# Random Geometric Graph: Some recent developments and perspectives

**Quentin Duchemin and Yohann De Castro**

**Keywords** Random Geometric Graphs • Concentration inequality for U-statistics • Random matrices • Non-parametric estimation • Spectral clustering • Coupling • Information inequalities

**Abstract**

The Random Geometric Graph (RGG) is a random graph model for network data with an underlying spatial representation. Geometry endows RGGs with a rich dependence structure and often leads to desirable properties of real-world networks such as the small-world phenomenon and clustering. Originally introduced to model wireless communication networks, RGGs are now very popular with applications ranging from network user profiling to protein-protein interactions in biology. RGGs are also of purely theoretical interest since the underlying geometry gives rise to challenging mathematical questions. Their resolutions involve results from probability, statistics, combinatorics or information theory, placing RGGs at the intersection of a large span of research communities.
This paper surveys the recent developments in RGGs from the lens of high dimensional settings and non-parametric inference. We also explain how this model differs from classical community based random graph models and we review recent works that try to take the best of both worlds. As a by-product, we expose the scope of the mathematical tools used in the proofs.

# 1 Introduction

## 1.1 Random graph models

Graphs are nowadays widely used in applications to model real world complex systems. Since they are high dimensional objects, one needs to assume some structure on the data of interest to be able to efficiently extract information on the studied system. To this purpose, a large number of models of random graphs have been already introduced. The most simple one is the Erdös-Renyi model $G(n,p)$ in which each edge between pairs of $n$ nodes is present in the graph with some probability $p \in (0,1)$. One can also mention the scale-free network model of Barabasi and Albert (Barabási, 2009) or the small-world networks of Watts and Strogatz (Watts and Strogatz, 1998). We refer to Channarond (2015) for an introduction to the most famous random graph models. On real world problems, it appears that there often exist some relevant variables accounting for the heterogeneity of the observations. Most of the time, these explanatory variables are unknown and carry a precious information on the system studied. To deal with such cases, latent space models for network data emerged (see Smith et al. (2019)). Ones of the most studied latent models are the *community based random graphs* where each node is assumed to belong to one (or multiple) community while the connection probabilities between two nodes in the graph depend on their respective membership. The well-known Stochastic Block Model has received increasing attention in the recent years and we refer to Abbe (2018) for a nice introduction to this model and the statistical and algorithmic questions at stake. In the previous mentioned latent space models the intrinsic geometry of the problem is not taken into account. However, it is known that the underlying spatial structure of network is an important property since geometry affects drastically the topology of networks (see Barthélemy (2011) and Smith et al. (2019)). To deal with embedded complex systems, spatial random graph models have been studied such as the Random Geometric Graph (RGG). This paper surveys the recent developments in the theoretical analysis of RGGs through the prism of modern statistics and applications.

The theoretical analysis of random graph models is interesting by itself since it often involves elegant and important information theoretic, combinatorial or probabilistic tools. In the following, we adopt this mindset trying to provide a faithful picture of the state of the art results on RGGs focusing mainly on high dimensional settings and non-parametric inference while underlining the main technical tools used in the proofs. We want to illustrate how the theory can impact real data applications. To this end, we will essentially be focused on the following questions:

- **Detecting Geometry in RGGs.** Nowadays real world problems often involve high-dimensional feature spaces. A first natural work is to identify the regimes where the geometry is lost in the dimension (see Eq.(1) for a formal definition). Several recent papers have made significant progress towards the resolution of this question that can be formalized as follows. Given a graph of $n$ nodes, a latent geometry of dimension $d = d(n)$ and edge density $p = p(n)$, for what triples $(n, d, p)$ is the model indistinguishable from $G(n, p)$?

- **Non-parametric estimation in RGGs.** By considering other rules for connecting latent points, the RGG model can be naturally extended to cover a larger class of networks. In such a framework, we will wonder what can be learned in an adaptive way from graphs with an underlying spatial structure. We will address non-parametric estimation in RGGs and its extension to growth model.

- **Connections between RGGs and community based latent models.** Until recently, community and geometric based random graph models have been mainly studied separately. Recent works try to investigate graph models that account for both cluster and spatial structures. We present some of them and we sketch interesting research directions for future works.

## 1.2 Brief historical overview of RGGs

The RGG model was first introduced by Gilbert (1961) to model the communications between radio stations. Gilbert's original model was defined as follows: pick points in $\mathbb{R}^2$ according to a Poisson Point Process of intensity one and join two if their distance is less than some parameter $r > 0$. The Gilbert model has been intensively studied and we refer to Walters (2011) for a nice survey of its properties including connectivity, giant component, coverage or chromatic number. The most closely related model is the Random Geometric Graph where $n$ nodes are independently and identically distributed on the space. A lot of results are actually transferable from one model to the other as presented in (Penrose et al., 2003, Section 1.7). In this paper we will focus on the $n$ points i.i.d. model which is formally defined in the next subsection (see Definition 1). The Random Geometric Graph model was extended to other latent spaces such as the hypercube $[0, 1]^d$, the Euclidean sphere or compact Lie group Méliot (2019). A large body of literature has been devoted to studying the properties of low-dimensional Random Geometric Graphs Penrose et al. (2003), Dall and Christensen (2002), Bollobás (2001). RGGs have found applications in a very large span of fields. One can mention wireless networks Haenggi et al. (2009), Mao and Anderson (2012), gossip algorithms Wang and Lin (2014), consensus Estrada and Sheerin (2016), spread of a virus Preciado and Jadbabaie (2009), protein-protein interactions Higham et al. (2008), citation networks Xie et al. (2016). One can also cite an application to motion planning in Solovey et al. (2018), a problem which consists in finding a collision-free path for a robot in a workspace cluttered with static obstacles. The ubiquity of this random graph model to faithfully represent real world networks has motivated a great interest for its theoretical study.

## 1.3 Outline

In Section 2, we formally define the RGG and several variant models that will be useful for this article. In Sections 3, 4 and 5, we describe recent results related to high-dimensional statistic, non-parametric estimation and temporal prediction. Note that in these three sections, we will be working with the $d$-dimensional sphere $\mathbb{S}^{d-1}$ as latent space. $\mathbb{S}^{d-1}$ will be endowed with the Euclidean metric $\|\cdot\|$ which is the norm induced by the inner product $\langle\cdot,\cdot\rangle : (x, y) \in \left(\mathbb{S}^{d-1}\right)^2 \mapsto \sum_{i=1}^{d} x_i y_i$. The choice of this latent space is motivated by both recent theoretical developments in this framework Bubeck et al. (2016), De Castro et al. (2020), Allen-Perkins (2018), Issartel et al. (2021) and by applications Pereda and Estrada (2019), Perry et al. (2020). We further discuss in Section 6 recent works that investigate the connections between

community based random graph models and RGGs. Contrary to the previous sections, our goal is not to provide an exhaustive review of the literature in Section 6 but rather to shed light on some pioneering papers.

| Section | Questions tackled | Model |
|---|---|---|
| 3 | Geometry detection | RGG on $\mathbb{S}^{d-1}$ |
| 4 | Non-parametric estimation | TIRGG on $\mathbb{S}^{d-1}$ |
| 5 | Non-parametric estimation & Temporal prediction | MRGG on $\mathbb{S}^{d-1}$ |
| 6 | Connections between community based models and RGGs | |

Table 1: Outline of the paper. Models are defined in Section 2.

