





**Random geometric graphs in high dimension**Vittorio Erba <sup>1,2,\*</sup>, Sebastiano Ariosto <sup>1</sup>, Marco Gherardi <sup>1,2</sup> and Pietro Rotondo <sup>2</sup><sup>1</sup>*Dipartimento di Fisica dell'Università di Milano, via Celoria 16, 20100 Milano, Italy*<sup>2</sup>*INFN, sezione di Milano, via Celoria 16, 20100 Milano, Italy*

(Received 6 March 2020; revised 10 May 2020; accepted 23 June 2020; published 14 July 2020)

Many machine learning algorithms used for dimensional reduction and manifold learning leverage on the computation of the nearest neighbors to each point of a data set to perform their tasks. These proximity relations define a so-called geometric graph, where two nodes are linked if they are sufficiently close to each other. Random geometric graphs, where the positions of nodes are randomly generated in a subset of  $\mathbb{R}^d$ , offer a null model to study typical properties of data sets and of machine learning algorithms. Up to now, most of the literature focused on the characterization of low-dimensional random geometric graphs whereas typical data sets of interest in machine learning live in high-dimensional spaces ( $d \gg 10^2$ ). In this work, we consider the infinite dimensions limit of hard and soft random geometric graphs and we show how to compute the average number of subgraphs of given finite size  $k$ , e.g., the average number of  $k$  cliques. This analysis highlights that local observables display different behaviors depending on the chosen ensemble: soft random geometric graphs with continuous activation functions converge to the naive infinite-dimensional limit provided by Erdős-Rényi graphs, whereas hard random geometric graphs can show systematic deviations from it. We present numerical evidence that our analytical results, exact in infinite dimensions, provide a good approximation also for dimension  $d \gtrsim 10$ .

DOI: [10.1103/PhysRevE.102.012306](https://doi.org/10.1103/PhysRevE.102.012306)**I. INTRODUCTION**

Random geometric graphs (RGGs) are networks whose nodes are  $d$ -dimensional randomly generated vectors from some probability distribution over  $\mathbb{R}^d$ , and edges link nodes only if their distance does not exceed a threshold distance  $r$  [1]. As such, their connectivity structure encodes information about the spatial structure of the nodes, and on the space they are embedded in: for this reason they are widely used in modeling complex systems in which geometric constraints play a fundamental role, such as transport [2,3], wireless [4], and social networks [5,6].

Most of the results on RGGs have been established in the low-dimensional regime  $d \leq 3$  [1,2,7–10]. However, the high-dimensional limit  $d \rightarrow \infty$  has recently gathered interest. Indeed in the era of big data and machine learning, typical data sets are made of vectors of hundreds of components (think for instance to the workhorse model in computer vision, the MNIST data set of handwritten digits); understanding how high-dimensional geometry works, and how it affects the proximity structure of data sets is crucial for the correct usage of manifold learning algorithms (from dimensional reduction protocols [11,12] to intrinsic dimension estimators [13]), and for the creation of novel procedures tailored for the high-dimensional regime with benefits for dimensional reduction and clustering algorithms. With this idea in mind, high-dimensional RGGs become a perfect null model for unstructured data, to benchmark and compare against real world data sets [14,15].

On the more mathematical side, it is an open problem to understand whether high-dimensional RGGs converge (as a statistical ensemble) to Erdős-Rényi graphs; rigorous results for RGGs with nodes uniformly distributed on the sphere can be found in Refs. [16–18] and suggest that high- $d$  RGGs are similar to Erdős-Rényi graphs. On the other hand, the clustering coefficient of RGGs with nodes uniformly distributed on the hypercube shows systematic deviations from the ERG prediction [19]. A related but different question is whether the critical behavior of high-dimensional RGGs converges to that of Erdős-Rényi graphs: see Ref. [20] for more information.

In this work, we present a general framework for the computation of the average value of local observables of high-dimensional hard and soft RGGs. To this end, we exploit a multivariate version of the central limit theorem (CLT) to show that the joint probability of rescaled distances between nodes is normal distributed, and we compute and characterize its correlation matrix.

We evaluate the average number of  $M$  cliques, i.e., of fully connected subgraphs with  $M$  vertices, in high-dimensional RGGs. We point out that these local observables show systematic deviations from the ERG prediction in hard RGGs (whenever the hypothesis of the CLT are satisfied), whereas we observe convergence to ERG for soft RGGs with continuous activation functions. This implies that the form of the activation function as well as the probability distribution on the nodes are crucial elements in studying the convergence of RGGs to ERGs.

Finally, we present numerical evidence that our analytical results do not hold only for  $d \rightarrow \infty$ , but provide a good approximation even in finite dimensions as low as  $d \sim 10$ . This suggests that the high-dimensional limit of RGGs could be seen as a zeroth-order term of a series expansion in  $d$ ,

\*vittorio.erba@unimi.it

possibly giving perturbative access to analytical results for low-dimensional RGGs.

In summary, the main results of our paper are the following.

(i) We systematically establish (under hypotheses on node positions resembling those of the CLT) when we should expect deviations from the ERG prediction in the infinite-dimensional limit, by studying the behavior of the average number of cliques for hard and soft RGGs. It is worth remarking that observing this deviation for  $k$  cliques is a strong indicator that most of the other subgraphs will display systematic deviations from the naive ERG prediction as well.

(ii) In the case where the average number of cliques does not converge to the ERG prediction (i.e., for hard RGGs), we provide a quantitative analysis, based on the multivariate CLT, that well reproduces the nontrivial limit behavior of the properties considered.

