Real-World Networks Are Low-Dimensional: Theoretical and Practical Assessment*

Tobias Friedrich[1], **Andreas Göbel**[1], **Maximilian Katzmann**[2] and **Leon Schiller**[3]
[1]Hasso Plattner Institute, University of Potsdam, Germany
[2]Karlsruhe Institute of Technology, Germany
[3]ETH, Zürich, Switzerland
{tobias.friedrich, andreas.goebel}@hpi.de, maximilian.katzmann@kit.edu, leon.schiller@inf.ethz.ch

Abstract

Recent empirical evidence suggests that real-world networks have very low underlying dimensionality. We provide a theoretical explanation for this phenomenon as well as develop a linear-time algorithm for detecting the underlying dimensionality of such networks.

Our theoretical analysis considers geometric inhomogeneous random graphs (GIRGs), a geometric random graph model, which captures a variety of properties observed in real-world networks. These properties include a heterogeneous degree distribution and non-vanishing clustering coefficient, which is the probability that two random neighbors of a vertex are adjacent. Our first result shows that the clustering coefficient of GIRGs scales inverse exponentially with respect to the number of dimensions d, when the latter is at most logarithmic in n, the number of vertices. Hence, for a GIRG to behave like many real-world networks and have a non-vanishing clustering coefficient, it must come from a geometric space of $o(\log n)$ dimensions.

Our analysis on GIRGs allows us to obtain a linear-time algorithm for determining the dimensionality of a network. Our algorithm bridges the gap between theory and practice, as it comes with a rigorous proof of correctness and yields results comparable to prior empirical approaches, as indicated by our experiments on real-world instances. The efficiency of our algorithm makes it applicable to very large data-sets. We conclude that very low dimensionalities (from 1 to 10) are needed to explain properties of real-world networks.

1 Introduction

A key technique for understanding and analyzing large complex data sets is to embed them into a low-dimensional geometric space. Hence, the search for embedding and dimensionality reduction algorithms has become an important direction in data analysis and artificial intelligence research [Sarveniazi, 2014; Camastra and Staiano, 2016; Nickel and Kiela, 2017]. Embedding algorithms commonly require a metric that captures the similarities between data points, which is often abstracted using a graph whose vertices represent the data points and two vertices are connected if they are close with respect to this metric. The algorithm then determines geometric positions for these vertices such that connected vertices are close together. Such approaches often require an a-priori knowledge of the dimensionality, which is unknown in most applications. Heuristic approaches try to determine the dimensionality of a dataset by embedding it in spaces of different dimensionality and choosing the value that yields the optimal embedding [Levina and Bickel, 2004; Yin and Shen, 2018; Gu *et al.*, 2021].

The recent work of [Almagro *et al.*, 2022] gives a new algorithm for learning the dimensionality that does not require embeddings. Instead, given a graph as input, their method counts the number of short – i.e. length 3, 4 and 5 – cycles of a graph. It then generates a search space consisting of random graphs that are generated from a geometric model of varying parameters, including the dimensionality of the space. Finally, a data-driven classifier finds the random graph of the search space that resembles the input graph the most and returns its dimensionality. A remarkable observation, that comes from using their algorithm to learn the dimensionality of real-world networks, is that the vast majority of networks has very low dimensionality, which is independent of the size of the network.

A downside of the aforementioned approach is that it relies on machine learning techniques that are computationally heavy with no asymptotic run-time guarantees and lacks theoretical explanation for its solution quality. In order to argue with mathematical rigour, one requires to work with well-defined mathematical objects. A common approach to incorporate such an object is that of average-case analysis, that is, assume that the input graph comes from a well-defined random graph model. The random graph model that has been mostly considered so far in the literature is that of spherical random graphs, where vertices are generated independently and uniformly at random as points on the surface of a d-dimensional sphere and two vertices are connected if their angle is bellow a certain threshold. It can be easily shown that, as the number of dimensions increases, spherical random graphs converge to Erdős–Rényi graphs, the classical

*The full version which includes all proofs can be found at https://arxiv.org/abs/2302.06357.

random graph model where edges are drawn independently. A series of works considers the statistical testing problem of detecting whether a given graph is a spherical random graph or an Erdős–Rényi graph and determines the parameter regime under which this can be done [Devroye *et al.*, 2011; Bubeck *et al.*, 2016; Brennan *et al.*, 2020; Liu and Rácz, 2023b; Liu *et al.*, 2022]. Follow-up works consider noisy settings [Liu and Rácz, 2023a] or anisotropic geometric random graphs [Eldan and Mikulincer, 2020; Brennan *et al.*, 2024], where each dimension has a different influence on the drawing of edges. The techniques of the aforementioned results can also be used for determining the dimensionality of the given graph [Bubeck *et al.*, 2016, Theorem 5].

A characteristic of the random graph models considered in the aforementioned works, i.e. spherical random graphs and Erdős–Rényi graphs, is that the degree distributions of the generated graphs is concentrated around its expected value; this contrasts the power-law degree distributions observed in real-world networks [Faloutsos *et al.*, 1999]. While a latent geometric space appears to be a fundamental requirement for a random graph model that captures the high clustering coefficient [Krioukov, 2016; Boguna *et al.*, 2021] and small diameter [Friedrich *et al.*, 2013] observed in real-world networks, one needs to also consider the heterogeneity observed in the vertex degrees. A popular model in network theory capturing all previous properties is based on generating points on the hyperbolic plane instead of Euclidean space [Boguna *et al.*, 2010]. However, it is not clear what the non-geometric counterpart to test against is in this case.

Our Contribution. In this article we bring theory and practice closer together and provide a mathematical explanation for the very low dimensionality of real-world networks that has been observed in practice [Almagro *et al.*, 2022]. To this end, we build our analysis on *Geoemtric inhomogeneous random graphs (GIRGs)*, a model which was recently shown to capture many quantifiable properties of real-world networks [Bläsius and Fischbeck, 2022]. Our rigorous proofs give new insights about the model with which we are able to design a linear-time algorithm for learning the dimensionality of a network and show that when the input is a GIRG our algorithm returns the correct answer with high probability. We note that the small running time is an important advantage compared to the previously mentioned methods, as it enables working with the increasingly large networks that we encounter in practice.