## 2 The Random Geometric Graph model and its variants

The questions that we tackle here can require some additional structure on the model. In this section, we define the variants of the RGG that will be useful for our purpose. Figure 2 shows the connections between these different models.

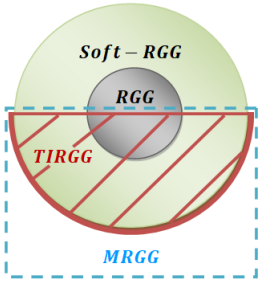


Figure 1: Venn diagram of the different random graph models.

### 2.1 (Soft-) Random Geometric Graphs

**Definition 1.** *(Random Geometric Graph: RGG)*
*Let $(\mathcal{X},\rho)$ be a metric space, and $m$ be a Borel probability measure on $\mathcal{X}$. Given a positive real number $r>0$, the Random Geometric Graph with $n\in\mathbb{N}\backslash\{0\}$ points and level $r>0$ is the random graph $G$ such that*

- *the $n$ vertices $X_1,\dots,X_n$ of $G$ are chosen randomly in $\mathcal{X}$ according to the probability measure $m^{\otimes n}$ on $\mathcal{X}^n$.*
- *for any $i,j\in[n]$ with $i\neq j$, an edge between $X_i$ and $X_j$ is present in $G$ if and only if $\rho(X_i,X_j)\leq r$.*

*We denote* $\mathrm{RGG}(n,m,(\mathcal{X},\rho))$ *the distribution of such random graphs.*

Motivated by wireless *ad hoc* networks, Soft-RGGs have been more recently introduced (see Penrose (2016)). In such models, we are given some function $H:\mathbb{R}_+\to[0,1]$ and two nodes at distance $\rho$ in the graph are connected with probability $H(\rho)$.

**Definition 2.** *(Soft Random Geometric Graph: Soft-RGG)*
*Let $(\mathcal{X},\rho)$ be a metric space, $m$ be a Borel probability measure on $\mathcal{X}$ and consider some function $H:\mathbb{R}_+\to[0,1]$. The Soft (or probabilistic) Random Geometric Graph with $n\in\mathbb{N}\backslash\{0\}$ points with connection function $H$ is the random graph $G$ such that*

- *the $n$ vertices $X_1,\dots,X_n$ of $G$ are chosen randomly in $\mathcal{X}$ according to the probability measure $m^{\otimes n}$ on $\mathcal{X}^n$.*
- *for any $i,j\in[n]$ with $i\neq j$, we draw an edge between nodes $X_i$ and $X_j$ with probability $H\left(\rho(X_i,X_j)\right)$.*

*We denote Soft-RGG$(n,m,(\mathcal{X},\rho))$ the distribution of such random graphs.*

Note that the RGG model with level $r>0$ is a particular case of the Soft-RGG model where the connection function $H$ is chosen as $\rho\mapsto\mathbb{1}_{\rho\leq r}$. The obvious next special case to consider of Soft-RGG is the so-called percolated RGG introduced in Müller and Prałat (2015) which is obtained by retaining each

edge of a RGG of level $r > 0$ with probability $p \in (0,1)$ (and discarding it with probability $1-p$). This reduces to consider the connection function $H : \rho \mapsto p \times \mathbb{1}_{\rho \leq r}$. Particular common choices of connection function are the *Reyleigh fading* activation functions which take the form

$$H^{Rayleigh}(\rho) = \exp\left[-\zeta\left(\frac{\rho}{r}\right)^{\eta}\right], \quad \zeta > 0, \eta > 0.$$

We refer to Dettmann and Georgiou (2016) and references therein for a nice overview of Soft-RGGs in particular the most classical connection functions and the question of connectivity in the resulting graphs.

## 2.2 Translation Invariant Random Geometric Graphs

One possible non-parametric generalization of the (Soft)-RGG model is given by the $W$ random graph model (see for example Diaconis and Janson (2007)) based on the notion of graphon. In this model, given latent points $x_1, \dots, x_n$ uniformly and independently sampled in $[0,1]$, the probability to draw an edge between $i$ and $j$ is $\Theta_{i,j} := W(x_i, x_j)$ where $W$ is a symmetric function from $[0,1]^2$ onto $[0,1]$, referred to as a graphon. Hence, the adjacency matrix $A$ of this graph satisfies

$$\forall i, j \in [n], \quad A_{i,j} \sim \mathrm{Ber}(\Theta_{i,j}),$$

where for any $p \in [0,1]$, $\mathrm{Ber}(p)$ is the Bernoulli distribution with parameter $p$.
**Remark.** Let us point out that graphon models can also be defined by replacing the latent space $[0,1]$ by the Euclidean sphere $\mathbb{S}^{d-1} := \{x \in \mathbb{R}^d \,|\, \|x\|_2 = 1\}$ in which case latent points are sampled independently and uniformly on $\mathbb{S}^{d-1}$.
This model has been widely studied in the literature (see Lovász (2012)) and it is now well-known that, by construction, graphons are defined on an equivalent class *up to a measure preserving homomorphism*. More precisely, two graphons $U$ and $W$ define the same probability distribution if and only if there exist measure preserving maps $\varphi, \psi : [0,1] \to [0,1]$ such that $U(\varphi(x), \varphi(y)) = W(\psi(x), \psi(y))$ almost everywhere. Hence it can be challenging to have a simple description from an observation given by sampled graph—since one has to deal with all possible composition of a bivariate function by any measure preserving homomorphism. Such difficulty arises in Wolfe and Olhede (2013) or in Klopp and Verzelen (2019) that use respectively Maximum Likelihood and least-square estimators to approximate the graphon $W$ from the adjacency matrix $A$. In those works, the error measures are based on the so-called *cut-distance* that is defined as an infimum over all measure-preserving transformations. This statistical issue motivates the introduction of (Soft)-RGGs with latent metric spaces for which the distance is invariant by translation (or conjugation) of pairs of points. This natural assumption leads to consider that the latent space has some group structure, namely it is a compact Lie group or some compact symmetric space.