(iii) We numerically show that the high-dimensional approximation under which we derive our results gives accurate results even in moderately low dimension  $d \sim 10$ .

The paper is organized as follows. In Sec. II we introduce the notation and define the ensembles of RGGs that we will study. In Sec. III we use a multivariate version of the central limit theorem to derive an explicit expression for the joint probability distribution of the distances of  $M$  randomly drawn vectors in the limit of high dimension. This will be the crucial tool to compute averages of observables in high-dimensional RGGs. Finally, in Sec. IV, we present our results on the average number of  $M$  cliques for hard and soft RGGs alongside with numerical simulations.

## II. HARD AND SOFT RANDOM GEOMETRIC GRAPHS

Note on terminology: In the literature, random geometric graphs are those with hard activation function (see later in this section). Here, when omitting the adjectives “hard” or “soft” we generically refer to both.

A random geometric graph is a graph whose nodes are random points in  $\mathbb{R}^d$ , and whose edges are randomly generated based on the mutual distances between the nodes (see Fig. 1). Let us be more precise, starting by nodes. We consider a probability distribution  $\nu$  over  $\mathbb{R}^d$ , and we draw  $N$  i.i.d. samples  $\{\vec{x}_i\}_{i=1}^N$  from  $\nu$ ; these will be the nodes of the random geometric graph. Among the possible choices of  $\nu$ , a very common one is the uniform distribution on the  $d$ -dimensional hypercube  $[0, 1]^d$ , i.e.,

$$\nu^{\text{cube}}(\vec{x}) = \prod_{k=1}^d \theta(x^k) \theta(1 - x^k), \quad (1)$$

where  $\theta$  is the Heaviside theta, and superscripts denote coordinates. We will consider more in general probability distributions  $\nu$  that are factorized and equally distributed over the coordinates, i.e.,

$$\nu(\vec{x}) = \prod_{k=1}^d \tau(x^k), \quad (2)$$

where  $\tau$  is a probability distribution on  $\mathbb{R}$  with finite first and second moments. In this case, the coordinates of all nodes  $\{x_i^k\}$ , with  $1 \leq i \leq N$  and  $1 \leq k \leq d$  are i.i.d. random variables with law  $\tau$ .

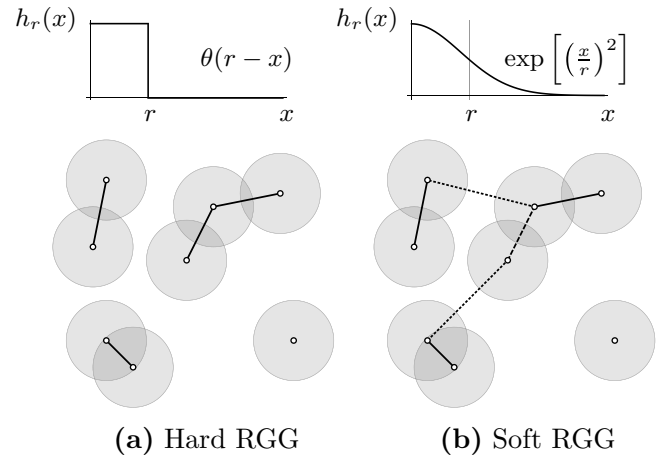


FIG. 1. Example of hard and soft random geometric graphs. Small circles denote nodes embedded in  $\mathbb{R}^2$  drawn randomly with the uniform measure on  $[0, 1]^2$  and shaded circles highlight a region of radius  $r/2$  around nodes. Solid lines highlight the actual edges of the represented graphs. On the top of the graph representations, the activation function used to build them are displayed. (a) In a hard random geometric graph at cutoff  $r$ , the only selected edges are those with nodes closer than  $r$  (in the picture, the nodes whose shaded regions intersect). (b) In a soft random geometric graph, edges are selected based on a continuous activation function  $h_r(x)$ . If two nodes are at distance  $d$  between each other, then the edge that connects them will be chosen with probability  $h_r(d)$ . In the picture, dotted edges are those edges that have been chosen by the soft random geometric graphs even though the distance between nodes was larger than  $r$ . Vice versa, dashed edges are those at distance smaller than  $r$ , but not selected in that specific instance of the soft random geometric graph.

Now, for each pair of nodes  $x, y$  we compute the distance  $d(x, y)$  and we add the link  $e = (x, y)$  to the edge set of the random geometric graph with probability  $h[d(x, y)]$ , where  $h : \mathbb{R}^+ \rightarrow [0, 1]$  is the so-called activation function of the random geometric graph. The activation function describes how likely it is for two nodes to be linked based on their distance, and will typically be a monotone decreasing function, with the idea that closer nodes will be linked with higher probability than further ones; we will consider monotone decreasing activation functions, with  $h(0) = 1$  and  $h(+\infty) = 0$ .

Usually, the activation function is labeled by a parameter  $r \in \mathbb{R}^+$  that describes the typical distance at which a pair of nodes will be considered close enough to be linked with a nontrivial probability, for example  $h_r(r) = \frac{1}{2}$ . In this case, the statistical properties of random geometric graphs can be investigated as functions of  $r$ .