2 Preliminaries

We use standard Landau notation to describe the asymptotic behavior of functions for sufficiently large n. That is, for functions f, g, we write $f(n) = \mathcal{O}(g(n))$ if there is a constant $c > 0$ such that for all sufficiently large n, $f(n) \leq cg(n)$. Similarly, we write $f(n) = \Omega(g(n))$ if $f(n) \geq cg(n)$ for sufficiently large n. If both statements are true, we write $f(n) = \Theta(g(n))$. Regarding our study of the clustering coefficient, some results make a statement about the asymptotic behavior of a function with respect to a sufficiently large d. These are marked by $\mathcal{O}_d(\cdot), \Omega_d(\cdot), \Theta_d(\cdot)$, respectively.

Geometric Inhomogeneous Random Graphs & Chung–Lu Graphs. Geometric inhomogeneous random graphs (GIRGs) as introduced in [Bringmann *et al.*, 2017] are defined as follows. Let $G(n, d, \beta, w_0, \lambda) = (V, E)$ denote the n-vertex graph obtained in the following way. For each $v \in V$, we independently sample a weight w_v from the Pareto distribution $\mathcal{P}$ with parameters $1 - \beta$ and w_0, and we draw a position $\mathbf{x}_v$ uniformly on the d-dimensional torus. Then, edges are formed depending on the weights of the vertices, the distances between the corresponding positions, and the constant $\lambda > 0$.

More precisely, the distribution of the weights is described by the distribution function

$$\mathbb{P}[w_v \leq x] = 1 - \left(\frac{x}{w_0}\right)^{1-\beta},$$

where $w_{\min} > 0$ and $\beta > 2$ are assumed to be constants, the constraint on the latter ensuring that that a single weight has finite expectation (and thus the average degree in the graph is constant), but possibly infinite variance. The density of w_v is $\rho_{w_v}(x) = (\beta - 1/w_0^{1-\beta})x^{-\beta}$. We denote the sequence of the drawn weights by $\{w\}_1^n$.

The position $\mathbf{x}_v$ in the d-dimensional torus $\mathbb{T}^d$ is drawn uniformly at random according to the standard Lebesgue measure. We denote the i-th component of $\mathbf{x}_v$ by $\mathbf{x}_{vi}$. Distances on the torus are then measured using the L_p-norm for a constant $1 \leq p \leq \infty$. That is, we define

$$\|\mathbf{x}_u - \mathbf{x}_v\|_p := \begin{cases} \left(\sum_{i=1}^d |\mathbf{x}_{ui} - \mathbf{x}_{vi}|^p\right)^{1/p} & \text{if } p < \infty \\ \max_i\{|\mathbf{x}_{ui} - \mathbf{x}_{vi}|\} & \text{otherwise.} \end{cases}$$

Note that L_∞ is a natural metric on the torus as $B_\infty(r)$, the ball of radius r under this norm, is a cube and "fits" entirely into $\mathbb{T}^d$ for all $0 \leq r \leq 1$.

Given the weights and positions, two vertices u and v are adjacent if and only if their distance $\|\mathbf{x}_u - \mathbf{x}_v\|_p$ is at most the *connection threshold* t_{uv}, which is defined such that the marginal connection probability of u, v is

$$\mathbb{P}[u \sim v] = \min\left\{1, \frac{\lambda w_u w_v}{n}\right\} = \frac{\kappa_{uv}}{n}, \tag{1}$$

for $\kappa_{uv} = \min\{n, \lambda w_u w_v\}$. Under L_∞-norm, and for $\mu = 2^d/\lambda$, this connection threshold is

$$t_{uv} = \frac{1}{2}\left(\frac{\lambda w_u w_v}{n}\right)^{1/d} = \left(\frac{w_u w_v}{\mu n}\right)^{1/d}.$$

The GIRG model has a natural non-geometric counterpart where the weight distribution of the vertices is the same as in GIRGs but the edges are now sampled independently, with probability

$$\mathbb{P}[u \sim v] = \min\left\{1, \frac{\lambda w_u w_v}{n}\right\}.$$

This inhomogeneous random graph model is known as the Chung–Lu random graph model and has been extensively studied in literature [Aiello *et al.*, 2001; Chung and Lu, 2002a; Chung and Lu, 2002b].

Hence for our analysis, we are now equipped with an appropriate geometric random graph model and its non-geometric variant. Note that the two models converge as the number of dimensions in the GIRG model goes to infinity, i.e., the total variation distance of the two models goes to zero, as it was shown in [Friedrich *et al.*, 2023a, Theorem 1]. Finally, let us note that the model is very versatile as one can consider other variants with different degree distributions or metric spaces. In the paper, we refer to GIRGs with a given arbitrary weight sequence $\{w\}_1^n$ as $\mathcal{G}(n, d, \{w\}_1^n, \lambda)$. We note, however, that our choice of the Pareto distribution for the vertex weights and of the torus for the geometric space is the one considered most frequently in literature (see e.g. [Bläsius and Fischbeck, 2022]).

Triangles and Clustering Coefficient. The number of triangles and related properties of a graph are common statistics used in the analysis of networks [Gupta *et al.*, 2014], especially for detecting underlying geometry. In fact a related statistic[1] was used in [Bubeck *et al.*, 2016] to efficiently test for geometry and the number of dimensions of a spherical random geometric graph.

When dealing with heterogeneous degree distributions, however, triangles that are attributed to large degree vertices have a significant influence on the total number of triangles. In fact as shown in [Friedrich *et al.*, 2023a], the number of triangles in GIRGs and in Chung–Lu graphs are asymptotically equivalent if $\beta \leq 7/3$, which is not an unrealistic choice for many real-world networks. In [Michielan *et al.*, 2022], it is suggested weighting each triangle by the inverse degrees of the involved vertices, a statistic called *weighted triangles*. A normalized version of the number of triangles as well as cordless squares and pentagons was also used in [Almagro *et al.*, 2022] to determine the dimensionality of a given network.

A natural statistic, observed in many real-world networks that is however strikingly absent in non-geometric random graphs where edges are drawn independently, is the existence of a non-vanishing clustering coefficient, i.e. the probability that two randomly chosen neighbours of a vertex are adjacent. The clustering coefficient is the central focus of our analysis and we use the following common definition, also used in [Keusch, 2018, Definition 5.1].

