplicity we only consider the case $p = 1/2$. For different values of p a similar argument carries over though a few straightforward but technical details need to be taken care of. It is not our goal here to optimize the power of the test and for the purpose of illustration the case $p = 1/2$ is sufficient.

Lemma 10. *Let $\delta_n \in (0,1)$ and consider the random geometric graph with $p = 1/2$ defined by the points $\boldsymbol{X}_i = \boldsymbol{Z}_i/\|\boldsymbol{Z}_i\|$, $i = 1,\dots,n$. Suppose $0 < \sigma^2 \le 1$. Under the alternative hypothesis, with probability at least $1 - \delta_n$, the graph contains a clique of size m whenever*

$$\binom{m}{2} \le \delta_n \exp\left(\frac{d\sigma^4}{10}\right).$$

The proof of the lemma is given in Appendix D.

Thus, under the alternative hypothesis, if d and σ are such that $d\sigma^4 \ge 10 \ln\left(\binom{m}{2}/\delta_n\right)$, then, with probability at least $1 - \delta_n$, there exists a clique of size m. On the other hand, Theorem 8 implies that under the null hypothesis, the largest clique is not larger than $3\log_2 n$ as long as d is at least a constant multiple of $\log^3 n$. To summarize, we may combine Theorem 8 with Lemma 10 to obtain the following consistency result for the test based on the clique number of the random geometric graph.

Corollary 11. *Consider the hypothesis testing problem described above and the test that accepts the null hypothesis if and only if the clique number of the random geometric graph $\overline{G}(n,d,1/2)$ defined by the points $\boldsymbol{X}_i = \boldsymbol{Z}_i/\|\boldsymbol{Z}_i\|$ is at most $3\log_2 n$. There exists a constant C and a sequence $\epsilon_n \to 0$ such that if*

$$d \ge C \max\left(\frac{\ln m}{\sigma^4}, \log_2^3 n\right) \quad \text{and} \quad m > 3\log_2 n$$

then the probability that the test makes a mistake is less than ϵ_n under both the null and alternative hypotheses.

We do not claim that the proposed test is optimal in any way. Its main drawback is that computing, or even approximating, the clique number of a graph is a notoriously difficult problem and there is no hope of applying the test unless n is quite small. It is a non-trivial challenge to design computationally efficient, yet powerful tests for the hypothesis testing problem described in this section. To understand why this is a difficult problem, one may compare it to the closely related problem of finding a hidden clique in a (Erdős-Rényi) random graph $G(n,1/2)$ – see Alon, Krivelevich, and Sudakov [1].

We also note here that the hypothesis testing problem considered here is a special case of a general class of signal detection problems where noisy observations of linearly transformed signal vectors are available and the goal is to determine the dimension of the signal vectors (e.g., to decide if there is some signal at all). These problems are usually solved via spectral analysis of some whitened covariance matrix of the observations, resulting in asymptotically optimal tests, see, for example, [14] and the references therein.

A Proof of asymptotic equivalence with $G(n,p)$

Here we prove Theorem 2 based on a multivariate central limit theorem. In particular, we use the following multivariate Berry-Esséen inequality (see Bentkus [4], Raič [17]) for convex sets.

Proposition 12. *(Raič [17].) Let $\boldsymbol{Y}_1,\dots,\boldsymbol{Y}_d$ be independent zero-mean random variables, taking values in $\mathbb{R}^m$ such that the covariance matrix of $\sum_{t=1}^d \boldsymbol{Y}_t$ is the identity matrix. Then for any convex set $A\subset\mathbb{R}^m$,*

$$\left|\mathbb{P}\left\{\sum_{t=1}^d \boldsymbol{Y}_t\in A\right\}-\gamma(A)\right|\le\left(47+42m^{1/4}\right)\sum_{t=1}^d\mathbb{E}\|\boldsymbol{Y}_t\|^3\,,$$

where γ is the standard Gaussian measure on $\mathbb{R}^m$.