**Definition 3.** *(Translation Invariant Random Geometric Graph: TIRGG)*
*Let $(S, \gamma)$ be a compact Lie group with an invariant Riemannian metric $\gamma$ normalized so that the range of $\gamma$ equals $[0, \pi]$. Let $m$ be the uniform probability measure on $S$ and let us consider some map $\mathbf{p} : [-1,1] \to [0,1]$, called the envelope function. The Translation Invariant Random Geometric Graph with $n \in \mathbb{N}\backslash\{0\}$ points is the random graph $G$ such that*

- *the $n$ vertices $X_1, \dots, X_n$ of $G$ are chosen randomly in $S$ according to the probability measure $m^{\otimes n}$ on $S^n$.*
- *for any $i, j \in [n]$ with $i \neq j$, we draw an edge between nodes $X_i$ and $X_j$ with probability $\mathbf{p}\big(\cos\gamma(X_i, X_j)\big)$.*

In Section 4, we present recent results regarding non-parametric estimation in the TIRGG model with $S := \mathbb{S}^{d-1}$ the Euclidean sphere of dimension $d$ from the observation of the adjacency matrix. A related question was addressed in Klopp and Verzelen (2019) where the authors derived sharp rates of convergence for the $L^2$ loss for the Stochastic Block Model (which belongs to the class of graphon models). Let us point out that a general approach to control the $L^2$ loss between the probability matrix and a eigenvalue-tresholded version of the adjacency matrix is the USVT method introduced by Chatterjee (2015), which was further investigated by Xu (2018). In Section 4, another line of work is presented to estimate the envelope function $\mathbf{p}$ where the difference between the adjacency matrix and the *matrix of probabilities* $\Theta$ is controlled in operator norm. The cornerstone of the proof is the convergence of the

spectrum of the matrix of probabilities towards the spectrum of some integral operator associated with the envelope function $\mathbf{p}$. Based on the analysis of Koltchinskii and Giné (2000), the proof of this convergence includes in particular matrix Bernstein inequality from Tropp (2015) and concentration inequality for order 2 U-statistics with bounded kernels that was first studied by Arcones and Gine (1993) and remains an active field of research (see Giné et al. (2000), Houdré and Reynaud-Bouret (2002) or Joly and Lugosi (2016)).

## 2.3 Markov Random Geometric Graphs

In the following, we will refer to *growth models* to denote random graph models in which a node is added at each new time step in the network and is connected to other vertices in the graph according to some probabilistic rule that needs to be specified. In the last decade, growth models for random graphs with a spatial structure have gained an increased interest. One can mention Jordan and Wade (2015), Papadopoulos et al. (2012) and Zuev et al. (2015) where geometric variants of the preferential attachment model are introduced with one new node entering the graph at each time step. More recently, Xie et al. (2015) and Xie and Rogers (2016) studied a growing variant of the RGG model. Note that in the latter works, the birth time of each node is used in the connection function while nodes are still sampled independently in $\mathbb{R}^2$. Still motivated by non-parametric estimation, the TIRGG model can be extended to a growth model by considering a Markovian sampling scheme of the latent positions. Considering a Markovian latent dynamic can be relevant to model customer behavior for item recommendation or to study bird migrations where animals have regular seasonal movement between breeding and wintering grounds (cf. Duchemin, 2022).

**Definition 4.** *(Markov Random Geometric Graph: MRGG)*
*Let $(S,\gamma)$ be a compact Lie group with an invariant Riemannian metric $\gamma$ normalized so that the range of $\gamma$ equals $[0,\pi]$. Let us consider some map $\mathbf{p}:[-1,1]\to[0,1]$, called the envelope function. The Markov Random Geometric Graph with $n\in\mathbb{N}\backslash\{0\}$ points is the random graph $G$ such that*

- *the sequence of $n$ vertices $(X_1,\dots,X_n)$ of $G$ is a Markov chain on $S$.*
- *for any $i,j\in[n]$ with $i\neq j$, we draw an edge between nodes $X_i$ and $X_j$ with probability $\mathbf{p}\big(\cos\gamma(X_i,X_j)\big)$.*

In Section 5, we shed light on a recent work from Duchemin and De Castro (2022) that achieves non-parametric estimation in MRGGs on the Euclidean sphere of dimension $d$. The theoretical study of such graphs becomes more challenging because of the dependence induced by the latent Markovian dynamic. Proving the consistency of the non-parametric estimator of the envelope function $\mathbf{p}$ proposed in Section 5 requires in particular a new concentration inequality for U-statistics of order 2 of uniformly ergodic Markov chains. By solving link prediction problems, Duchemin and De Castro (2022) also reveal that MRGGs are convenient tools to extract temporal information on growing graphs with an underlying spatial structure.

## 2.4 Other model variants

**Choice of the metric space.** The Euclidean Sphere or the unit square in $\mathbb{R}^d$ are the most studied latent spaces in the literature for RGGs. By the way, Allen-Perkins (2018) offers an interesting comparison of the different topological properties of RGGs working on one or the other of these two spaces. Nevertheless, one can find variants such as in Araya Valdivia (2020) where Euclidean Balls are considered. More recently, some researchers left the Euclidean case to consider negatively curved–i.e. hyperbolic–latent spaces. Random graphs with an hyperbolic latent space seem promising to faithfully model real world networks. Actually, Krioukov et al. (2010) showed that the RGG built on the hyperbolic geometry is a scale-free network, that is the proportion of node of degree $k$ is of order $k^{-\gamma}$ where $\gamma$ is between 2 and 3. The scale-free property is found in the most part of real networks as highlighted by Xie et al. (2015).

**Different degree distributions.** It is now well-known that the average degree of nodes in random graph models is a key property for their statistical analysis. Let us highlight some important regimes in the random graph community that will be useful in this paper. The dense regime corresponds to the case where the expected normalized degree of the nodes (*i.e.,* degree divided by $n$) is independent of the

number of nodes in the graph. The other two important regimes are the relatively sparse and the sparse regimes where the average degree of nodes scales respectively as $\log(n)/n$ and $1/n$ with the number of nodes $n$. A direct and important consequence of these definitions is that in the (relatively) sparse regime, the envelope function $\mathbf{p}$ from Definitions 3 and 4 depends on $n$ while it remains independent of $n$ in the dense regime. Similarly, in the (relatively) sparse regime, the radius threshold $r$ from Definition 1 (resp. the connection function $H$ from Definition 2) depend on $n$ contrary to the dense regime.

# 3 Detecting geometry in RGGs

To quote Bollobás (2001), "*One of the main aims of the theory of random graphs is to determine when a given property is likely to appear.*" In this direction, several works tried to identify structure in networks through testing procedure, see for example Bresler and Nagaraj (2018), Ghoshdastidar et al. (2020) or Gao and Lafferty (2017). Regarding RGGs, most of the results have been established in the low dimensional regime $d \leq 3$ Ostilli and Bianconi (2015), Penrose (2016), Penrose et al. (2003), Barthélemy (2011). Goel et al. (2005) proved in particular that all monotone graph properties (i.e. property preserved when adding edges to the graph) have a sharp threshold for RGGs that can be distinguished from the one of Erdös-Rényi random graphs in low-dimensions. However, applications of RGGs to cluster analysis and the interest in the statistics of high-dimensional data sets have motivated the community to investigate the properties of RGGs in the case where $d \to \infty$. If the ambitious problem of recognizing if a graph can be realized as a geometric graph is known to be NP-hard Breu and Kirkpatrick (1998), one can take a step back and wonder if a given RGG still carries some spatial information as $d \to \infty$ or if geometry is lost in high-dimensions (see Eq.(1) for a formal definition), a problem known as geometry detection. In the following, we present some recent results related to geometry detection in RGGs with latent space the Euclidean sphere $\mathbb{S}^{d-1}$ and we highlight several interesting directions for future research.