Given a graph $G = (V, E)$, its *local clustering coefficient* of a vertex v is

$$\mathrm{CC}_G(v) \coloneqq \begin{cases} \frac{|\{\{s,t\} \subseteq \Gamma(v) : s \sim t\}|}{\binom{\deg(v)}{2}} & \text{if } \deg(v) \geq 2 \\ 0 & \text{otherwise.} \end{cases}$$

The (global) *clustering coefficient* of G is the average of the local coefficient of each of G's vertices, that is,

$$\mathrm{CC}(G) \coloneqq \frac{\sum_{v \in V} \mathrm{CC}_G(v)}{|V|}.$$

For GIRGs it was shown in [Keusch, 2018] that, when the vertices of the generated graph are drawn on a torus of constant dimensionality, the generated graph has a constant clustering coefficient. On the other hand, on Chung–Lu graphs it was shown that the clustering coefficient goes to 0 as n, the number of vertices of the graph, grows [van der Hofstad *et al.*, 2017]. Our first result, which we discuss in the next section, extends the results on the clustering coefficient of GIRGs by giving an upper bound that explicitly depends on the dimension of the underlying space. This illustrates how the dimension limits the maximally achievable clustering and already explains why the dimension needs to be sufficiently low to reproduce the high clustering coefficients commonly found in real networks. We refine our result for the case of L_∞-norm in Section 3.1 and afterwards introduce an efficient statistical test for learning the dimension of a network based on the computation of clustering coefficients among nodes of similar degree (Section 4). Finally, we apply our test on a variety of real world networks and obtain results comparable to that of [Almagro *et al.*, 2022], while our test is applicable to much larger datasets. We find that the dimension corresponding to realistic networks is typically very low, ranging from 1 to around 20, supporting the hypothesis of previous work.

3 Upper Bounds on the Clustering Coefficient of GIRGs

Our first result shows that, when $d = o\log(n)$, the clustering coefficient of a GIRG scales inverse exponentially with respect to the number of dimensions d. Note that, due to space limitations, we will only discuss the technical difficulties of proving our result on a high level. For the complete proof with all details we refer the reader to the full version of the paper [Friedrich *et al.*, 2023b].

Theorem 3.1. *Asymptotically almost surely, if $d = o(\log(n))$, the clustering coefficient of G sampled from the GIRG model under some L_p-norm with $p \in [1, \infty]$ is*

$$\mathrm{CC}(G) = \exp(-\Omega_d(d)) + o(1).$$

For the case of L_∞-norm, we later derive a much more precise bound. Theorem 3.1 implies that if $d = \omega(1)$ and $d = o(\log(n))$ the clustering coefficient vanishes and that already for constant (sufficiently large) values of d, clustering decays at least exponentially in d. As most real-world networks have a non-vanishing clustering coefficient, our theorem suggests that their dimensionality must be at most constant in the number of vertices. This is the first rigorous theoretical explanation for the empirical observations of the low dimensionality of real-world networks by Almagro, Boguñá and Serrano [Almagro *et al.*, 2022].

Besides the results in [Keusch, 2018] for a constant number of dimensions, the clustering coefficient of random geometric graphs (i.e., our model in the case of homogeneous weights) under the L_2-norm as a function of d was previously analyzed in [Dall and Christensen, 2002]. However, Theorem 3.1 applies to inhomogeneous weights and arbitrary L_p-norms, which complicates the analysis. The main difficulty in proving Theorem 3.1 is that the probability that two random neighbours of a given vertex are connected is significantly dependent on their weights. Our proof circumvents this issue by showing that high-weight nodes only have a small influence on the global clustering coefficient of a power-law graph G.

[1] The statistic used, number of signed triangles, essentially measures by how much the number of triangles in the graph exceeds the expected value in the Erdős–Rényi graph model.

Via an application of the method of typical bounded differences (see e.g. [Warnke, 2016]) – a generalisation of McDiarmid's inequality [McDiarmid, 1989] and a powerful tool to showing concentration in high-dimensional spaces – we then show that the clustering coefficient of a GIRG concentrates around the expected clustering coefficient of a subgraph induced by vertices of small weight.

The bound on the clustering coefficient in this subgraph follows from a bound on the probability that two random vectors y_u, y_v uniformly distributed within the unit ball under L_p-norm have a distance larger than a certain threshold. Intuitively, the fact that this probability decays exponentially with d is a consequence of the law of large numbers: as d grows, with large probability, about half of the components of y_u and y_v have opposite sign, which already leads to a distance between u and v that is arbitrarily close to 1 with probability converging to 1 as d grows. Taking into account that the other components of y_u and y_v also contribute at least a constant increase in distance between u and v with large probability, we get that there is an exponentially increasing probability that the distance between u and v is strictly greater than one, which suffices to show an exponential upper bound on the clustering coefficient in G. To prove this exponential decay in terms of d, we use a coupling argument based on the observation that the "direction" $\mathbf{x}/\|\mathbf{x}\|_p$ and the norm $\|x\|_p$ of a random vector distributed in the unit ball under L_p-norm are independent. To analyze $\mathbf{x}/\|\mathbf{x}\|_p$, we define the following distribution and show that if $\mathbf{z}$ is a vector sampled from this distribution, then $\mathbf{z}/\|\mathbf{z}\|_p$ is distributed just as $\mathbf{x}/\|\mathbf{x}\|_p$. This has the advantage that the components of $\mathbf{z}$ are now independent, allowing us to apply sharp tail bounds from which our statements follow. While these arguments are only valid for random vectors distributed in $\mathbb{R}^d$, we show that they remain valid on the torus if we restrict ourselves to vertices of sufficiently low weight.

The χ^p-Distribution. Let $p \in \mathbb{R}, p \geq 1$. We say that a random vector $\mathbf{x} \in \mathbb{R}^d$ is $\chi_p(d)$ distributed if each of its components x_i is independently distributed according to the density function $\rho(x_i) \coloneqq \gamma e^{-\frac{1}{2}|x_i|^p}$, with normalising constant $\gamma = p(2^{1/p+1}\Gamma(1/p))^{-1}$, where $\Gamma(s) = \int_0^\infty x^{s-1}e^{-x}\mathrm{d}x$ is the gamma function. If $\mathbf{x} \sim \chi_p(d)$, then we denote the distribution of $\left(\|\mathbf{x}\|_p\right)^p = \sum_{i=1}^d |x_i|^p$ by $\chi^p(d)$.

This distribution is a generalisation of the χ^2 distribution and a simplification of the one proposed by [Livadiotis, 2014]. In the full version of the paper [Friedrich *et al.*, 2023b], we determine its moment generating function. This not only gives us its expectation, which is $2d/p$, but also allows us to prove the following concentration bound, which we use in the arguments used for the proof of Theorem 3.1.