Proof of Theorem 2. Since the total variation distance equals half of the L_1 distance

$$\sum_g\left|\mathbb{P}\{\overline{G}(n,d,p)=g\}-\mathbb{P}\{G(n,p)=g\}\right|$$

(where the summation is over all graphs g on n vertices) and because n and p are fixed, it suffices to show that for any fixed graph g with vertex set $\{1,\dots,n\}$,

$$\lim_{d\to\infty}\mathbb{P}\{\overline{G}(n,d,p)=g\}=\mathbb{P}\{G(n,p)=g\}\,.$$

In order to show this, we may define $G(n,p)$ as a function of $\binom{n}{2}$ independent standard normal random variables $N_{(i,j)}$ for all i,j with $1\le i<j\le n$. If one connects vertices i and j if and only if $N_{(i,j)}>\Phi^{-1}(1-p)$ then the obtained graph is just an Erdős-Rényi random graph $G(n,p)$. Now for a fixed graph g, let $g_{(i,j)}=1$ if edge (i,j) is present in g and let $g_{(i,j)}=0$ otherwise.

Introduce the $m=\binom{n}{2}$-dimensional vector $\boldsymbol{Y}_t$ whose components are $Y_{(i,j),t}=\sqrt{d}X_{i,t}X_{j,t}$ for all $1\le i<j\le n$ and $1\le t\le d$. Then we have

$$\mathbb{P}\{\overline{G}(n,d,p)=g\}=\mathbb{P}\left\{\bigcap_{1\le i<j\le n}\left\{\mathbb{I}_{\{\sum_{y=1}^d Y_{(i,j),t}>\sqrt{d}t_{p,d}\}}=g_{(i,j)}\right\}\right\}$$

and

$$\mathbb{P}\{G(n,p)=g\}=\mathbb{P}\left\{\bigcap_{1\le i<j\le n}\left\{\mathbb{I}_{\{N_{(i,j)}>\Phi^{-1}(1-p)\}}=g_{(i,j)}\right\}\right\}\,.$$

Denoting by $\widehat{p}_d=1-\Phi(\sqrt{d}t_{p,d})$, it follows from (2) that $\lim_{d\to\infty}\widehat{p}_d=p$ and therefore

$$\lim_{d\to\infty}\mathbb{P}\{G(n,\widehat{p}_d)=g\}=\mathbb{P}\{G(n,p)=g\}$$

so it suffices to prove that $|\mathbb{P}\{\overline{G}(n,d,p)=g\}-\mathbb{P}\{G(n,\widehat{p}_d)=g\}|\to 0$. But

$$\mathbb{P}\{G(n,\widehat{p}_d)=g\}=\mathbb{P}\left\{\bigcap_{1\le i<j\le n}\left\{\mathbb{I}_{\{N_{(i,j)}>\sqrt{d}t_{p,d}\}}=g_{(i,j)}\right\}\right\}$$

and the two probabilities may be compared with the help of Proposition 12.

Clearly, the random variables $Y_{(i,j),t}$ have zero mean, and $Y_{(i,j),t}$ and $Y_{(k,l),t}$ are uncorrelated if $(i,j) \neq (k,l)$ and $\mathbb{E}Y_{(i,j),t}^2 = 1/d$. Therefore, the $m = \binom{n}{2}$-dimensional random vector $\sum_{t=1}^d \boldsymbol{Y}_t$ whose components are $\sum_{t=1}^d Y_{(i,j),t}$ satisfies the assumptions of Proposition 12. Taking A as the m-dimensional rectangle

$$A = \left\{ \boldsymbol{y} \in \mathbb{R}^m : \prod_{(i,j):g_{(i,j)}=1} (\sqrt{d}\, t_{p,d}, \infty) \times \prod_{(i,j):g_{(i,j)}=0} (-\infty, \sqrt{d}\, t_{p,d}) \right\} ,$$

we have

$$\mathbb{P}\{\overline{G}(n,d,p) = g\} = \mathbb{P}\left\{ \sum_{t=1}^d \boldsymbol{Y}_t \in A \right\}$$

and

$$\mathbb{P}\{G(n,\widehat{p}_d) = g\} = \gamma(A)$$

and therefore Proposition 12 implies that

$$\left| \mathbb{P}\{\overline{G}(n,d,p) = g\} - \mathbb{P}\{G(n,hp_d) = g\} \right| \leq (47 + 42\sqrt{n/2}) \sum_{t=1}^d \mathbb{E}\|\boldsymbol{Y}_t\|^3 .$$