**Notations** Given two sequences $(a_n)_{n\in\mathbb{N}}$ and $(b_n)_{n\in\mathbb{N}}$ of positive numbers, we write $a_n = \mathcal{O}_n(b_n)$ or $b_n = \Omega_n(a_n)$ if the sequence $(a_n/b_n)_{n\geq 0}$ is bounded and we write $a_n = o_n(b_n)$ or $b_n = \omega_n(a_n)$ if $a_n/b_n \underset{n\to+\infty}{\to} 0$. We further denote $a_n = \Theta(b_n)$ if $a_n = \mathcal{O}_n(b_n)$ and $b_n = \mathcal{O}_n(a_n)$. In the following, we will denote $G(n,p,d)$ the distribution of random graphs of size $n$ where nodes $X_1,\dots,X_n$ are sampled uniformly on $\mathbb{S}^{d-1}$ and where distinct vertices $i \in [n]$ and $j \in [n]$ are connected by an edge if and only if $\langle X_i, X_j\rangle \geq t_{p,d}$. The threshold value $t_{p,d} \in [-1,1]$ is such that $\mathbb{P}\left(\langle X_1, X_2\rangle \geq t_{p,d}\right) = p$. Note that $G(n,p,d)$ is the distribution of RGGs on $(\mathbb{S}^{d-1}, \|\cdot\|)$ sampling nodes uniformly with connection function $H : t \mapsto \mathbb{1}_{t\leq\sqrt{2-2t_{p,d}}}$. In the following, we will also use the notation $G(n,d,p)$ to denote a graph sampled from this distribution. We also introduce some definitions of standard information theoretic tools with Definition 5.

**Definition 5.** *Let us consider two probability measures $P$ and $Q$ defined on some measurable space $(E,\mathscr{E})$. The total variation distance between $P$ and $Q$ is given by*

$$\mathrm{TV}(P,Q) := \sup_{A\in\mathscr{E}} |P(A) - Q(A)|.$$

*Assuming further that $P \ll Q$ and denoting $dP/dQ$ the density of $P$ with respect to $Q$,*

- *the $\chi^2$-divergence between $P$ and $Q$ is defined by*

$$\chi^2(P,Q) := \int_E \left(\frac{dP}{dQ} - 1\right)^2 dQ.$$

- *the Kullback-Leibler divergence between $P$ and $Q$ is defined by*

$$\mathrm{KL}(P,Q) := \int_E \log\left(\frac{dP}{dQ}\right) dP.$$

Considering that both $p$ and $d$ depend on $n$, we will say in this paper that *geometry is lost* if the distributions $G(n,p)$ and $G(n,p,d)$ are indistinguishable, namely if

$$\mathrm{TV}(G(n,p), G(n,p,d)) \to 0 \quad \text{as } n \to \infty. \tag{1}$$

## 3.1 Detecting geometry in the dense regime

Devroye et al. (2011) is the first to consider the case where $d \to \infty$ in RGGs. In this paper, the authors proved that the number of cliques in the dense regime in $G(n,p,d)$ is close to the one of $G(n,p)$ provided $d \gg \log n$ in the asymptotic $d \to \infty$. This work allowed them to show the convergence of the total variation (see Definition 5) between RGGs and Erdös-Renyi graphs as $d \to \infty$ for fixed $p$ and $n$. Bubeck et al. (2016) closed the question of geometry detection in RGGs in the dense regime showing that a phase transition occurs when $d$ scales as $n^3$ as stated by Theorem 1.

**Theorem 1.** *(Bubeck et al., 2016, Theorem 1)*

*(i) Let $p \in (0,1)$ be fixed, and assume that $d/n^3 \to 0$. Then*

$$\mathrm{TV}(G(n,p), G(n,p,d)) \to 1 \quad \text{as } n \to \infty.$$

*(ii) Furthermore, if $d/n^3 \to \infty$, then*

$$\sup_{p\in(0,1)} \mathrm{TV}(G(n,p), G(n,p,d)) \to 0 \quad \text{as } n \to \infty.$$

The proof of Theorem.1.$(i)$ relies on a count of *signed* triangles in RGGs. Denoting $A$ the adjacency matrix of the RGG, the number of triangles in $A$ is $\mathrm{Tr}(A^3)$, while the total number of signed triangles is defined as

$$\tau(G(n,p,d)) := \mathrm{Tr}((A-p(J-I))^3) = \sum_{\{i,j,k\}\in\binom{[n]}{3}} \big(A_{i,j}-p\big)\big(A_{i,k}-p\big)\big(A_{j,k}-p\big),$$

where $I$ is the identity matrix and $J \in \mathbb{R}^{n\times n}$ is the matrix with every entry equals to 1. The analogous quantity in Erdös Renyi graphs $\tau(G(n,p))$ is defined similarly. Bubeck et al. (2016) showed that the variance of $\tau(G(n,p,d))$ is of order $n^3$ while the one of the number of triangles is of order $n^4$. This smaller variance for signed triangles is due to the cancellations introduced by the centering of the adjacency matrix. Lemma 1 provides the precise bounds obtained on the expectation and the variance of the statistic of signed triangles. Theorem.1.$(i)$ follows from the lower-bounds (resp. the upper-bounds) on the expectations (resp. the variances) of $\tau(G(n,p))$ and $\tau(G(n,p,d))$ presented in Lemma 1.

**Lemma 1.** *(Bubeck et al., 2016, Section 3.4) For any $p \in (0,1)$ and any $n, d \in \mathbb{N}\backslash\{0\}$ it holds*

$$\mathbb{E}[\tau(G(n,p))] = 0, \quad \mathbb{E}[\tau(G(n,p,d))] \geq \binom{n}{3}\frac{C_p}{\sqrt{d}}$$
$$\text{and} \quad \max\{\mathbb{V}ar[\tau(G(n,p))], \mathbb{V}ar[\tau(G(n,p,d))]\} \leq n^3 + \frac{3n^4}{d},$$

*where $C_p > 0$ is a constant depending only on $p$.*

Let us now give an overview of the proof of the indistinguishable part of Theorem 1. Bubeck et al. (2016) proved that in the dense regime, the phase transition for geometry detection occurs at the regime at which Wishart matrices becomes indistinguishable from GOEs (Gaussian Orthogonal Ensemble). In the following, we draw explicitly this link in the case $p = 1/2$.
An $n \times n$ Wishart matrix with $d$ degrees of freedom is a matrix of inner products of $n$ $d$-dimensional Gaussian vectors denoted by $W(n,d)$ while an $n \times n$ GOE random matrix is a symmetric matrix with i.i.d. Gaussian entries on and above the diagonal denoted by $M(n)$. Let $\mathbb{X}$ be an $n \times d$ matrix where the entries are i.i.d. standard normal random variables, and let $W = W(n,d) = \mathbb{X}\mathbb{X}^\top$ be the corresponding $n \times n$ Wishart matrix. Then recalling that for $X_1 \sim \mathcal{N}(0, I_d)$ a standard gaussian vector of dimension $d$, $X_1/\|X_1\|_2$ is uniformly distributed on the sphere $\mathbb{S}^{d-1}$, we get that the $n \times n$ matrix $A$ defined by

$$\forall i,j \in [n], \quad A_{i,j} = \begin{cases} 1 & \text{if } W_{i,j} \geq 0 \text{ and } i \neq j \\ 0 & \text{otherwise.} \end{cases}$$

has the same distribution as the adjacency matrix of a graph sampled from $G(n,1/2,d)$. We denote $H$ the map that takes $W$ to $A$. Analogously, one can prove that $G(n,1/2)$ can be seen as a function of an $n \times n$