In this work, we will consider two types of activation functions. The first one is that of hard random geometric graphs, i.e.,

$$h_r^{\text{hard}}(x) = \theta(r - x). \quad (3)$$

In this case, all pairs of nodes with distance smaller than  $r$  will be deterministically linked by an edge. The second one is that of soft random geometric graphs (also called random connection models in the literature), i.e., those with  $h_r(x)$  at least continuous in  $x$ . A common choice in the literature (see,

for example, Refs. [4,21]) is to employ the so-called Reyleigh fading activation functions, i.e.,

$$h_r^{\text{rayleigh}}(x) = \exp\left[-\xi\left(\frac{x}{r}\right)^\eta\right], \quad (4)$$

where  $\xi = \log(2)$  guarantees that  $h_r(r) = \frac{1}{2}$ .

The last ingredient to be discussed is the distance function  $d(x, y)$ . We will consider the  $p$  norms  $\mathbb{R}^d$

$$\|\vec{x}\|_p = \sqrt[p]{\sum_{i=1}^d |x_i|^p}. \quad (5)$$

Notice that  $p$  norms are norms only for  $p \geq 1$ , as for  $0 < p < 1$  the triangle inequality is not satisfied. In this case, one can show that  $\|\vec{x} - \vec{y}\|_p^p$  defines nonetheless a distance. Thus, we will define and consider the distances

$$d_p(\vec{x}, \vec{y}) = \|\vec{x} - \vec{y}\|_p^{\min(1,p)}. \quad (6)$$

### III. A CENTRAL LIMIT THEOREM FOR DISTANCES IN HIGH DIMENSION

As a first step in our analysis, we are interested in computing the high-dimensional limit of the joint probability distribution of the distances between  $M$  random points  $\{\vec{x}_i\}_{i=1}^M \subset \mathbb{R}^d$ , drawn independently from the factorized distribution  $\nu$  in Eq. (2):

$$\begin{aligned} & \Pi(d_{(1,2)}, d_{(1,3)}, \dots, d_{(M-1,M)}) \\ &= \int \prod_{i=1}^M \nu(\vec{x}_i) dx_i \prod_{1 \leq i < j \leq M} \delta(d_p(\vec{x}_i, \vec{x}_j) - d_{(i,j)}). \end{aligned} \quad (7)$$

Since the distance  $d_p(\vec{x}, \vec{y})$  between two vectors  $\vec{x}, \vec{y}$  is a function of the sum of  $d$  i.i.d. random variables, we expect that for  $d \rightarrow \infty$  it converges to its average value  $d\mu$  by the law of large numbers. Correspondingly, let us define the rescaled variables

$$\begin{aligned} q_{(i,j)} &= \frac{[d_p(\vec{x}_i, \vec{x}_j)]^{\max(1,p)} - d\mu}{\sqrt{d}} \\ &= \frac{1}{\sqrt{d}} \sum_{k=1}^d (|x_i^k - x_j^k|^p - \mu) \\ &= \frac{1}{\sqrt{d}} \sum_{k=1}^d q_{(i,j)}^k, \end{aligned} \quad (8)$$

where

$$\mu = \int dx dy \tau(x) \tau(y) |x - y|^p. \quad (9)$$

Notice that the random vectors  $\mathbf{q}_k = (q_{(1,2)}^k, q_{(1,3)}^k, \dots, q_{(M-1,M)}^k) \in \mathbb{R}^{\binom{M}{2}}$ , with  $1 \leq k \leq d$ , are statistically independent and identically distributed, and that by definition the expected value of  $\mathbf{q}_k$  is the null vector. Notice also that the components of the vectors  $\mathbf{q}_k$  are naturally indexed by lexicographically ordered multi-indices, as they are related to the distances between pairs of points along the  $k$ th dimension; to distinguish such vectors from the Euclidean ones, we type them in boldface.

The vector  $\mathbf{q} = (q_{(1,2)}, q_{(1,3)}, \dots, q_{(M-1,M)})$  is a sum of i.i.d. multivariate random variables, and satisfies the following central limit theorem.

*Theorem 1 (Multivariate central limit theorem).* Let  $\mathbf{q}^1, \mathbf{q}^2, \dots, \mathbf{q}^d$  be i.i.d. random vectors in  $\mathbb{R}^{\binom{M}{2}}$  with null mean and covariance matrix  $\Sigma_{(i,j),(k,l)} = \mathbb{E}[q_{(i,j)}^1 q_{(k,l)}^1]$ . Then

$$\mathbf{q} = \frac{1}{\sqrt{d}} \sum_{k=1}^d \mathbf{q}^k \quad (10)$$

is Gaussian distributed with null mean and covariance  $\Sigma$  in the limit  $d \rightarrow \infty$ .

The general formal proof can be found in Ref. [22] (see Proposition 2.17). A ‘‘physicist’’ approach to the proof would be to compute the characteristic function of  $\mathbf{q}$  and to expand it to the leading order for large  $d$ . It is worth noticing that the first neglected term in the expansion is of relative order  $1/\sqrt{d}$ , and may depend on  $M$ . Thus, this  $d \rightarrow \infty$  limit is to be intended at fixed  $M$ , and the result can be used either to treat generic observables for graphs where the total number of nodes is fixed, or to treat observables that depend only on a finite number of nodes at a time in graphs where the total number of nodes may scale with  $d$ .

The CLT presented above holds for the variable  $\mathbf{q}$ , and not for the actual distances. However, this is not an issue as the joint distribution for distances can be derived by a simple coordinate change, factorized over each direction. Moreover, as we will see in the following, it is often easy to obtain the observables of interest in terms of the  $\mathbf{q}$  variable.