Corollary 3.2. *Let $X_i, \dots X_d$ be i.i.d. random variables from $\chi_p(1)$ and define $Z = \sum_{i=1}^k |X_i|^p \sim \chi^p(d)$. Then, for every $\delta > 0$ and $\varepsilon = \sqrt{2\delta} + \delta$, it holds that*

$$\mathbb{P}[Z \geq (1+\varepsilon)\mathbb{E}[Z]] \leq \exp\left(-\frac{2\delta}{p}\cdot d\right),$$

$$\mathbb{P}[Z \leq (1-\varepsilon)\mathbb{E}[Z]] \leq \exp\left(-\frac{2\delta}{p}\cdot d\right).$$

We believe our analysis of the $\chi_p(d)$ and $\chi^p(d)$ distributions to be of independent interest, as many random spaces can be related to vectors drawn uniformly at random within the d-dimensional unit ball of some L_p-norm.

3.1 Improved Bounds for the L_∞-Norm

When using L_∞-norm as a distance measure for GIRGs we obtain more precise results and are able to further determine the base of the exponent.

Theorem 3.3. *Asymptotically almost surely, if $d = o(\log(n))$, then the clustering coefficient of G sampled from the GIRG model with L_∞-norm fulfils*

$$\mathrm{CC}(G) \leq 3\left(\frac{3}{4}\right)^{d\left(1-\frac{1}{\beta}\right)} + o(1).$$

Recall that the L_∞ norm is not only a natural distance measure on the torus from a mathematical point of view, but also one that yields graphs that closely resemble real-world networks [Bläsius and Fischbeck, 2022]. The stronger result for L_∞-norm is based on an application of the following theorem of Friedrich et al. [Friedrich *et al.*, 2023a, Theorem 3] that estimates the probability that three random vertices that have similar weights are a triangle if the vertex of minimal weight is adjacent to the other two and the weights are sufficiently small compared to n. The complete proof of Theorem 3.3 can be found in [Friedrich *et al.*, 2023b].

Theorem 3.4. *Let G be a GIRG generated under L_∞-norm. Let $U = \{v, s, t\}$ be a set of 3 vertices with weights w_v, w_s, w_t such that $w_v \leq w_s \leq w_t$ and $w_v \leq cw_t$ for some constant $c > 0$. If $\left(w_t^2/(\mu n)\right)^{1/d} \leq 1/4$, we have*

$$\left(\frac{3}{4}\right)^d \leq \mathbb{P}[U \text{ is a triangle} \mid v \sim s, t] \leq c\left(\frac{3}{4}\right)^d.$$

4 Learning the Dimensionality

While the previous results show that the dimension needs to be sufficiently low to allow for large clustering coefficients, it does not explicitly allow us to recover the dimension of a network as the hidden constants in our statements may depend on other parameters such as β, w_0, λ. In the following, we address this issue by presenting a statistical test for recovering the dimension of a GIRG with high probability.

[Michielan *et al.*, 2022] study the problem of detecting the underlying geometry in GIRGs of constant dimension under L_∞-norm. As previously mentioned, they observe that the number of triangles of a GIRG is close to that of a Chung–Lu graph when β is close to 2 as then, the number of triangles is dominated by those forming among high-degree vertices. The authors therefore suggest to count the number of weighted triangles instead, where each triangle contributes a weight that is inversely proportional to the product of the degrees of its vertices. Weighted triangles thus counteract the effect of high-degree vertices as the influence of triangles forming among such vertices is diminished. However, this approach

only allows to decide whether the network has an underlying metric structure, but not its dimensionality.

We take a similar (yet more direct) approach for excluding the effect of high-degree vertices and introduce a test that is able to infer the dimension of the underlying metric space. Namely, using Theorem 3.4 together with the method of typical bounded differences [Keusch, 2018, Theorem 2.5] we can show that the average local clustering coefficient in the induced subgraph of all vertices with at least two neighbors and weight in the interval $[w_c, cw_c]$ (where $w_c \geq w_0, 0 < c < 2/\sqrt{3}$) concentrates tightly around a value that only depends on d and not on β, w_0 or λ.

We use the following lemma.

Lemma 4.1. *Let $G = (V, E)$ be a GIRG and let $V_{\leq \log(n)}$ be the set of all vertices with weight at most $\log(n)$. With probability at least $1 - n^{-\Omega(\log^2(n))}$, we have for all $v \in V_{\leq \log(n)}$ that $\deg(v) \leq \log^3(n)$.*

The proof is in the full version of our paper [Friedrich *et al.*, 2023b]. Our main result follows.

Theorem 4.2. *Let G be a GIRG generated under L_∞-norm. Let further $1 < c, w_c \geq w_0$ be constants, and let G' be the subgraph of G consisting of all vertices with weight in $[w_c, cw_c]$. Assume that d is an integer with $d = o(\log(n))$. Define the set S as the set of nodes in G' that have at least two neighbors in G' and the random variable $\mathrm{CC}^{(+)}(G')$ as*

$$\mathrm{CC}^{(+)}(G') = \frac{1}{|S|} \sum_{v \in S} \frac{|\{\{s,t\} \subseteq \Gamma(v) \mid s \sim t\}|}{\binom{\deg(v)}{2}} = \frac{1}{|S|} \sum_{v \in S} \mathrm{CC}_{G'}(v).$$

Then,

$$\mathrm{CC}^{(+)}(G') \in \left(\frac{1}{c} \left(\frac{3}{4}\right)^d, c \left(\frac{3}{4}\right)^d \right) \pm n^{-1/5}$$

with probability at least $1 - o(1/n)$.

Proof. We start by estimating the expectation of $\mathrm{CC}^{(+)}(G')$. It is not hard to see that by linearity of expectation

$$\mathbb{E}\Big[|S| \, \mathrm{CC}^{(+)}(G')\Big] = \sum_{v \in G'} \mathbb{E}[\mathbb{1}(v \in S) \, \mathrm{CC}_{G'}(v)] = \mathbb{E}[|S|] \, \mathbb{P}[\Delta \mid v \sim s, t]$$

where v, s, t are three random vertices in G', and Δ is the event that v, s, t are a triangle. Our proof proceeds in two steps: (1) we show that $\mathbb{P}[\Delta \mid v \sim s, t]$ is in the interval $(c^{-1}(3/4)^d, c(3/4)^d)$, and (2) we show that $\mathrm{CC}^{(+)}(G')$ concentrates around its expectation using the method of typical bounded differences.