GOE matrix. Let $M(n)$ be a symmetric $n \times n$ random matrix where the diagonal entries are i.i.d. normal random variables with mean zero and variance 2, and the entries above the diagonal are i.i.d.standard normal random variables, with the entries on and above the diagonal all independent. Then $B = H(M(n))$ is distributed as the adjacency matrix of $G(n, 1/2)$. We then get

$$\mathrm{TV}(G(n,1/2,d), G(n,1/2)) = \mathrm{TV}(H(W(n,d)), H(M(n))) \leq \mathrm{TV}(W(n,d)), M(n)). \tag{2}$$

If a simple application of the multivariate Central Limit Theorem proves that the right hand side of (2) goes to zero as $d \to \infty$ for fixed $n$, more work is necessary to address the case where $d = d(n) = \omega_n(n^3)$ and $n \to \infty$. The distributions of $W(n,d)$ and $M(n)$ are known and allow explicit computations leading to Theorem 2. This proof can be adapted for any $p \in (0,1)$ leading to Theorem 1.$(ii)$ from (2).

**Theorem 2.** *(Bubeck et al., 2016, Theorem 7)*
*Define the random matrix ensembles $W(n,d)$ and $M(n)$ as above. If $d/n^3 \to \infty$, then*

$$\mathrm{TV}(W(n,d), M(n)) \to 0.$$

**Extensions** Considering $\mathbb{R}^d$ as latent space endowed with the Euclidean metric, Bubeck and Ganguly (2015) extended Theorem 2 and proved an information theoretic phase transition. To give an overview of their result, let us consider the $n \times n$ Wigner matrix $\mathscr{M}_n$ with zeros on the diagonal and i.i.d. standard Gaussians above the diagonal. For some $n \times d$ matrix $\mathbb{X}$ with i.i.d. entries from a distribution $\mu$ on $\mathbb{R}^d$ that has mean zero and variance 1, we also consider the following rescaled Wishart matrix associated with $\mathbb{X}$

$$\mathscr{W}_{n,d} := \frac{1}{\sqrt{d}} \left( \mathbb{X}\mathbb{X}^\top - \mathrm{diag}(\mathbb{X}\mathbb{X}^\top) \right),$$

where the diagonal was removed. Using an high-dimensional entropic Central Limit Theorem, Bubeck and Ganguly (2015) proved Theorem 3 which implies that geometry is lost in $RGG(n, \mu, (\mathbb{R}^d, \|\cdot\|))$ as soon as $d \gg n^3 \log^2(d)$ provided that the measure $\mu$ is sufficiently smooth (namely log-concave) and the rate is tight up to logarithmic factors. We refer to Racz and Bubeck (2016) for a friendly presentation of this result. Note that the comparison between Wishart and GOE matrices also naturally arise when dealing with covariance matrices. For example, Theorem 2 was used in Brennan and Bresler (2019) to study the informational-computational tradeoff of sparse Pincipal Component Analysis.

**Theorem 3.** *(Bubeck and Ganguly, 2015, Theorem 1)*
*If the distribution $\mu$ is log-concave and $\frac{d}{n^3 \log^2(d)} \to \infty$, then* $\mathrm{TV}(\mathscr{W}_{n,d}, \mathscr{M}_n) \to 0$.
*On the other hand, if $\mu$ has a finite fourth moment and $\frac{d}{n^3} \to 0$, then* $\mathrm{TV}(\mathscr{W}_{n,d}, \mathscr{M}_n) \to 1$.

## 3.2 Failure to extend the proof techniques to the sparse regime

Bubeck et al. (2016) provided a result in the sparse regime where $p = \frac{c}{n}$ showing that one can distinguish between $G(n, \frac{c}{n})$ and $G(n, \frac{c}{n}, d)$ as long as $d \ll \log^3 n$. The authors conjectured that this rate is tight for the sparse regime (see Conjecture 1).

**Conjecture 1.** *(Bubeck et al., 2016, Conjecture 1)*
*Let $c > 0$ be fixed, and assume that $d/\log^3(n) \to \infty$. Then*

$$\mathrm{TV}\left( G\left(n, \frac{c}{n}\right), G\left(n, \frac{c}{n}, d\right) \right) \to 0 \quad \text{as } n \to \infty.$$

The testing procedure from Bubeck et al. (2016) to prove the distinguishability result in the sparse regime was based on a simple counting of triangles. Indeed, when $p$ scales as $\frac{1}{n}$, the signed triangle statistic $\tau$ does not give significantly more power than the triangle statistic which simply counts the number of triangles in the graph. Recently, Avrachenkov and Bobu (2020) provided interesting results that give credit to Conjecture 1. First, they proved that in the sparse regime, the clique number of $G(n,p,d)$ is almost surely at most 3 under the condition $d \gg \log^{1+\varepsilon} n$ for any $\varepsilon > 0$. This means that in the sparse regime, $G(n,p,d)$ does not contain any complete subgraph larger than a triangle like Erdös-Renyi graphs. Hence it is hopeless to prove that Conjecture 1 is false considering the number of $k$-cliques

for $k \geq 4$. Nevertheless, one could believe that improving the work of Bubeck et al. (2016) by deriving sharper bounds on the number of 3-cliques (i.e. the number of triangles), it could possible to statistically distinguish between $G(n,p,d)$ and $G(n,p)$ in the sparse regime even for some $d \gg \log^3 n$. In a regime that can be made arbitrarily close to the sparse one, Avrachenkov and Bobu (2020) proved that this is impossible as stated by Theorem 4.

**Theorem 4.** *(Avrachenkov and Bobu, 2020, Theorem 5)*
*Let us suppose that $d \gg \log^3 n$ and $p = \theta(n)/n$ with $n^m \leq \theta(n) \ll n$ for some $m > 0$. Then the expected number of triangles–denoted $\mathbb{E}[T(n,p,d)]$–in RGGs sampled from $G(n,p,d)$ is of order $\binom{n}{3}p^3$, meaning that there exist two universal constants $c, C > 0$ such that for $n$ large enough it holds*

$$c\binom{n}{3}p^3 \leq \mathbb{E}[T(n,p,d)] \leq C\binom{n}{3}p^3.$$

In a nutshell, the work from Avrachenkov and Bobu (2020) suggests that a negative result regarding Conjecture 1 cannot be obtained using statistics based on clique numbers. This discussion naturally gives rise to the following more general question.

*Given a random graph model with $n$ nodes, latent geometry in dimension $d = d(n)$ and edge density $p = p(n)$, for what triples $(n,d,p)$ is the model $G(n,p,d)$ indistinguishable from $G(n,p)$?* (Q)

## 3.3 Towards the resolution of geometry detection

### 3.3.1 A first improvement when $d > n$

A recent work from Brennan et al. (2020) tackled the general problem (Q) and proved Theorem 5.