We now focus on the explicit form of the covariance matrix  $\Sigma$  (notice that, as the vectors  $\mathbf{q}_k$ , the covariance matrix is indexed by multi-indices). By definition, one has

$$\Sigma_{(i,j),(k,l)} = \mathbb{E}[(|y_i - y_j|^p - \mu)(|y_k - y_l|^p - \mu)], \quad (11)$$

where  $y_i, y_j, y_k, y_l$  are all i.i.d. random variables with distribution  $\tau$ , and  $1 \leq i < j \leq M$ ,  $1 \leq k < l \leq M$ . By permutational symmetry, only three different cases are possible.

(i) Diagonal correlations ( $i = k$  and  $j = l$ )

$$\alpha = \Sigma_{(i,j),(i,j)} = \int dx dy \tau(x) \tau(y) |x - y|^{2p} - \mu^2; \quad (12)$$

(ii) Triangular correlations ( $i = k$  and  $j \neq l$  or  $i \neq k$  and  $j = l$ )

$$\begin{aligned} \beta &= \Sigma_{(i,j),(i,k)} = \Sigma_{(i,j),(k,j)} \\ &= \int dx dy dz \tau(x) \tau(y) \tau(z) |x - y|^p |x - z|^p - \mu^2; \end{aligned} \quad (13)$$

(iii) Pair-pair correlations ( $i, j, k, l$  are all distinct)

$$\gamma = \Sigma_{(i,j),(k,l)} = \left( \int dx dy \tau(x) \tau(y) |x - y|^p \right)^2 - \mu^2. \quad (14)$$

Notice that  $\gamma = 0$  due to the definition of  $\mu$ .

In the case of the hypercube  $\nu = \nu^{\text{cube}}$ ,  $\tau(x) = \theta(x)\theta(1-x)$ , the coefficients  $\alpha$  and  $\beta$  are given by:

$$\begin{aligned} \alpha^{\text{cube}} &= \frac{p^2(p+5)}{(p+1)^2(p+2)^2(2p+1)} \\ \beta^{\text{cube}} &= \frac{2}{(p+1)^2} \left[ \frac{p^2-2}{(2p+3)(p+2)^2} + \frac{\Gamma(p+2)^2}{\Gamma(2p+4)} \right], \end{aligned} \quad (15)$$

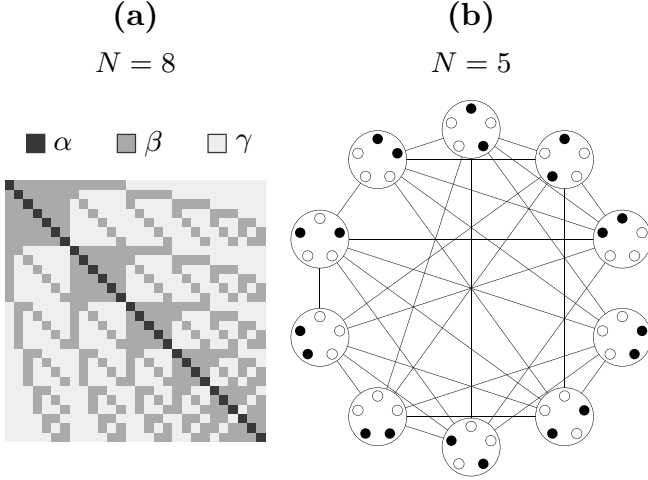


FIG. 2. (a) Example of a matrix  $\Delta(N, \alpha, \beta, \gamma)$  for  $N = 8$ . The entries with value equal to  $\beta$  have the same structure of the adjacency matrix of the Johnson graph. (b) Example of Johnson graph with  $N = 5$ . The Johnson graph  $J(N, 2)$  is the line graph of the complete graph over  $N$  nodes. It has all the distinct pairs of the original nodes as its vertices, and the vertices are linked if their pairs share an original node.

where  $\Gamma(x)$  is the Euler gamma function. In general,  $\alpha$  and  $\beta$  depend only on the choice of  $\tau$ .

The general form of a matrix with the symmetries of  $\Sigma$  is given by (see Fig. 2).

$$\Delta_{(i,j)(k,l)}(M, \alpha, \beta, \gamma) = (\alpha - 2\beta + \gamma)\delta_{i,k}\delta_{j,l} + (\beta - \gamma) \times (\delta_{i,k} + \delta_{i,l} + \delta_{j,k} + \delta_{j,l}) + \gamma, \quad (16)$$

where  $\delta_{i,j}$  is the Kronecker delta, and  $\binom{M}{2} \times \binom{M}{2}$  is the size of the matrix. We collect properties of such matrices in the following Proposition.

*Proposition 1.* Let  $\Delta$  be a matrix of the form of Eq. (16), then:

(i) it can be written as

$$\Delta(M, \alpha, \beta, \gamma) = (\alpha - \gamma)\mathbf{I} + (\beta - \gamma)\mathbf{J} + \gamma\mathbf{U}, \quad (17)$$

where  $\mathbf{I}$  is the identity matrix,  $\mathbf{U}$  is the matrix with all elements equal to one and  $\mathbf{J}$  is the adjacency matrix (with null diagonal) of the Johnson graph  $J(M, 2)$ , which is the line graph of the complete graph over  $M$  vertices;

(ii) the eigenvalues are

(a)  $\lambda_1 = \alpha + 2(N - 2)\beta + \frac{(N-2)(N-3)}{2}\gamma$  with multiplicity 1;

(b)  $\lambda_2 = \alpha + (N - 4)\beta - (N - 3)\gamma$  with multiplicity  $N - 1$ ;

(c)  $\lambda_3 = \alpha - 2\beta + \gamma$  with multiplicity  $\frac{N(N-3)}{2}$ ;

(iii) the inverse matrix  $\Delta^{-1}$  is of the same form of  $\Delta$  with inverse eigenvalues, and its parameters  $\alpha'$ ,  $\beta'$ , and  $\gamma'$  can be found by solving the linear system

$$\lambda_i[\Delta(M, \alpha, \beta, \gamma)] \times \lambda_i[\Delta(M, \alpha', \beta', \gamma')]^{-1} = 1, \quad (18)$$

for  $i = 1, 2, 3$ .