For part (1), we apply Theorem 3.4 and note that – since v, s, t are in G' – the weights of v, s, t differ by at most a factor of c. However, Theorem 3.4 only yields a bound on the probability of Δ if v is the vertex of minimal weight among v, s, t. However, we can express $\mathbb{P}[\Delta \mid v \sim s, t]$ as

$$\frac{\mathbb{P}[\Delta]}{\mathbb{P}[v \sim s, t]} = \mathbb{P}\big[\Delta \mid \hat{v} \sim \hat{s}, \hat{t}\big] \, \frac{\mathbb{P}\big[\hat{v} \sim \hat{s}, \hat{t}\big]}{\mathbb{P}[v \sim s, t]}$$

where $\hat{v}, \hat{s}, \hat{t}$ are the vertices v, s, t reordered such that $\hat{v}$ is of minimal weight. Since $\mathbb{P}[v \sim s, t] = \lambda^2 w_v^2 w_s w_t / n^2$ (we can ignore the minimum in eq. (1) here because the weights are constant), we can conclude that the fraction in the above equation is at least $1/c$ and at most 1. For the remaining term, Theorem 3.4 is applicable so that we can conclude that

$$\mathbb{P}[\Delta \mid v \sim s, t] \in \left(\frac{1}{c} \left(\frac{3}{4}\right)^d, c \left(\frac{3}{4}\right)^d \right)$$

For the second part of the proof, we first show that $|S|$ is linear and concentrates well around its expectation. This will ensure that we can apply the method of typical bounded differences to $|S| \cdot \mathrm{CC}^{(+)}(G')$. We start by showing that there is a constant $\alpha > 0$ such that $|S|$ is at least αn with probability $1 - n^{-\omega(1)}$. Consider a fixed vertex v from G and denote the number of its neighbors in G' by X_v. We note that every vertex in G has a constant probability of being in G' and a probability of at least $\lambda w_0^2 / n$ to connect to v. X_v is therefore lower bounded by the sum of n independent Bernoulli distributed random variables with success probability in $\Theta(1/n)$. Denote this sum by X'_v and note that $\mathbb{E}[X'_v] = \Theta(1)$. By [Cam, 1960, Proposition 1], X'_v converges to a Poisson distributed random variable with constant expectation. Accordingly, $\mathbb{P}[X'_v \geq 2]$ is constant as well. This shows that every vertex in G has at least a constant probability of having two neighbors in G'. As the probability that v is in G' is constant as well, this implies that $\mathbb{E}[|S|] = \Omega(n)$. We continue with showing concentration of this random variable using [Keusch, 2018, Theorem 2.5]. We note that the random variables $x_1, x_2, \ldots, x_n$ (the positions of all vertices), and $w_1, w_2, \ldots, w_n$ (the weights of all vertices) are independent and define a product probability space Ω as in [Keusch, 2018, Theorem 2.5]. Each $\omega \in \Omega$ defines a graph $G(\omega)$, and $f(\omega)$ is defined as the value of $|S|$ in this graph. We consider the "bad" event

$$\mathcal{B} = \{\omega \in \Omega \mid \text{the max degree in } G' \text{ is } > \log^3(n)\}.$$

By Lemma 4.1, $\mathcal{B}$ happens with probability $n^{-\omega(1)}$. Now, let $\omega, \omega' \in \overline{\mathcal{B}}$ be such that they differ in at most two coordinates. Changing the weight or coordinate of one vertex can only decrease the number of vertices in G' with at least two neighbors by at most $2\log^3(n)$ as the weight or coordinate change only influences vertices that are neighbours of the changed vertex before or after the change. Accordingly, two coordinate or weight changes can only change $|S|$ by at most $c' := 4\log^3(n)$. Using $t = n^{3/4}$ further fulfills the condition $t \geq 2M\mathbb{P}[\mathcal{B}]$ as $M \leq n$ and $\mathbb{P}[\mathcal{B}] = n^{-\omega(1)}$. As $m = 2n$, we get,

$$\mathbb{P}\Big[|S| - \mathbb{E}[|S|] \mid \geq n^{3/4}\Big] \leq 2\exp\left(-\frac{n^{1/2}}{32 \cdot 2 \cdot 16 \log^6(n)}\right) + \left(\frac{n^2}{\log^3(n)} + 1\right) n^{-\omega(1)} = n^{-\omega(1)}.$$

Similarly, we can show concentration of $f(S) = |S| \cdot \mathrm{CC}^{(+)}(G') = \sum_{v \in S} |\{\{s,t\} \subseteq \Gamma(v) \mid s \sim t\}| / \binom{\deg(v)}{2}$. Again, changing the coordinate or weight of any two vertices

can only increase or decrease the local clustering coefficient of at most $4\log^3(n)$ vertices by a value of at most one. Hence, we can again choose $c' := 4\log^3(n)$ and $t = n^{3/4}$ to obtain

$$\mathbb{P}\Big[|f(S) - \mathbb{E}[f(S)]| \geq n^{3/4}\Big] \leq n^{-\omega(1)}.$$

Combining these two concentration results, we get that

$$f(S) = \mathbb{E}[|S|]\,\mathbb{P}[\Delta \mid v \sim s,t] \pm n^{3/4} \text{ and}$$
$$|S| = \mathbb{E}[|S|] \pm n^{3/4}$$

both hold with probability $1 - o(1/n)$. Dividing by $|S|$ and using $\mathbb{E}[|S|] = \Theta(n)$ then yields

$$\begin{aligned}&\mathrm{CC}^{(+)}(G')\\ &= \frac{\mathbb{E}[|S|]}{\mathbb{E}[|S|] \mp n^{3/4}}\mathbb{P}[\Delta \mid v \sim s,t] \pm \frac{n^{3/4}}{\mathbb{E}[|S|] \mp n^{3/4}}\\ &= \mathbb{P}[\Delta \mid v \sim s,t] \pm n^{-1/5}.\end{aligned}$$

Using our initial estimate for $\mathbb{P}[\Delta \mid v \sim s,t]$ concludes the proof. □

Theorem 4.2 can be viewed as a linear-time algorithm for the following statistical testing problem (assuming that w_c is constant). We are given a graph G on n vertices, its weight sequence, and an integer $d = o(\log(n))$. Under the null hypothesis, G is a GIRG with dimension d, whereas under the alternative hypothesis, G was generated in dimension $d_1 \neq d$ or it is a Chung–Lu graph. Here, we allow d_1 to be any integer (potentially larger than $\log(n)$). Consider the following testing procedure for this problem. Fix a constant $1 < c < 2/\sqrt{3}$ and a weight $w_c \geq w_0$. Now, consider the induced subgraph G' of G consisting of all nodes with weight in $[w_c, c \cdot w_c]$. For every node $v \in G'$ that has at least two neighbours in G', compute its local clustering coefficient $CC_{G'}(v)$ and denote by $\mathrm{CC}^{(+)}(G')$ the mean over all these values. We accept the null hypothesis if and only if $-n^{-1/5} + \frac{1}{c}\left(\frac{3}{4}\right)^d \leq \mathrm{CC}^{(+)}(G') \leq c\left(\frac{3}{4}\right)^d + n^{-1/5}$.