**Theorem 5.** *(Brennan et al., 2020, Theorem 2.4)*
*Suppose $p = p(n) \in (0, 1/2]$ satisfies that $n^{-2}\log n = \mathcal{O}_n(p)$ and*

$$d \gg \min\left\{pn^3 \log p^{-1}\,,\; p^2 n^{7/2} (\log n)^3 \sqrt{\log p^{-1}}\right\},$$

*where $d$ also satisfies that $d \gg n\log^4 n$. Then*

$$\mathrm{TV}(G(n,p), G(n,p,d)) \to 0 \quad \text{as } n \to \infty.$$

**Remarks** In the dense regime, Theorem 5 recovers the optimal guarantee from Theorem 1. In the sparse regime, Theorem 5 states that if $d \gg n^{3/2}(\log n)^{7/2}$, then geometry is lost in $G(n, \frac{c}{n}, d)$ (where $c > 0$). This result improves the work from Bubeck et al. (2016). Nevertheless, regarding Conjecture 1, it remains a large gap between the rates $\log^3 n$ and $n^{3/2}(\log n)^{7/2}$ where nothing is known up to date. Let us sketch the main elements of the proof of Theorem 5. In the following we denote $G = G(n,p,d)$ with set of edges $E(G)$ and for any $i, j \in [n]$, $i \neq j$, we denote $G_{\sim\{i,j\}}$ the set of edges other than $\{i,j\}$ in $G$. One first important step of their approach is the following tenzorization Lemma for the Kullback-Leibler divergence.

**Lemma 2.** *(Kontorovich and Raginsky, 2017, Lemma 3.4)*
*Let us consider $(X, \mathcal{B})$ a measurable space with $X$ a Polish space and $\mathcal{B}$ its Borel $\sigma$-field. Consider some probability measure $\mu$ on the product space $X^k$ with $\mu = \mu_1 \otimes \mu_2 \otimes \cdots \otimes \mu_k$. Then for any other probability measure $\nu$ on $X^k$ it holds*

$$\mathrm{KL}(\nu||\mu) \leq \sum_{i=1}^{k} \mathbb{E}_{x\sim\nu}\left[\mathrm{KL}\left(\nu_i(\cdot|x_{\sim i})||\mu_i\right)\right],$$

*where $\nu_i$ is the probability distribution corresponding to the $i$-th marginal of $\nu$ and where $x_{\sim i} := (x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_k)$.*

$$\begin{aligned}
2\mathrm{TV}(G(n,p,d), G(n,p))^2 &\leq \mathrm{KL}(G(n,p,d)||G(n,p)) \qquad \text{from Pinsker's inequality} \\
&\leq \sum_{1\leq i<j\leq n} \mathbb{E}\left[\mathrm{KL}\left(\mathscr{L}(\mathbb{1}_{\{i,j\}\in E(G)}|\sigma(G_{\sim\{i,j\}}))||\mathrm{Bern}(p)\right)\right] \qquad \text{from Lemma 2} \\
&\leq \binom{n}{2} \times \mathbb{E}\left[\chi^2\left(\mathscr{L}(\mathbb{1}_{e_0\in E(G)}|\sigma(G_{\sim e_0})), \mathrm{Bern}(p)\right)\right] \\
&= \binom{n}{2} \times \mathbb{E}\left[\frac{(Q-p)^2}{p(1-p)}\right],
\end{aligned}$$

where $Q := \mathbb{P}\left(e_0 \in E(G) | G_{\sim e_0}\right)$ is a $\sigma(G_{\sim e_0})$-measurable random variable corresponding to the probability that a specific edge is included in the graph given the rest of the graph. The proof then consists in showing that with high probability, $Q$ concentrates near $p$. To do so, they use a coupling argument that gives an alternative way to generate $X_1$ that provides a direct description of $\mathbb{1}_{e_0 \in E(G)}$ in terms of the random variables introduced in the coupling. If this step may seem computationally involved, it is not conceptually difficult since it turns out to be a simple re-parametrization of the problem. An integration of this concentration result for $Q$ implies that the convergence of Theorem 5 holds when $d \gg pn^3 \log p^{-1}$. To get the convergence result in the regime where $d \gg p^2 n^{7/2} (\log n)^3 \sqrt{\log p^{-1}}$ — which gives the improvement over Bubeck et al. (2016) in the sparse case — one additional step of coupling is required. More precisely, they decompose $\mathbb{E}[(Q-p)^2]$ as $\mathbb{E}[(Q-p) \times (Q-p)]$. The previous coupling argument gives a concentration inequality allowing to bound with high probability the first term $|Q-p|$. It remains then to upper bound $\mathbb{E}[|Q-p|]$ which relies on a simple observation given by the following proposition.

**Proposition 1.** *(Brennan et al., 2020, Proposition 5.3) Let $\nu_{\sim e_0}$ denote the marginal distribution of $G$ restricted to all edges that are not $e_0$, and let $\nu^+_{\sim e_0}$ denote the distribution of $G$ conditioned on the event $e_0 \in E(G)$. It holds*

$$\mathbb{E}[|Q-p|] = 2p \times \mathrm{TV}\left(\nu^+_{\sim e_0}, \nu_{\sim e_0}\right). \tag{3}$$

The proof is then concluded by using another coupling argument between $\nu^+_{\sim e_0}$ and $\nu_{\sim e_0}$ to upper-bound the total variation distance involved in Eq.(3) and we give a sketch of proof in the following. Given latent positions $X_1, \dots, X_n$ uniformly and independently sampled on $\mathbb{S}^{d-1}$, we can consider without loss of generality that $X_1 = (1, 0, \dots 0)$. Denoting $X_2 = (X_{2,j})_{j \in [d]}$ and $\varphi_d$ the density of $X_{2,1}$[1], one can define $\gamma = \sqrt{\frac{1-\tau^2}{1-X_{2,1}^2}}$ and $X_2^+ := (\tau, \gamma X_{2,2}, \dots, \gamma X_{2,d})$ where $\tau$ is a random variable in $[-1,1]$ with density $\varphi^+_{d,p}(x) = p^{-1} \mathbb{1}_{x \geq t_{p,d}} \varphi_d(x)$. Denoting further $G_{\sim e_0}$ (resp. $G^+_{\sim e_0}$) the RGG with threshold $t_{p,d}$ induced by the latent points $(X_i)_{i \in [n]}$ (resp. $(X_1, X_2^+, X_3, \dots, X_n)$) without the edge $e_0 = \{1,2\}$, $G_{\sim e_0}$ (resp.$G^+_{\sim e_0}$) is distributed as $\nu_{\sim e_0}$ (resp. $\nu^+_{\sim e_0}$). Hence it holds,

$$\mathbb{E}[|Q-p|] \leq 2p \times \mathrm{TV}\left(\nu^+_{\sim e_0}, \nu_{\sim e_0}\right) \leq 2p \times \mathbb{P}(G_{\sim e_0} \neq G^+_{\sim e_0}) \leq 2p \sum_{i=3}^{n} \mathbb{P}\left(\mathbb{1}_{\langle X_2, X_i \rangle \geq t_{p,d}} \neq \mathbb{1}_{\langle X_2^+, X_i \rangle \geq t_{p,d}}\right).$$

The proof is concluded using standard concentration arguments.