Now

$$\begin{aligned}
\mathbb{E}\|\boldsymbol{Y}_t\|^3 &= \mathbb{E}\left(\sum_{1\leq i<j\leq n} Y_{(i,j),t}^2 \right)^{3/2} \\
&= \frac{1}{d^{3/2}} \mathbb{E}\left(\sum_{1\leq i<j\leq n} (X_{i,t}\sqrt{d})^2 (X_{j,t}\sqrt{d})^2 \right)^{3/2} \\
&\leq \frac{1}{d^{3/2}} \left(\mathbb{E}\left(\sum_{1\leq i<j\leq n} (X_{i,t}\sqrt{d})^2 (X_{j,t}\sqrt{d})^2 \right)^2 \right)^{3/4} \\
&= \frac{1}{d^{3/2}} \left(\sum_{1\leq i<j\leq n, 1\leq i'<j'\leq n} \mathbb{E}\left[(X_{i,t}\sqrt{d})^2 (X_{j,t}\sqrt{d})^2 (X_{i',t}\sqrt{d})^2 (X_{j',t}\sqrt{d})^2 \right] \right)^{3/4} .
\end{aligned}$$

By the joint asymptotic normality of $\sqrt{d}(X_{i,t}, X_{j,t}, X_{i',t}, X_{j',t})$,

$$\begin{aligned}
\lim_{d\to\infty} \sum_{1\leq i<j\leq n, 1\leq i'<j'\leq n} & \mathbb{E}\left[(X_{i,t}\sqrt{d})^2 (X_{j,t}\sqrt{d})^2 (X_{i',t}\sqrt{d})^2 (X_{j',t}\sqrt{d})^2 \right] \\
&= \sum_{1\leq i<j\leq n, 1\leq i'<j'\leq n} \mathbb{E}\left[N_{i,t}^2 N_{j,t}^2 N_{i',t}^2 N_{j',t}^2 \right] ,
\end{aligned}$$

where $N_1, \ldots, N_n$ are i.i.d. standard normal random variables. But

$$\mathbb{E} \sum_{1 \le i < j \le n, 1 \le i' < j' \le n} N_{i,t}^2 N_{j,t}^2 N_{i',t}^2 N_{j',t}^2 \le 9 \binom{n}{2}^2 ,$$

where we used that the fourth moment of a standard normal random variable equals 3. Summarizing, we have that for any fixed graph g,

$$\limsup_{d \to \infty} \sqrt{d} \left| \mathbb{P}\{\overline{G}(n,d,p) = g\} - \mathbb{P}\{G(n,\widehat{p}_d) = g\} \right| \le 146 n^3 + 92 n^{7/2} .$$

which concludes the proof. □

B Proof of Theorem 5

In the proof we use the following classical result of Jung [12] (see also Danzer, Grünbaum, and Klee [8, Theorem 2.6]):

Proposition 13. (JUNG'S THEOREM.) *For every set $A \subset \mathbb{R}^d$ of diameter at most 1 there exists a closed ball of radius $\sqrt{d/(2(d+1))}$ that includes A.*

Proof of Theorem 5. To prove the upper bound, notice that the vectors $\boldsymbol{X}_i$ corresponding to any clique K in $\overline{G}(n,d,p)$ form a set of diameter at most $\sqrt{2(1-t_{p,d})}$. Therefore, Jung's theorem implies that K is contained in a ball of radius $\sqrt{\frac{d}{d+1}(1-t_{p,d})}$. Furthermore, since the points lie on the surface of the unit ball, K is contained in a spherical cap $C_{d-1}(\boldsymbol{u}, s_p)$ for some $\boldsymbol{u} \in S_{d-1}$ and $s_p = \sqrt{(d\,t_{p,d}+1)/(d+1)}$. Therefore, if $\mathscr{C}_{d-1}(t) = \{C_{d-1}(\boldsymbol{u}, t) : \boldsymbol{u} \in S_{d-1}\}$ denotes the set of all spherical caps of the unit ball with height $1-t$ and $\mu_{d-1,n}$ denotes the empirical measure defined by the vectors $\boldsymbol{X}_i$, that is, for any set $A \subset \mathbb{R}^d$ we define

$$\mu_{d-1,n}(A) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{I}_{\{\boldsymbol{X}_i \in A\}} ,$$

then the clique number of a random geometric graph can be upper bounded as

$$\omega(d,n,p) \le n \sup_{C \in \mathscr{C}_{d-1}(s_p)} \mu_{d-1,n}(C) .$$

The expected number of points $\mathbb{E}\{\mu_{d-1,n}(C)\}$ falling in any set $C \in \mathscr{C}_{d-1}(s_p)$ is $n\mu_{d-1}(C)$ where μ_{d-1} denotes the uniform distribution on S_{d-1}. An inequality of Vapnik-Chervonenkis [18] states that

$$\mathbb{P}\left\{ \exists C \in \mathscr{C}_{d-1}(s_p) : \mu_{d-1,n}(C) - \mu_{d-1}(C) > \varepsilon \sqrt{\mu_{d-1,n}(C)} \right\} \le 4 \left(\frac{2ne}{V} \right)^V e^{-n\varepsilon^2/4} .$$