Due to Theorem 4.2 the probability that our test makes a mistake assuming that the null hypothesis is true is only $n^{-\omega(1)}$. Under the alternative hypothesis, assume that d_1 is the ground truth dimension G came from, and assume further without loss of generality that $d_1 \geq d + 1$. We have to show that asymptotically,

$$\begin{aligned}&\frac{1}{c}\left(\frac{3}{4}\right)^d - n^{-1/5} > c\left(\frac{3}{4}\right)^{d+1} + n^{-1/5}\\ &\qquad\Leftrightarrow 1 > \frac{3c^2}{4} + 2\left(\frac{4}{3}\right)^d n^{-1/5}.\end{aligned}$$

Accordingly, if $d = o(\log(n))$ and if we choose $c < 2/\sqrt{3} \approx 1.155$, this inequality is true for sufficiently large n.

Finally, we observe that the running time of this test is linear, as we have to compute the local clustering coefficient of vertices of constant weight and as the degree of a vertex with weight cw_c is constant in expectation. Iterating this statistical test over the range of d we can recover the dimensionality of the input graph with high probability. Let us note that our result is not restricted to constant number of dimensions but applies to the whole regime with $d = o(\log(n))$, which as Theorem 3.3 implies, is the only relevant one.

5 Application to Real-World Networks

In addition to our theoretical results, we implemented our algorithm and evaluated it on both real-world networks and GIRGs[2]. For estimating the vertex weights, we used the maximum likelihood estimator derived in [Boguna *et al.*, 2010, Appendix B.2]. The outcome of our experiments is summarised in Figure 1, where we plot our test statistic for real and synthetic networks over different choices of w_c. We also include a histogram showing the frequencies of the inferred dimensions over our whole dataset of about 3k real-world networks. The dimension was inferred by taking a weighted median (weighted by the size of the induced subgraph of vertices with weight in $[w_c, cw_c]$) over different choices of w_c ranging in $\{2, \ldots, 300\}$. Our dataset of real-world networks for the plots in Figure 1 outside the histogram is a collection of 65 networks from the SNAP-dataset [Leskovec and Krevl,] and Network Repository [Rossi and Ahmed, 2015] with sizes between 10k and 4M vertices. We mainly use social, citation, collaboration and biological networks. The histogram in Figure 1 additionally uses a dataset of 2976 real-world networks from [Bläsius and Fischbeck, 2022]. The results we obtain are indeed similar to the results in [Almagro *et al.*, 2022, Fig. 5] while our algorithm has stronger theoretical foundations and is much more efficient. In fact, we are able to handle datasets of orders of magnitude larger than that of [Almagro *et al.*, 2022]. We observe in Figure 1 that some real-world networks show an overall similar behaviour as the GIRGs (e.g. `soc-academia`, `fb-pages-artist`, `ca-AstroPh`,`bio-WormNet`, cf. Figure 1) and that very low dimensions (ranging from 1 to around 10) are required to explain the properties of the networks in our dataset. Our results in Figure 1 confirm the hypothesis that the vast majority of realistic networks corresponds to very low dimensionalities although there are outliers. It is not a surprise that real-world data can be noisy and, therefore, often exhibit a behaviour that differs from synthetic networks. Roughly speaking the variance of the clustering coefficients over different weight ranges $[w_c, cw_c]$ is higher in those real-world networks than it is on GIRGs. However, our results on these networks still yield an estimate of their dimensionality, or at least a range of realistic dimensionalities. Such difficulties were also encountered in [Almagro *et al.*, 2022] (see their supplementary material). A similar noisy behaviour can also be observed in small generated GIRGs, where the number of vertices is not high enough for the concentration results to be strong. Our algorithm performs exceptionally well at recovering the ground truth for the number of dimensions when the number of vertices is sufficiently high and the test is performed among vertices of sufficiently low weight, as predicted by our theoretical results. We further remark that the GIRG model in the literature is often introduced with an additional temperature parameter $T \in [0, 1]$ (e.g. in [Bläsius *et al.*, 2022]) that controls the influence of the underlying geometry and influences our test, with higher temperatures leading to lower clustering. While our theoretical results technically only apply if the temperature is 0, experiments show that the test remains robust if $T \leq 1/2$ with inferred dimensions on synthetic networks that

[2]Code: https://github.com/leon-schi/dimensionality-estimation.