*Proof.*

(i) Follows from the explicit expression of  $\mathbf{I}$ ,  $\mathbf{J}$  and  $\mathbf{U}$ .

(ii)  $\mathbf{I}$ ,  $\mathbf{J}$ , and  $\mathbf{U}$  commute between each other, and can be diagonalized simultaneously. The contribution of  $\mathbf{I}$  is trivial.  $\mathbf{J}$  and  $\mathbf{U}$  share a nondegenerate eigenvector (that with all components equal to one) that accounts for  $\lambda_1$ . In the orthogonal subspace,  $\mathbf{U}$  represents the null operator, and does not contribute. Thus, the remainder of the spectrum is determined by that of  $\mathbf{J}$ , which is known [23].

(iii) Follows from the fact that a matrix and its inverse share the same eigenvectors. ■

#### IV. NUMBER OF CLIQUES REVEALS NONTRIVIAL STRUCTURE OF HARD GEOMETRIC GRAPHS

We are now ready to compute observables on random geometric graphs in the limit of infinite dimensions; in particular, we aim to characterize the average number of subgraphs with a given structure. Recall that the adjacency matrix of a graph  $g$  with  $M$  nodes is the  $M \times M$  matrix with entry  $A_{ij}(g) = 1$  if  $(i, j)$  is an edge of  $g$ , and  $A(g)_{ij} = 0$  otherwise.

In general, the average number of a certain subgraph  $g$  with  $M$  nodes of a random geometric graph can be factored in two terms. The first one is a combinatorial factor  $\binom{N}{M}$ , that accounts for the number of ways in which one can extract  $M$  nodes from a set of  $N$  of them. The second one is the so-called density  $\rho_g(r)$  of the subgraph  $g$  at scale  $r$ , that is the probability that  $M$  random points are close enough with respect to the cutoff radius  $r$  to form a subgraph with the same adjacency matrix of  $g$ . Recalling the definition of the joint probability of the distances between  $M$  points given in Eq. (7), we have that

$$\rho_g(r) = \int d\mathbf{y} \Pi(\mathbf{y}) \prod_{1 \leq i < j \leq M} [h_r(y_{(i,j)})]^{A_{ij}(g)}, \quad (19)$$

where  $y_{(i,j)}$  is the distance between nodes  $i$  and  $j$ . We can rescale the variables  $y_{(i,j)}$  as in Eq. (8), and exploit the fact that for large dimension  $d\mathbf{y} \Pi(\mathbf{y}) \sim d\mathbf{q} \mathcal{N}(\mathbf{0}, \Sigma)(\mathbf{q})$  (see Theorem 1) to obtain an expression for  $\rho_g(r)$  that is valid in the limit of large dimension:

$$\begin{aligned} \rho_g(r) &= \int d\mathbf{y} \Pi(\mathbf{y}) \prod_{1 \leq i < j \leq M} [h_r(y_{(i,j)})]^{A_{ij}(g)} \\ &\sim \int d\mathbf{q} \mathcal{N}(\mathbf{0}, \Sigma)(\mathbf{q}) \\ &\times \prod_{1 \leq i < j \leq M} [h_r([d\mu + \sqrt{d} q_{(i,j)}]^{\min(1, \frac{1}{p})})]^{A_{ij}(g)}, \end{aligned} \quad (20)$$

where  $\mathcal{N}(\mathbf{0}, \Sigma)$  is the multivariate Gaussian with null mean and covariance  $\Sigma$  [given in Eq. (11)], i.e.,

$$\mathcal{N}(\mathbf{0}, \Sigma)(\mathbf{q}) = \frac{e^{-\frac{1}{2}\mathbf{q}^T \Sigma \mathbf{q}}}{\sqrt{(2\pi)^{\binom{M}{2}} \det \Sigma}}. \quad (21)$$

In the rest of the section, all results are to be intended in the limit of large dimension.

As a paradigmatic example, we consider the average density of  $M$  cliques  $\rho_M(r)$ , i.e., fully connected subgraphs with  $M$  vertices, on random geometric graphs with generic activation function  $h_r(x)$ ; in this specific case,  $A_{ij}$  has only unit

entries, so that

$$\rho_M(r) = \int d\mathbf{q} \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma})(\mathbf{q}) \times \prod_{1 \leq i < j \leq M} h_r([d\mu + \sqrt{d}q_{(i,j)}]^{\min(1, \frac{1}{p})}). \quad (22)$$

In the case of hard activation function  $h^{\text{hard}}$ , we observe that

$$\begin{aligned} h_r^{\text{hard}}(x) &= h_{r^p}^{\text{hard}}(x^p) \\ h_r^{\text{hard}}(x+c) &= h_{r-c}^{\text{hard}}(x), \quad \forall c \in \mathbb{R} \\ h_r^{\text{hard}}(x) &= h_{c^r}^{\text{hard}}(cx), \quad \forall c \in \mathbb{R}^+ \end{aligned} \quad (23)$$

so that the  $p$ th root can be discarded along with a factor of  $\sqrt{d}$ , and the integral reduces to

$$\rho_M^{\text{hard}}(r) = \bar{\rho}_M^{\text{hard}}\left(\frac{r^{\max(1,p)} - d\mu}{\sqrt{d}}\right), \quad (24)$$

with

$$\begin{aligned} \bar{\rho}_M^{\text{hard}}(x) &= \int d\mathbf{q} \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma})(\mathbf{q}) \prod_{1 \leq i < j \leq M} h_x^{\text{hard}}(q_{i,j}) \\ &= \int d\mathbf{q} \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma})(\mathbf{q}) \prod_{1 \leq i < j \leq M} \theta(x - q_{i,j}), \end{aligned} \quad (25)$$

which is a multivariate Gaussian cumulative distribution function. Equation (24) highlights the simple dependence of  $\bar{\rho}_M^{\text{hard}}$  on the parameters  $p$ ,  $d$  and  $\mu$ .