Here V is the VC-dimension of the class of all closed balls in $\mathbb{R}^d$ which is well-known to equal $V = d+1$ (Dudley [9]). Therefore, for any $0 < \eta < 1$ we have that, with probability at least $1-\eta$, for all $C \in \mathscr{C}_{d-1}(s_p)$,

$$\mu_{d-1,n}(C) \le \mu_{d-1}(C) + \sqrt{\mu_{d-1,n}(C) \left(\frac{4(d+1)}{n} \ln \frac{2ne}{d+1} + \frac{4}{n} \ln \frac{4}{\eta} \right)} .$$

Solving this inequality for $\mu_{d-1,n}$ we obtain (from $a \le b + \sqrt{ac}$ we get $a \le b + c/2 + \sqrt{c^2 + 4bc}/2 \le 2b + c$)

$$\mu_{d-1,n}(C) \le 2\mu_{d-1}(C) + \frac{4(d+1)}{n} \ln \frac{2ne}{d+1} + \frac{4}{n} \ln \frac{4}{\eta} .$$

Using $1 + x \le e^x$ and the upper bound of (1), we have that, with probability at least $1 - \eta$,

$$\sup_{C \in \mathscr{C}_{d-1}(s_p)} \mu_{d-1,n}(C) \le \frac{1}{s_p \sqrt{d}} e^{-s_p^2 (d-1)/2} + \frac{4(d+1)}{n} \ln \frac{2ne}{d+1} + \frac{4}{n} \ln \frac{4}{\eta} .$$

This implies the theorem. □

C Proof of Lemma 1

Proof of Lemma 1. We show that for any $0 < \gamma < 1/2$ and $d \ge \max\left\{(4\ln(2/p))^{1/(1-2\gamma)}, 2^{1/\gamma}\right\}$,

$$-\frac{3}{2} d^{-\gamma} \Phi^{-1}(1-p) - \sqrt{2\pi} e^{-\frac{d^{1-2\gamma} - 2C_l}{4}} \le t_{p,d}\sqrt{d} - \Phi^{-1}(1-p) \le 2d^{-\gamma}\Phi^{-1}(1-p) + 2\sqrt{2\pi} e^{-\frac{d^{1-2\gamma} - 2C_u}{4}} , \tag{27}$$

where $C_l = \max\left\{\left(\Phi^{-1}(1-p)\right)^2, \left(\Phi^{-1}(1-\frac{3}{2}p)\right)^2\right\}$ and $C_u = (\Phi^{-1}(1-p/2))^2$. The statement of the lemma then follows easily by setting $\gamma = \frac{1}{2}(1 - \log_d(2\ln d))$ and noting that $\left|\Phi^{-1}(1-p/2)| = \max\left\{\left|\Phi^{-1}(1-p/2)\right|, |\Phi^{-1}(1-p)\right|, \left|\Phi^{-1}(1-\frac{3}{2}p)\right|\right\}$ for $0 < p \le 1/2$ and that $\sqrt{d/(2\ln d)} > 2$ if $d \ge 27$.

To show (27), we use the same technique as in Theorems 6 and 7. First we prove the upper bound. Let $\boldsymbol{Z}$ be a d-dimensional standard normal random variable and let $\boldsymbol{e} = (1, 0, \ldots, 0)$ denote the first standard basis vector. Then

$$p = \mu_{d-1}(C_{d-1}(\boldsymbol{e}, t_{p,d})) = \mathbb{P}\left\{\left(\frac{\boldsymbol{Z}}{\|\boldsymbol{Z}\|}, \boldsymbol{e}\right) \ge t_{p,d}\right\} .$$

Now for any $0 < \alpha < 1$, we have, similarly to (7),

$$p \le 1 - \Phi\left(\alpha t_{p,d}\sqrt{d}\right) + e^{-\frac{(1-\alpha)^2 d}{4}} .$$