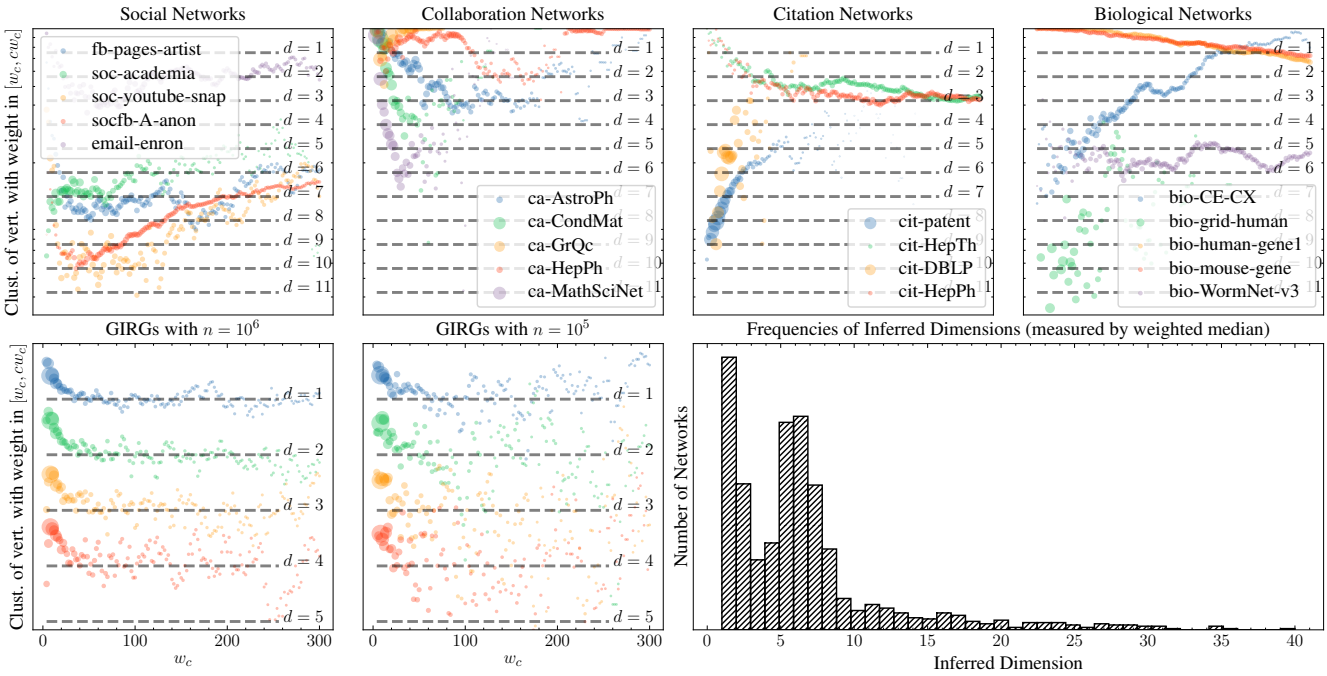

Figure 1: First row: our test statistic for different choices of w_c in real-world networks. The size of the circles indicates the number of vertices in the induced subgraph of vertices with weight in $[w_c, cw_c]$ for $c = 1.15$. Dashed lines show a lower bound on the expected value of our test statistic for a GIRG generated in dimension d, i.e $(3/4)^d$. Second row, left: the same procedure for synthetic networks, generated with the algorithm of [Bläsius *et al.*, 2022] using $\beta = 2.5, \alpha = 10$ and average degree 10. Second row, right: distribution of the inferred dimension of 2976 real-world networks from the dataset of [Bläsius and Fischbeck, 2022].

only exceed the ground truth by at most 2. Furthermore, as T increases, so does the inferred dimension, hence our experiments are still an indicator of the maximally admissable dimension of a network, even for arbitrarily large temperatures. Our results on real-world networks further indicate that the GIRG model does not fully capture the properties of real networks, motivating further research.

6 Future Work

A large body of work has been devoted to understanding for which asymptotic behavior of d geometry is detectable in spherical random geometric graphs (SRGGs) for homogeneous weights. While the parameter regime where these graphs lose geometry in the dense case, i.e. the case where the edge probability is constant and does not depend on n, is well understood [Devroye *et al.*, 2011; Bubeck *et al.*, 2016; Liu and Rácz, 2023b], it is unclear what happens in the sparse case (where the marginal connection probability is proportional to $1/n$) and progress has been made only recently [Brennan *et al.*, 2020; Liu *et al.*, 2022].

On the other hand, research devoted to studying the influence of the dimension on random geometric graphs in the case of inhomogeneous weights is sparse. We gave first results in this regard by studying how the clustering coefficient depends on d and showed that the dimension can be detected by means of statistical testing assuming that $d = o(\log(n))$, which is a valid assumption for realistic data. It is interesting to study under which conditions the geometry remains detectable if $d = \Omega(\log(n))$ and under which circumstances the model converges to its non-geometric counterpart with respect to the total variation distance of the distributions over the produced graphs as previously studied for spherical random graphs. Also, to determine what differences arise when using the torus instead of the sphere as the underlying metric space and which influence different norms have in this regard.

Noisy settings have also been considered in the context of testing for geometry in random graphs [Liu and Rácz, 2023a]. The GIRG model comes with a natural way of modelling noise in the form of an (inverse) temperature parameter $\alpha > 1$ [Keusch, 2018]. As higher temperatures diminish clustering in addition to the effects induced by the dimension, it would be relevant to generalize our results to the case of non-zero temperatures and to investigate how the effects of high temperature can be distinguished from that of high dimension.

Finally, our experiments indicate that some real-world networks behave differently than GIRGs when being subjected to our test. One possible explanation is the non-zero assortativity of many real networks. Therefore, we would like to understand how this parameter influences our test and how the GIRG model can be extended by further parameters that allow adjusting assortativity and related properties.

Ethical Statement

There are no ethical issues.

Aknowledgements

Tobias Friedrich was funded by project No. 390859508 of the German Research Foundation. Andreas Göbel was funded by the project PAGES (project No. 467516565) of the German Research Foundation (DFG). We thank Marcos Kiwi for fruitful discussions related to this work.

References

[Aiello *et al.*, 2001] William Aiello, Fan Chung, and Linyuan Lu. A random graph model for power law graphs. *Experimental Mathematics*, 10(1):53–66, 2001.

[Almagro *et al.*, 2022] Pedro Almagro, Marián Boguñá, and M Ángeles Serrano. Detecting the ultra low dimensionality of real networks. *Nature communications*, 13(1):6096, 2022.

[Bläsius and Fischbeck, 2022] Thomas Bläsius and Philipp Fischbeck. On the external validity of average-case analyses of graph algorithms. In *30th Annual European Symposium on Algorithms (ESA 2022)*. Schloss Dagstuhl-Leibniz-Zentrum für Informatik, 2022.

[Bläsius *et al.*, 2022] Thomas Bläsius, Tobias Friedrich, Maximilian Katzmann, Ulrich Meyer, Manuel Penschuck, and Christopher Weyand. Efficiently generating geometric inhomogeneous and hyperbolic random graphs. *Network Science*, 10(4):361–380, 2022.

[Boguna *et al.*, 2010] Marian Boguna, Fragkiskos Papadopoulos, and Dmitri Krioukov. Sustaining the internet with hyperbolic mapping. *Nature Communications*, 1(1):62, Dec 2010. arXiv: 1009.0267.

[Boguna *et al.*, 2021] Marian Boguna, Ivan Bonamassa, Manlio De Domenico, Shlomo Havlin, Dmitri Krioukov, and M Ángeles Serrano. Network geometry. *Nature Reviews Physics*, 3(2):114–135, 2021.

[Brennan *et al.*, 2020] Matthew Brennan, Guy Bresler, and Dheeraj Nagaraj. Phase transitions for detecting latent geometry in random graphs. *Probability Theory and Related Fields*, 178(3-4):1215–1289, 2020.

[Brennan *et al.*, 2024] Matthew Brennan, Guy Bresler, and Brice Huang. Threshold for detecting high dimensional geometry in anisotropic random geometric graphs. *Random Structures & Algorithms*, 64(1):125–137, 2024.

[Bringmann *et al.*, 2017] Karl Bringmann, Ralph Keusch, and Johannes Lengler. Sampling Geometric Inhomogeneous Random Graphs in Linear Time. In *25th Annual European Symposium on Algorithms (ESA 2017)*, Leibniz International Proceedings in Informatics (LIPIcs), pages 20:1–20:15, 2017.