### 3.3.2 Reaching the polylogarithmic regime

Very recently, Liu et al. (2021) came with novels ideas and improved upon the previous bounds for geometry detection by polynomial factors in the sparse regime. This significant breakthrough presented in Theorem 6 almost solves Conjecture 1.

**Theorem 6.** *(Liu et al., 2021, Theorem 1.2) For any fixed constant $c \geq 1$, if $d \gg \log^{36} n$, then*

$$\mathrm{TV}\left(G\left(n, \frac{c}{n}\right), G\left(n, \frac{c}{n}, d\right)\right) \to 0 \quad \text{as } n \to \infty.$$

The authors do not limit their analysis to the sparse regime but also provide results holding for any regime interpolating between the sparse and the dense cases as shown with Theorem 7.

**Theorem 7.** *(Liu et al., 2021, Theorem 1.1 and Lemma A.1)*

- *For any fixed constant $c > 0$, if $\frac{c}{n} < p < \frac{1}{2}$ and $d \gg p^2 n^3$, then*

$$\mathrm{TV}\left(G(n,p), G(n,p,d)\right) \to 0 \quad \text{as } n \to \infty.$$

- *If $\frac{1}{n^2} \ll p \leq 1 - \delta$ for any fixed constant $\delta > 0$, then as long as $d \ll (nH(p))^3$,*

$$\mathrm{TV}\left(G\left(n, \frac{c}{n}\right), G\left(n, \frac{c}{n}, d\right)\right) \to 1 \quad \text{as } n \to \infty,$$

*where $H(p) = p \log \frac{1}{p} + (1-p) \log \frac{1}{1-p}$ is the binary entropy function. This result can be achieved using the signed triangle statistic following an approach strictly analogous to Bubeck et al. (2016).*

[1] i.e. $\varphi_d$ is the density of a one-dimensional marginal of a uniform random point on $\mathbb{S}^{d-1}$.

Liu et al. (2021) extend the work from Bubeck et al. (2016) and prove that the signed statistic distinguishes between $G(n,p)$ and $G(n,p,d)$ not only in the sparse and dense cases but also for most $p$, as long as $d \ll (nH(p))^3$. We provide in the Appendix A a synthetic description of the proofs of Theorems 6 and 7. Let us mention that the proofs rely on a new concentration result for the area of the intersection of a random sphere cap with an arbitrary subset of $\mathbb{S}^{d-1}$, which is established using optimal transport maps and entropy-transport inequalities on the unit sphere. Liu et al. (2021) make use of this set-cap intersection concentration lemma for the theoretical analysis of the Belief Propagation algorithm.

### 3.4 Open problems and perspectives

The main results we have presented so far look as follows:

| Task | Current state of knowledge | Ref. |
|---|---|---|
| Recognizing if a graph can be realized as a RGG | NP-hard | 1998 |
| Testing between $G(n,p,d)$ and $G(n,p)$ in high-dimension for $p \in (0,1)$ fixed | $0$ — Polynomial time test — $n^3$ — Undistinguishable — $d$ | 2016 |
| Testing between $G(n,\frac{c}{n},d)$ and $G(n,\frac{c}{n})$ in high-dimension for $c>0$ | $0$ — Polynomial time test — $\log^3 n$ — ? — $\log^{36} n$ — Undistinguishable — $d$ | 2016 & 2021 |

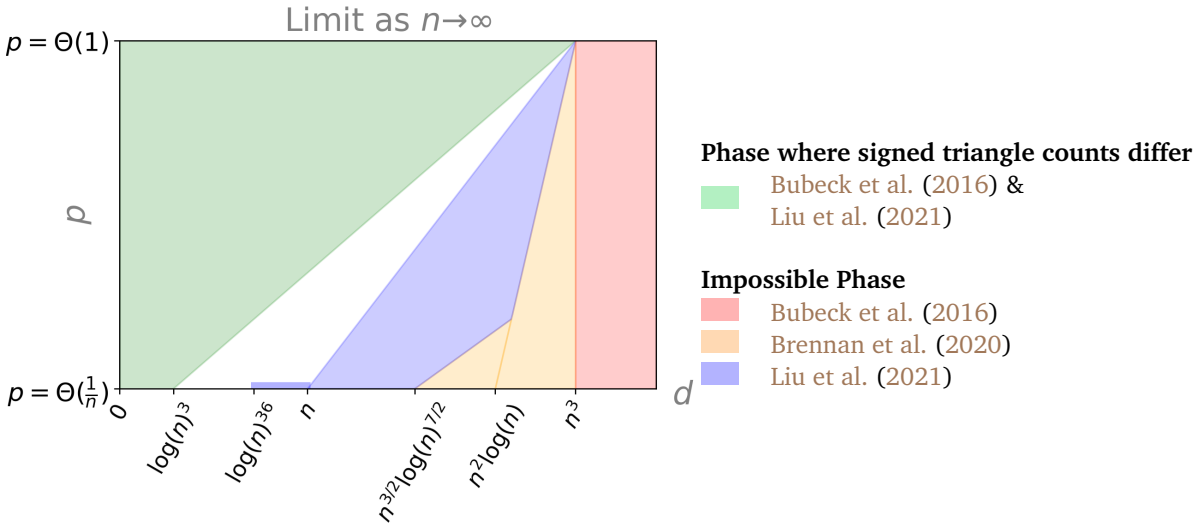


Figure 2: Phase-Diagram of the $(d,p)$ regions where geometry detection (on the Euclidean sphere) is known to be information theoretically impossible or possible (in polynomial time). Note that the figure only presents a simplified illustration of the current state of knowledge for the problem of geometry detection on $\mathbb{S}^{d-1}$ since the true scales are not respected.

With new proof techniques based on combinatorial arguments, direct couplings and applications of information inequalities, Brennan et al. (2020) were the first to make a progress towards Conjecture 1. Nevertheless, their proof was heavily relying on a coupling step involving a De Finetti-type result that requires the dimension $d$ to be larger than the number of points $n$. Liu et al. (2021) improved upon the previous bounds by polynomial factors with innovative proof arguments. In particular, their analysis makes use of the Belief Propagation algorithm and the cavity method, and relies on a new sharp estimate for the area of the intersection of a random sphere cap with an arbitrary subset of $\mathbb{S}^{d-1}$. The proof of this

new concentration result is an application of optimal transport maps and entropy-transport inequalities. Despite this recent progress, a large span of research directions remain open and we discuss some of them in the following.