From here, letting $\alpha = 1 - d^{-\gamma}$, we obtain

$$\begin{aligned} t_{p,d}\sqrt{d} &\le \frac{\Phi^{-1}\left(1 - p + e^{-\frac{(1-\alpha)^2 d}{4}}\right)}{\alpha} = \frac{\Phi^{-1}\left(1 - p + e^{-d^{1-2\gamma}/4}\right)}{1 - d^{-\gamma}} \\ &\le \frac{\Phi^{-1}(1-p) + \sqrt{2\pi} e^{-d^{1-2\gamma}/4} e^{C_u/2}}{1 - d^{-\gamma}} \\ &\le \Phi^{-1}(1-p) + 2d^{-\gamma}\Phi^{-1}(1-p) + 2\sqrt{2\pi} e^{-\frac{d^{1-2\gamma} - 2C_u}{4}} , \end{aligned}$$

where the second inequality holds since $\Phi^{-1}(x+y) \le \Phi^{-1}(x) + y\sqrt{2\pi} e^{(\Phi^{-1}(x+y))^2}$ for any $x > 1/2, y > 0$, as our assumptions imply $p/2 < e^{-d^{1-2\gamma}/4}$, while the last inequality follows since our assumptions on d imply $1/(1 - d^{-\gamma}) \le 1 + 2d^{-\gamma} \le 2$. This proves the upper bound of the lemma.

The proof of the lower bound is similar. As in (21), we can prove for any $\alpha > 1$,

$$p \geq 1 - \Phi(\alpha t_{p,d}\sqrt{d} - e^{-\frac{\alpha^2 - \sqrt{2\alpha^2-1}}{2}d} .$$

Letting $\alpha = 1 + d^{-\gamma} + d^{-2\gamma}/2$, a similar reasoning as above yields

$$\begin{aligned} t_{p,d}\sqrt{d} &\geq \frac{\Phi^{-1}\left(1 - p - e^{-d^{1-2\gamma}/4}\right)}{1 + d^{-\gamma} + d^{-2\gamma}/2} \\ &\geq \frac{\Phi^{-1}(1-p) - \sqrt{2\pi}e^{-d^{1-2\gamma}/4}e^{C_l/2}}{1 + d^{-\gamma} + d^{-2\gamma}/2} \\ &\geq \Phi^{-1}(1-p) - \frac{3}{2}d^{-\gamma}\Phi^{-1}(1-p) - \sqrt{2\pi}e^{-\frac{d^{1-2\gamma}-2C_l}{4}} , \end{aligned}$$

where in the last step we used $1/(1 + d^{-\gamma} + d^{-2\gamma}/2) \geq 1 - d^{-\gamma} - d^{-2\gamma}/2 \geq 1 - 3d^{-\gamma}/2$. □

D Proof of Lemma 10

Proof of Lemma 10. It suffices to show that, if i and j both belong to S then

$$\mathbb{P}\{(\boldsymbol{X}_i, \boldsymbol{X}_j) < 0\} \leq e^{-d\sigma^4/10} . \tag{28}$$

The lemma then follows by the union bound applied for the $\binom{m}{2}$ pairs or vertices of S. Since

$$\begin{aligned} \mathbb{P}\{(\boldsymbol{X}_i, \boldsymbol{X}_j) < 0\} &= \mathbb{P}\left\{\frac{1}{d}\sum_{t=1}^{d}(N_{i,t} + Y_t)(N_{j,t} + Y_t) < 0\right\} \\ &= \mathbb{P}\left\{\frac{1}{d}\sum_{t=1}^{d}\Big((N_{i,t} + Y_t)(N_{j,t} + Y_t) - \mathbb{E}(N_{i,t} + Y_t)(N_{j,t} + Y_t)\Big) < -\sigma^2\right\} , \end{aligned}$$

the problem boils down to finding appropriate left-tail bounds for independent sums of products of correlated normal random variables. To this end, we use the well-know fact (which is easy to obtain by direct calculation) that if ξ and ζ are jointly normal zero-mean random variables with variances s_ξ^2 and s_ζ^2, respectively, and correlation $r = \mathbb{E}[\xi\zeta]/(s_\xi s_\zeta)$ then the cumulant generating function of their product equals

$$\ln \mathbb{E}\left[\exp(\lambda\xi\zeta)\right] = \frac{1}{2}\ln\frac{1 - r^2}{1 - \Big(r + (1-r^2)s_\xi s_\zeta\lambda\Big)^2}$$

for all λ such that $\Big|r + (1-r^2)s_\xi s_\zeta\lambda\Big| < 1$.
Writing $\rho = \sigma^2/(1+\sigma^2)$ for the correlation of $N_{i,t} + Y_t$ and $N_{j,t} + Y_t$, this implies that