[Bubeck *et al.*, 2016] Sébastien Bubeck, Jian Ding, Ronen Eldan, and Miklós Z Rácz. Testing for high-dimensional geometry in random graphs. *Random Structures and Algorithms*, 49(3):503–532, 2016.

[Cam, 1960] Lucien Le Cam. An approximation theorem for the poisson binomial distribution. *Pacific Journal of Mathematics*, 10(4):1181–1197, Jan 1960.

[Camastra and Staiano, 2016] Francesco Camastra and Antonino Staiano. Intrinsic dimension estimation: Advances and open problems. *Information Sciences*, 328:26–41, 2016.

[Chung and Lu, 2002a] Fan Chung and Linyuan Lu. The average distances in random graphs with given expected degrees. *Proceedings of the National Academy of Sciences*, 99(25):15879–15882, 2002.

[Chung and Lu, 2002b] Fan Chung and Linyuan Lu. Connected Components in Random Graphs with Given Expected Degree Sequences. *Annals of Combinatorics*, 6(25):125–145, 2002.

[Dall and Christensen, 2002] Jesper Dall and Michael Christensen. Random geometric graphs. *Physical Review E*, 66(1):016121, Jul 2002. arXiv:cond-mat/0203026.

[Devroye *et al.*, 2011] Luc Devroye, András György, Gábor Lugosi, and Frederic Udina. High-dimensional random geometric graphs and their clique number. *Electronic Journal of Probability*, 16:2481–2508, Jan 2011.

[Eldan and Mikulincer, 2020] Ronen Eldan and Dan Mikulincer. Information and dimensionality of anisotropic random geometric graphs. In *Geometric Aspects of Functional Analysis: Israel Seminar (GAFA) 2017-2019 Volume I*, pages 273–324. Springer, 2020.

[Faloutsos *et al.*, 1999] Michalis Faloutsos, Petros Faloutsos, and Christos Faloutsos. On power-law relationships of the internet topology. *ACM SIGCOMM computer communication review*, 29(4):251–262, 1999.

[Friedrich *et al.*, 2013] Tobias Friedrich, Thomas Sauerwald, and Alexandre Stauffer. Diameter and broadcast time of random geometric graphs in arbitrary dimensions. *Algorithmica*, 67:65–88, 2013.

[Friedrich *et al.*, 2023a] Tobias Friedrich, Andreas Göbel, Maximilian Katzmann, and Leon Schiller. Cliques in high-dimensional geometric inhomogeneous random graphs. In Kousha Etessami, Uriel Feige, and Gabriele Puppis, editors, *50th International Colloquium on Automata, Languages, and Programming, ICALP*, pages 62:1–62:13, 2023.

[Friedrich *et al.*, 2023b] Tobias Friedrich, Andreas Göbel, Maximilian Katzmann, and Leon Schiller. A simple statistic for determining the dimensionality of complex networks. *CoRR*, abs/2302.06357, 2023.

[Gu *et al.*, 2021] Weiwei Gu, Aditya Tandon, Yong-Yeol Ahn, and Filippo Radicchi. Principled approach to the selection of the embedding dimension of networks. *Nature Communications*, 12(1):3772, 2021.

[Gupta *et al.*, 2014] Rishi Gupta, Tim Roughgarden, and Comandur Seshadhri. Decompositions of triangle-dense graphs. In *Proceedings of the 5th conference on Innovations in theoretical computer science*, pages 471–482, 2014.

[Keusch, 2018] Ralph Keusch. *Geometric Inhomogeneous Random Graphs and Graph Coloring Games*. Doctoral thesis, ETH Zurich, 2018.

[Krioukov, 2016] Dmitri Krioukov. Clustering implies geometry in networks. *Physical review letters*, 116(20):208302, 2016.

[Leskovec and Krevl,] Jure Leskovec and Andrej Krevl. SNAP Datasets: Stanford large network dataset collection. http://snap.stanford.edu/data. Accessed: 2024-01-12.

[Levina and Bickel, 2004] Elizaveta Levina and Peter Bickel. Maximum likelihood estimation of intrinsic dimension. *Advances in neural information processing systems*, 17, 2004.

[Liu and Rácz, 2023a] Suqi Liu and Miklós Z Rácz. Phase transition in noisy high-dimensional random geometric graphs. *Electronic Journal of Statistics*, 17(2):3512–3574, 2023.

[Liu and Rácz, 2023b] Suqi Liu and Miklós Z Rácz. A probabilistic view of latent space graphs and phase transitions. *Bernoulli*, 29(3):2417–2441, 2023.

[Liu *et al.*, 2022] Siqi Liu, Sidhanth Mohanty, Tselil Schramm, and Elizabeth Yang. Testing thresholds for high-dimensional sparse random geometric graphs. In *Proceedings of the 54th Annual ACM SIGACT Symposium on Theory of Computing*, STOC 2022, page 672–677, New York, NY, USA, 2022. Association for Computing Machinery.

[Livadiotis, 2014] George Livadiotis. Chi-p distribution: characterization of the goodness of the fitting using lp-norms. *Journal of Statistical Distributions and Applications*, 1(1):4, Jun 2014.

[McDiarmid, 1989] Colin McDiarmid. On the method of bounded differences. *Surveys in combinatorics*, 141(1):148–188, 1989.

[Michielan *et al.*, 2022] Riccardo Michielan, Nelly Litvak, and Clara Stegehuis. Detecting hyperbolic geometry in networks: Why triangles are not enough. *Physical Review E*, 106(5):054303, 2022.

[Nickel and Kiela, 2017] Maximillian Nickel and Douwe Kiela. Poincaré embeddings for learning hierarchical representations. *Advances in neural information processing systems*, 30, 2017.

[Rossi and Ahmed, 2015] Ryan A. Rossi and Nesreen K. Ahmed. The network data repository with interactive graph analytics and visualization. In *AAAI*, 2015.

[Sarveniazi, 2014] Alireza Sarveniazi. An actual survey of dimensionality reduction. *American Journal of Computational Mathematics*, 2014, 2014.

[van der Hofstad *et al.*, 2017] Remco van der Hofstad, A. J. E. M. Janssen, Johan S. H. van Leeuwaarden, and Clara Stegehuis. Local clustering in scale-free networks with hidden variables. *Physical Review E*, 95(2):022307, Feb 2017.

[Warnke, 2016] Lutz Warnke. On the method of typical bounded differences. *Combinatorics, Probability and Computing*, 25:269–299, 2016.

[Yin and Shen, 2018] Zi Yin and Yuanyuan Shen. On the dimensionality of word embedding. *Advances in neural information processing systems*, 31, 2018.