In the case  $M = 2$ , the integral in Eq. (25) can be explicitly solved as it reduces to the computation of an error function, giving

$$\bar{\rho}_2^{\text{hard}}(x) = \frac{1}{2} \left[ 1 + \text{Erf}\left(\frac{x}{\sqrt{2\alpha}}\right) \right]. \quad (26)$$

The simple dependence of  $\rho_M(r)$  on  $p$ ,  $d$ , and  $\mu$  suggests to study the quantities

$$\omega_M^{\text{hard}}(x) = (\rho_M^{\text{hard}} \circ (\rho_2^{\text{hard}})^{-1})(x) = (\bar{\rho}_M^{\text{hard}} \circ (\bar{\rho}_2^{\text{hard}})^{-1})(x). \quad (27)$$

Notice that  $\omega_M(x)$  is related to  $\rho_M(r)$  by the bijective change of variable  $x = \rho_2(r)$  that, to a cutoff radius  $r$ , assigns the probability  $x$  that a random pair of nodes in the graph will be linked. Thus,  $\omega_M(x)$  gives the probability that  $M$  random nodes will form a  $M$  clique as a function of the probability that two random nodes will be linked. In practice,  $\omega_M(x)$  can be plotted by producing a scatter plot of  $\rho_M(r)$  versus  $\rho_2(r)$ . With this change of variable, the dependence of  $\rho_M$  on  $p$ ,  $d$ , and  $\mu$  cancels out, and the curves at different values of the parameters all lie in the domain  $x \in [0, 1]$ .

In the case of soft random geometric graphs with continuous activation functions, one can expand  $h_r(x)$  to the zeroth order in powers of  $1/\sqrt{d}$ , obtaining that in the limit of

high dimension

$$\begin{aligned} \rho_M^{\text{regular}}(r) &= [\rho_2^{\text{regular}}(r)]^{\binom{M}{2}} \\ \rho_2^{\text{regular}}(r) &= h_r((d\mu)^{\min(1, \frac{1}{p})}). \end{aligned} \quad (28)$$

Here, the relation between  $\rho_M$  and  $\rho_2$  reduces to that of Erdős-Rényi graphs with linking probability  $\rho_2^{\text{regular}}(r)$ , i.e.,

$$\omega_M^{\text{soft}}(x) = \omega_M^{\text{ER}}(x) = x^{\binom{M}{2}}. \quad (29)$$

In the special case of Rayleigh fading activation function  $h^{\text{rayleigh}}$ , one has

$$\rho_2^{\text{rayleigh}}(r) = \exp\left[-\xi \left(\frac{d\mu}{r}\right)^{\eta \min(1, \frac{1}{p})}\right]. \quad (30)$$

Intuitively, the difference between hard and soft RGGs depends on the freedom in performing the rescaling of the cutoff radius in the former case [see Eq. (24)], which is lost in the latter.

We performed extensive numerical simulations to study Eq. (25) and to check our analytical results; a summary is provided in Fig. 3. The numerical methods are described in Appendix. We observe a very good qualitative agreement between our analytical predictions in infinite dimension and finite-dimensional simulations for both hard and soft random geometric graphs. The convergence to the limit is fast, and even at  $d = 20$  our analytical prediction provides a good approximation of the simulated observables. More quantitatively, we observe relative deviations from the analytical predictions of the order of  $\sim 10\%$  in  $d = 20$  and  $\sim 2\%$  in  $d = 200$  in both the hard and soft case for  $k = 3$ . For  $k = 4$ , 5 relative errors are slightly larger, mainly due to the fact that we are measuring  $\rho_k(r)$  by a random sampling procedure (see Appendix) that needs more and more samples as  $k$  increases.

## V. DISCUSSION

In this work we exploited a multivariate version of the central limit theorem to compute average observables of random geometric graphs in the limit of infinite dimension. In particular, we obtained the average number of  $M$  cliques in hard and soft RGGs for different distance functions induced by  $p$  norms.

Our approach highlights that convergence to the ERG prediction for local observables depends on the choice of the ensemble: soft RGGs in particular seem to approach this naive limit for  $d \rightarrow \infty$ , whereas hard RGGs whose probability distribution of the nodes fulfills the CLT hypothesis deviate systematically from it. This result suggests that the latter provide a nontrivial null model to benchmark empirical data.