$$\begin{aligned} F(\lambda) &\stackrel{\text{def}}{=} \ln \mathbb{E}\left[\exp(\lambda(N_{i,t} + Y_t)(N_{j,t} + Y_t))\right] \\ &= \frac{1}{2}\ln\frac{1 - \rho^2}{1 - \big(\rho + (1+\rho)\lambda\big)^2} \end{aligned}$$

for all λ such that $|\rho+(1+\rho)\lambda|<1$. Since we are interested in lower tail probabilities, we consider negative values of λ. Then $F(\lambda)$ is well defined for $\lambda\in(-1,0]$. By Taylor's theorem, for every such λ there exists $y\in(\lambda,0)$ such that

$$F(\lambda)=F(0)+\lambda F'(0)+\frac{\lambda^2}{2}F''(y)\,.$$

By straightforward calculation, $F(0)=0$, $F'(0)=\sigma^2$, and

$$F''(y)=(1+\rho)^2\frac{1+\big(\rho+(1+\rho)y\big)^2}{\Big(1-\big(\rho+(1+\rho)y\big)^2\Big)^2}$$

which is monotone increasing for $y\in(-\rho/(1+\rho),0]$ and therefore

$$F(\lambda)\le\lambda\sigma^2+\frac{\lambda^2}{2}F''(0)=\lambda\sigma^2+\frac{\lambda^2}{2}\frac{1+\rho^2}{(1-\rho)^2}\quad\text{for all }\lambda\in(-\rho/(1+\rho),0].$$

Thus, by Chernoff's bounding method (see Chernoff [7]), for all $\lambda\in(-\rho/(1+\rho),0]$,

$$\mathbb{P}\{(\boldsymbol{X}_i,\boldsymbol{X}_j)<0\}\le\exp(dF(\lambda))\le\exp\left(d\lambda\sigma^2+\frac{d\lambda^2}{2}\frac{1+\rho^2}{(1-\rho)^2}\right)\,.$$

The upper bound is minimized for $\lambda=-\sigma^2(1-\rho)^2/(1+\rho^2)$ which is a legal choice since $\sigma^2(1-\rho)^2/(1+\rho^2)<\rho/(1+\rho)$. The upper bound becomes

$$\mathbb{P}\{(\boldsymbol{X}_i,\boldsymbol{X}_j)<0\}\le\exp\left(-\frac{d\sigma^4(1-\rho)^2}{2(1+\rho^2)}\right)\,.$$

Since $\sigma^2\le1$, we have $\rho\le1/2$ and we obtain (28). □

Acknowledgments. We thank a referee for pointing out an embarrassing mistake in the original manuscript.

References

[1] N. Alon, M. Krivelevich, and B. Sudakov. Finding a large hidden clique in a random graph. *Random Structures and Algorithms*, 13:457–466, 1999. MR1662795

[2] N. Alon and J.H. Spencer. *The Probabilistic Method*. Wiley, New York, 1992. MR1140703

[3] C. Ané, S. Blachère, D. Chafaï, P. Fougères, I. Gentil, F. Malrieu, C. Roberto, and G. Scheffer. *Sur les inégalités de Sobolev logarithmiques*, volume 10 of *Panoramas et Synthèses*. Société Mathématique de France, Paris, 2000. MR1845806

[4] V. Bentkus. On the dependence of the Berry–Esséen bound on dimension. *Journal of Statistical Planning and Inference*, pages 385–402, 2003. MR1965117

[5] B. Bollobás. *Random Graphs*. Academic Press, London, 1985. MR0809996

[6] A. Brieden, P. Gritzmann, R. Kannan, V. Klee, L. Lovász, and M. Simonovits. Deterministic and randomized polynomial-time approximation of radii. *Mathematika. A Journal of Pure and Applied Mathematics*, 48(1-2):63–105, 2001. MR1996363

[7] H. Chernoff. A measure of asymptotic efficiency for tests of a hypothesis based on the sum of observations. *Annals of Mathematical Statistics*, 23:493–507, 1952. MR0057518

[8] L. Danzer, B. Grünbaum, and V. Klee. Helly's theorem and its relatives. In *Proc. Sympos. Pure Math., Vol. VII*, pages 101–180. Amer. Math. Soc., Providence, R.I., 1963. MR0157289

[9] R.M. Dudley. Central limit theorems for empirical measures. *Annals of Probability*, 6:899–929, 1979. Correction in 7:909–911, 1979. MR0512411

[10] P. Erdős and A. Rényi. On the evolution of random graphs. *Publications of the Mathematical Institute of the Hungarian Academy of Sciences, Ser. A*, 5:17–61, 1960. MR0125031

[11] S. Janson, T. Łuczak, and A. Ruciński. *Random Graphs*. John Wiley, New York, 2000. MR1782847

[12] H. Jung. Über die kleinste Kugel, die eine räumliche Figur einschliesst. *J. Reine Angew. Math.*, 123:241–257, 1901.

[13] P. Massart. *Concentration inequalities and model selection*. Ecole d'été de Probabilités de Saint-Flour 2003. Lecture Notes in Mathematics. Springer, 2006. MR2319879

[14] R.R. Nadakuditi and J.W. Silverstein. Fundamental limit of sample generalized eigenvalue based detection of signals in noise using relatively few signal-bearing and noise-only samples. Technical report, arXiv:0902.4250v1, 2009.

[15] E.M. Palmer. *Graphical Evolution*. John Wiley & Sons, New York, 1985. MR0795795

[16] M. Penrose. *Random Geometric Graphs*, volume 5 of *Oxford Studies in Probability*. Oxford University Press, Oxford, 2003. MR1986198

[17] M. Raič. *Normalna aproksimacija po Steinovi metodi*. PhD thesis, Univerza v Ljubljani, 2009.

[18] V.N. Vapnik and A.Ya. Chervonenkis. *Theory of Pattern Recognition*. Nauka, Moscow, 1974. (in Russian); German translation: *Theorie der Zeichenerkennung*, Akademie Verlag, Berlin, 1979. MR0474638

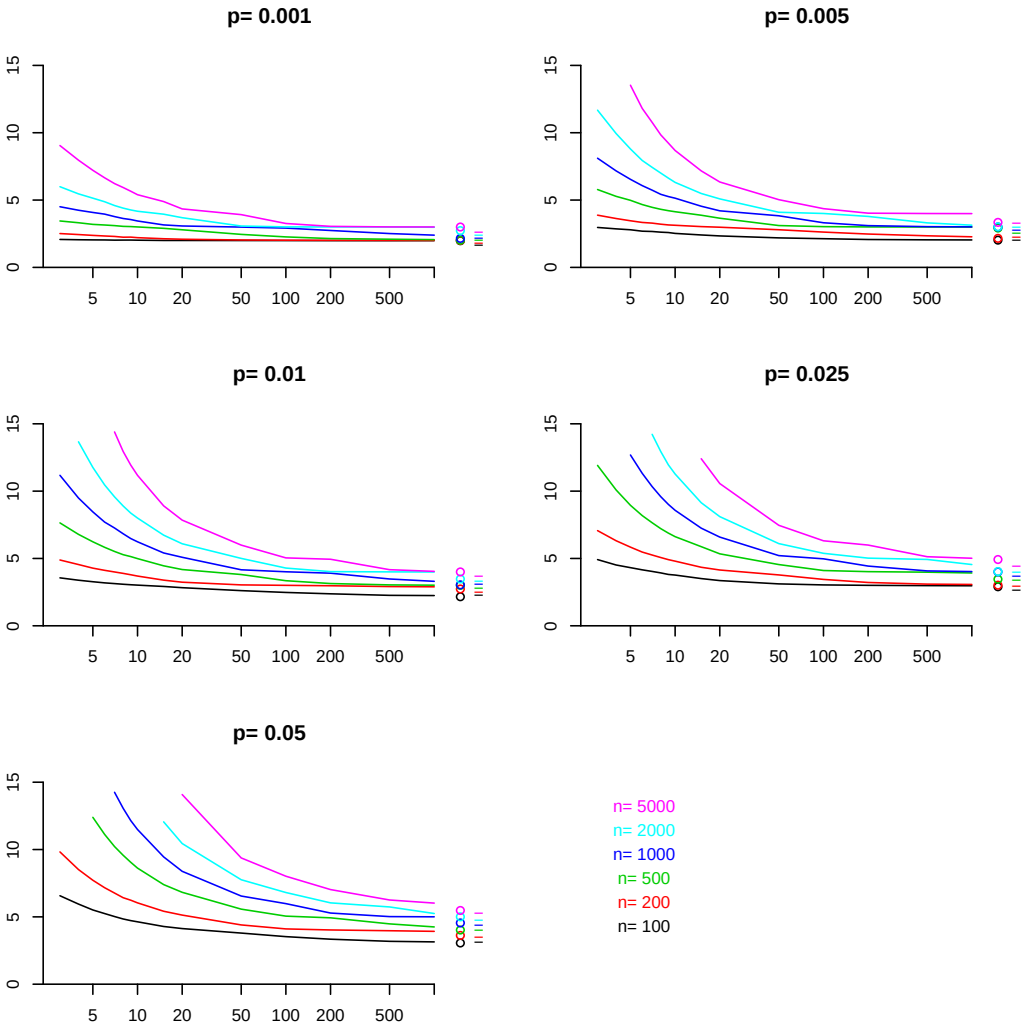

Figure 2: $\mathbb{E}\omega(n,d,p)$ as a function of d for various values of the parameters.