A potentially useful application of our results lies in their guidance with regard to the choice of null models, which are essential if one is to extract meaningful information from the data. For example, let us consider data points from an empirical data set (such as MNIST, for instance), and a graph constructed on these points, where a link exist whenever two data points are closer than a given cutoff radius (determining this graph is the starting point of algorithms for hierarchical clustering or manifold learning). Now, say the number of cliques in this graph deviates from the ER prediction. If

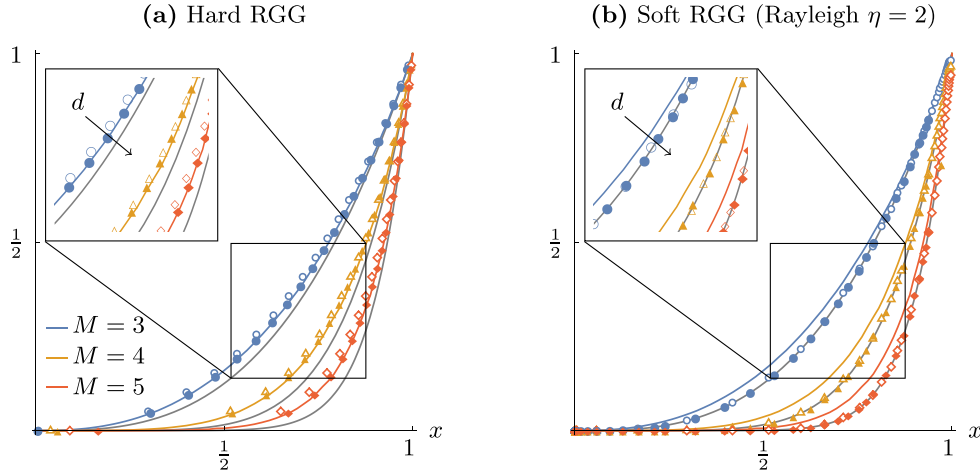


FIG. 3. Comparison between finite  $d$  simulations and infinite  $d$  analytical predictions. Colored solid lines represent the analytical predictions for  $\omega_M^{\text{hard}}(x)$  obtained from Eq. (25), for  $M = 3, 4, 5$  [blue (first line from above), orange (third line from above) and red (fifth line from above) respectively]. Gray solid lines represent  $\omega_M^{\text{ER}}(x)$  for the Erdős-Rényi graph [Eq. (29)] for comparison for the same values of  $M = 3, 4, 5$  (second, fourth, and sixth line from above, respectively). Open and filled markers show numerical simulations at  $d = 20, 200$ , respectively,  $p = 2$  and  $\nu = \nu^{\text{cube}}$ , for (a) hard RGG and (b) soft RGG with Rayleigh activation function  $\eta = 2$ , for the same values of  $M = 3, 4, 5$  (circles, triangles, and diamonds, respectively). In practice,  $\omega_M(x)$  can be represented by producing a scatter plot of  $\rho_M(r)$  versus  $\rho_2(r)$ .

we erroneously believe that RGGs in high dimension are ERGs, then we should conclude that the behavior is due to specificities of the data (e.g., deviations from the assumption of independence). This conclusion would be misleading, since, for the hard activation function, there are systematic deviations from the ER prediction even if the data points are uncorrelated and identically distributed. Our work makes clear that ruling out the null hypothesis of RGG in high dimension is fundamentally different from ruling out the hypothesis of being a ERG.

Since the CLT can be formulated in a much more general setting than the one reported in this paper, we expect that our findings hold (possibly with slight modifications) for several probability distributions of the nodes not included here, e.g., not factorized over coordinates, but with mild intercoordinate correlations; factorized over coordinates, but not identically distributed; factorized over coordinates, but with infinite second moment. The wide basin of attraction of the Gaussian limit hints to the possibility that the properties of high-dimensional structured data sets may be faithfully described by our approach. In this paper we worked with the simplest version of the CLT, as random geometric graphs are commonly studied with nodes that are independently drawn in the hypercube. The very relevant case of structured data [24–27] calls for more sophisticated CLTs, which can be addressed with the same tools developed here.

Another potentially interesting case is that of RGGs whose vertex measure is supported on low-dimensional manifolds but is embedded in a much higher-dimensional ambient space with noise. Which observables will be hidden by the added noise? And which will be robust, allowing the recovery of nontrivial properties of the underlying geometry?

Finally, our numerical simulations show that the infinite-dimensional limit is a good approximation even in finite

dimensions of order  $d \sim 10$ . This hints at the possibility to improve our results by computing higher-order corrections to the CLT, and using  $d$  as a perturbative parameter, to access the low-dimensional regime of RGGs.

**ACKNOWLEDGMENT**

P. R. acknowledges funding from the Fellini program under the H2020-MSCA-COFUND action, Grant Agreement No. 754496, INFN (IT).

**APPENDIX: NUMERICAL METHODS**

To numerically evaluate the integrals of Eq. (25), we implemented the algorithm described in Ref. [28], allowing very fast run times for the small values of  $M$  ( $M \lesssim 10$ ) we were interested in; notice that the dimension of the integral is already of order  $10^2$  for  $M = 10$ . Higher values of  $M$  would require finer techniques.

To compute the density of  $M$  cliques in simulated hard RGGs, we implemented a simple random sampling procedure, as exhaustive enumeration scales poorly, i.e., as  $O(N^M)$ , with the total number of nodes. For each realization of the nodes (with  $\nu^{\text{cube}}$  and  $N = 10^4$ ), we extracted  $\sim 5 \times 10^5$   $M$ -tuples of nodes, computing the minimum cutoff distance at which they formed a clique. The cumulative distribution of the minimal distances obtained, averaged over different realization of the nodes, reconstructs  $\rho_M^{\text{hard}}(r)$ . We noticed that as  $N$  grows, the last average is well approximated by a single realization of the nodes, suggesting a self-averaging property for the density of  $M$  cliques; in practice, not averaging does not affect the results of the simulations.

To compute the density of cliques in simulated soft RGGs with generic activation function, we implemented again a random sampling procedure. This time, for each realization of the nodes (as above) and for a fixed radius  $r$ , we counted how

many of  $\sim 10^4$   $M$ -tuples of nodes  $\{y_i\}_{i=1}^M$  where  $M$  cliques, considering each of them to be a  $M$  clique with probability

$$\prod_{1 \leq i < j \leq M} h_r[d(\bar{y}_i, \bar{y}_j)]. \quad (\text{A1})$$

Normalizing the count over the total number of candidate cliques and averaging over different realizations of the nodes (order  $10^2$ ) gives an empirical estimation for  $\rho_M(r)$  in the soft case.

- 
- [1] M. Penrose, *Random Geometric Graphs* (Oxford University Press, Oxford, 2003).
- [2] M. Barthélemy, Spatial networks, *Phys. Rep.* **499**, 1 (2011).
- [3] A. Bottinelli, R. Louf, and M. Gherardi, Balancing building and maintenance costs in growing transport networks, *Phys. Rev. E* **96**, 032316 (2017).
- [4] A. P. Giles, O. Georgiou, and C. P. Dettmann, Betweenness centrality in dense random geometric networks, in *2015 IEEE International Conference on Communications (ICC)*, pp. 6450–6455 (IEEE, London, 2015).
- [5] D. F. Nettleton, Data mining of social networks represented as graphs, *Comput. Sci. Rev.*, **7**, 1 (2013).
- [6] A. Bonato, J. Janssen, and P. Prałat, A geometric model for on-line social networks, in *Proceedings of the International Workshop on Modeling Social Media* (Association for Computing Machinery, Toronto, Ontario, Canada, 2010), pp. 1–2.
- [7] E. Estrada and M. Sheerin, Random rectangular graphs, *Phys. Rev. E* **91**, 042805 (2015).
- [8] A. Allen-Perkins, Random spherical graphs, *Phys. Rev. E* **98**, 032310 (2018).
- [9] M. D. Penrose, Connectivity of soft random geometric graphs, *Ann. Appl. Probab.* **26**, 986 (2016).
- [10] M. Ostilli and G. Bianconi, Statistical mechanics of random geometric graphs: Geometry-induced first-order phase transition, *Phys. Rev. E* **91**, 042136 (2015).
- [11] J. B. Tenenbaum, A global geometric framework for nonlinear dimensionality reduction, *Science* **290**, 2319 (2000).
- [12] S. T. Roweis, Nonlinear dimensionality reduction by locally linear embedding, *Science* **290**, 2323 (2000).
- [13] V. Erba, M. Gherardi, and P. Rotondo, Intrinsic dimension estimation for locally undersampled data, *Sci. Rep.* **9**, 1 (2019).
- [14] A. N. Gorban and I. Yu. Tyukin, Blessing of dimensionality: mathematical foundations of the statistical physics of data, *Philos. Trans. Roy. Soc. A* **376**, 20170237 (2018).
- [15] O. Bobrowski and M. Kahle, Topology of random geometric complexes: A survey, *J. Appl. Comput. Topology* **1**, 331 (2018).
- [16] L. Devroye, A. György, G. Lugosi, and F. Udina, High-dimensional random geometric graphs and their clique number, *Electronic J. Probab.* **16**, 2481 (2011).
- [17] S. Bubeck, J. Ding, R. Eldan, and M. Z. Rácz, Testing for high-dimensional geometry in random graphs, *Rand. Struct. Alg.* **49**, 503 (2016).
- [18] K. Avrachenkov and A. Bobu, Cliques in high-dimensional random geometric graphs, *Complex Networks Their Applications VIII*, Vol. 591 (Springer International Publishing, Cham, Switzerland, 2020).
- [19] J. Dall and M. Christensen, Random geometric graphs, *Phys. Rev. E* **66**, 016121 (2002).
- [20] M. Heydenreich, R. van der Hofstad, G. Last, and K. Matzke, Lace expansion and mean-field behavior for the random connection model, [arXiv:1908.11356](https://arxiv.org/abs/1908.11356).
- [21] A. P. Kartun-Giles, M. Barthélemy, and C. P. Dettmann, The shape of shortest paths in random spatial networks, *Phys. Rev. E* **100**, 032315 (2019).
- [22] A. W. Van der Vaart, *Asymptotic Statistics*, Vol. 3 (Cambridge University Press, Cambridge, 2000).
- [23] A. Burcroff, Johnson schemes and certain matrices with integral eigenvalues, University of Michigan, Tech. Rep, 2017.
- [24] F. Borra, M. C. Lagomarsino, P. Rotondo, and M. Gherardi, Generalization from correlated sets of patterns in the perceptron, *J. Phys. A: Math. Theor.* **52**, 384004 (2019).
- [25] P. Rotondo, M. C. Lagomarsino, and M. Gherardi, Counting the learnable functions of geometrically structured data, *Phys. Rev. Res.* **2**, 023169 (2020).
- [26] M. Pastore, P. Rotondo, V. Erba, and M. Gherardi, Statistical learning theory of structured data, [arXiv:2005.10002](https://arxiv.org/abs/2005.10002).
- [27] P. Rotondo, M. Pastore, and M. Gherardi, Beyond the storage capacity: Data driven satisfiability transition, [arXiv:2005.09992](https://arxiv.org/abs/2005.09992).
- [28] A. Genz, Numerical computation of multivariate normal probabilities, *J. Comput. Graph. Stat.* **1**, 141 (1992).