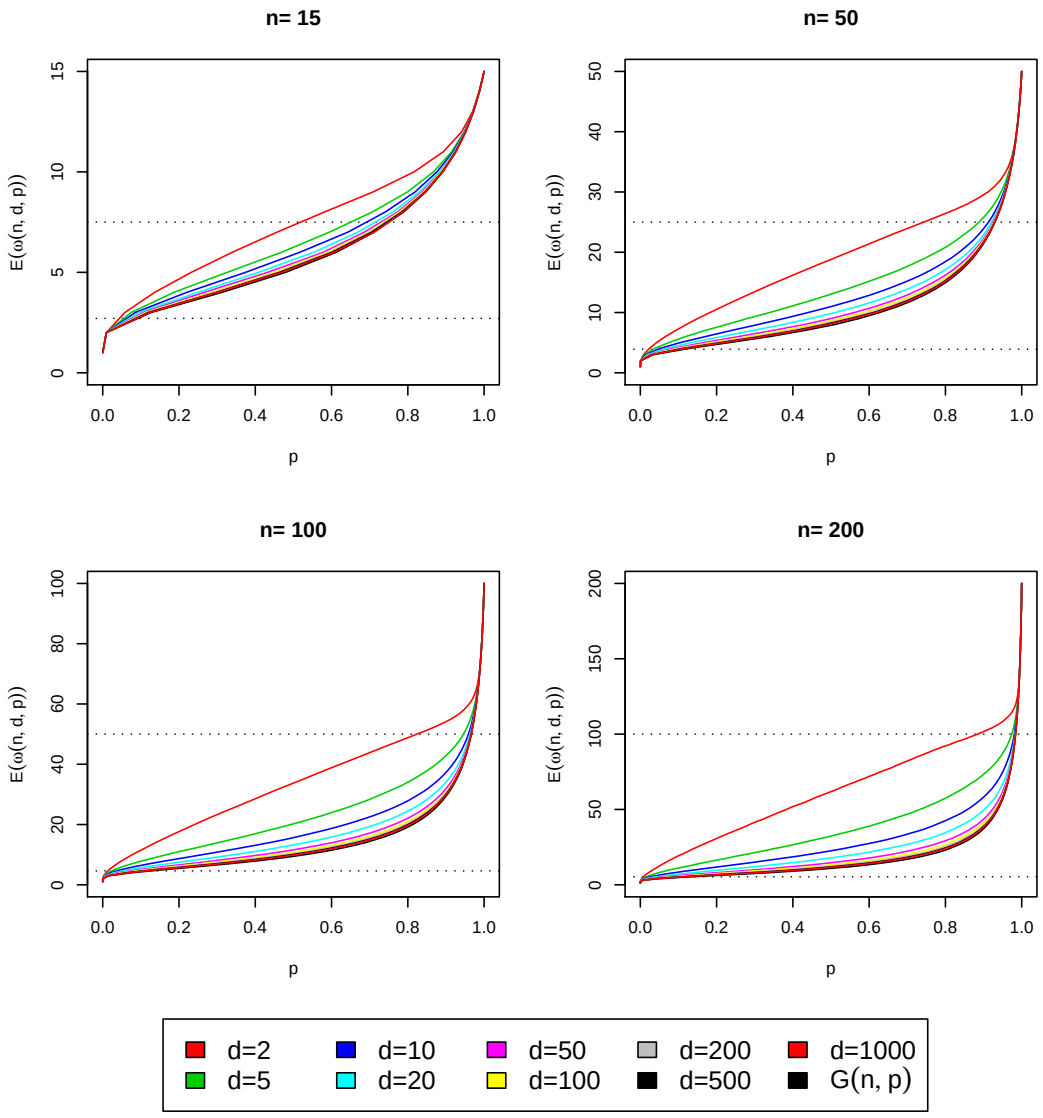

Figure 3: $\mathbb{E}\omega(n,d,p)$ as a function of p for some small values of n.