1. *Closing the gaps for geometry detection on the Euclidean sphere* $\mathbb{S}^{d-1}$.
   Figure 2 shows that there are still important research directions to investigate to close the question of geometry detection regarding RGGs on $\mathbb{S}^{d-1}$. First in the sparse regime, it would be desirable to finally know if Conjecture 1 is true, meaning that the phase transition occurs when the latent dimension is of the order of $\log^3 n$. It could be fruitful to see if some steps in the approach from Liu et al. (2021) could be sharpened in order to get down to the threshold $\log^3 n$. A question that seems even more challenging is to understand what happens in the regimes where $p = p(n) \in (\frac{1}{n}, 1)$ and $d = d(n) \in ([H(p)n]^3, p^2 n^3)$ (corresponding to the white region on Figure 2). To tackle this question, one could try to extend the methods used in the sparse case by Liu et al. (2021) to denser cases. Another possible approach to close this gap would be to dig deeper into the connections between the Wishart and GOE ensembles. One research direction to possibly improve the existing impossibility results regarding geometry detection would be to avoid the use of the data-processing inequality in Eq.(2) which makes us lose the fact that we do not observe the matrices $W(n,d)$ and $M(n)$ themselves. To some extent, we would like to take into account that some information is lost by observing only the adjacency matrices. In a recent work, Brennan et al. (2021) made a first step in this direction. They study the total variation distance between the Wishart and GOE ensembles when some given mask is applied beforehand. They proved that the combinatorial structure of the revealed entries, viewed as the adjacency matrix of a graph $G$, drives the distance between the two distributions of interest. More precisely, they provide regimes for the latent dimension $d$ based exclusively on the number of various small subgraphs in $G$, for which the total variation distance goes to either 0 or 1 as $n \to \infty$.

2. *How specific is the signed triangle statistic to RGGs?*
   Let us mention that the signed triangle statistic has found applications beyond the scope of spatial networks. In Jin et al. (2019), the authors study community based random graphs (namely the Degree Corrected Mixed Membership model) and are interested in testing whether a graph has only one community or multiple communities. They propose the Signed Polygon as a class of new tests. In that way, they extend the signed triangle statistic to $m$-gon in the network for any $m \geq 3$. Contrary to Bubeck et al. (2016), the average degree of each node is not known and the Degree Corrected Mixed Membership model allows degree heterogeneity. In Jin et al. (2019), the authors define the signal-to-noise ratio (SNR) using parameters of their model and they prove that a phase transition occurs, namely $i$) when the SNR goes to $+\infty$, the Signed Polygon test is able to separate the alternative hypothesis from the null asymptotically, and $ii$) when the SNR goes to 0 (and additional mild conditions), then the alternative hypothesis is inseparable from the null.

3. *How the phase transition phenomenon in geometry detection evolves when other latent spaces are considered?*
   This question is related to the robustness of the previous results with respect to the latent space. Inspired by Bubeck et al. (2016), Eldan and Mikulincer (2020) provided a generalization of Theorem 1 considering an ellipsoid rather than the sphere $\mathbb{S}^{d-1}$ as latent space. They proved that the phase transition also occurs at $n^3$ provided that we consider the appropriate notion of dimension which takes into account the anisotropy of the latent structure.
   In Dall and Christensen (2002), the clustering coefficient of RGGs with nodes uniformly distributed on the hypercube shows systematic deviations from the Erdos-Rényi prediction.

4. *What is inherent to the connection function?*

   Considering a fixed number of nodes, Erba et al. (2020) use a multivariate version of the central limit theorem to show that the joint probability of rescaled distances between nodes is normal-distributed as $d \to \infty$. They provide a way to compute the correlation matrix. This work allows them to evaluate the average number of $M$-cliques, i.e. of fully-connected subgraphs with $M$ vertices, in high-dimensional RGGs and Soft-RGGs. They can prove that the infinite dimensional limit of the average number of $M$-cliques in Erdös-Rényi graphs is the same of the one obtained from for

Soft-RGGs with a continuous activation function. On the contrary, they show that for classical RGGs, the average number of cliques does not converge to the Erdös-Rényi prediction. This paper leads to think that the behavior of local observables in Soft-RGGs can heavily depend on the connection function considered. The work from Erba et al. (2020) is one of the first to address the emerging questions concerning the high-dimensional fluctuations of some statistics in RGGs. If they focused on the number of $M$-cliques, one can also mention the recent work from Grygierek and Thäle (2020) that provide a central limit theorem for the edge counting statistic as the space dimension $d$ tends to infinity. Their work shows that the Malliavin–Stein approach for Poisson functionals that was first introduced in stochastic geometry can also be used to deal with spatial random models in high dimensions.

In a recent work, Liu and Racz (2021a) are interested in extending the previous mentioned results on geometry detection in RGGs to Soft RGGs with some specific connection functions. The authors consider the dense case where the average degree of each node scales with the size of the graph $n$ and study geometry detection with graphs sampled from Soft-RGGs that interpolate between the standard RGG on the sphere $\mathbb{S}^{d-1}$ and the Erdös-Rényi random graph. Hence, the null hypothesis remains that the observed graph $G$ is a sample from $G(n,p)$ while the alternative becomes that the graph is the Soft-RGG where we draw an edge between nodes $i$ and $j$ with probability

$$(1-q)p + q\mathbb{1}_{t_{p,d}\leq\langle X_i,X_j\rangle},$$

where $(X_i)_{i\geq 1}$ are randomly and independently sampled on $\mathbb{S}^{d-1}$ and where $q\in[0,1]$ can be interpreted as the geometric strength of the model. Denoting the random graph model $G(n,p,d,q)$, one can easily notice that $G(n,p,d,1)$ is the standard RGG on the Euclidean sphere $\mathbb{S}^{d-1}$ while $G(n,p,d,0)$ reduces to the Erdös-Rényi random graph. Hence, by taking $q=1$ in Theorem 8, we recover Theorem 1 from Bubeck et al. (2016). One can further notice that Theorem 8 depicts a polynomial dependency on $q$ for geometry detection but when $q<1$ there is a gap between the upper and lower bounds as illustrated by Figure 3 taken from Liu and Racz (2021a). As stated in Liu and Racz (2021a), "*[...] a natural direction of future research is to consider [geometry detection] for other connection functions or underlying latent spaces, in order to understand how the dimension threshold for losing geometry depends on them.*"

**Theorem 8.** *(Liu and Racz, 2021a, Theorem 1.1) Let $p\in(0,1)$ be fixed.*

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

$$\mathrm{TV}(G(n,p),G(n,p,d,q))\to 1 \quad \text{as } n\to\infty.$$

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

$$\mathrm{TV}(G(n,p),G(n,p,d,q))\to 0 \quad \text{as } n\to\infty.$$

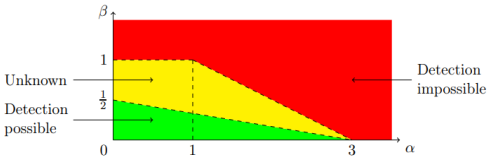


Figure 3: Phase diagram for detecting geometry in the soft random geometric graph $G(n,p,d,q)$. Here $d=n^\alpha$ and $q=n^{-\beta}$ for some $\alpha,\beta>0$.

The same authors in Liu and Racz (2021b) extend the model of the Soft-RGG by considering the latent space $\mathbb{R}^d$ where the latent positions $(X_i)_{i\in[n]}$ are i.i.d. sampled with $X_1\sim\mathcal{N}(0,I_d)$. Two different nodes $i,j\in[n]$ are connected with probability $\mathbf{p}(\langle X_i,X_j\rangle)$ where $\mathbf{p}$ is a monotone increasing connection function. More precisely, they consider a connection function $\mathbf{p}$ parametrized by $i$) a cumulative distribution function $F:\mathbb{R}\to[0,1]$ and $ii$) a scalar $r>0$ and given by

$$\mathbf{p}: t\mapsto F\left(\frac{t-\mu_{p,d,r}}{r\sqrt{d}}\right),$$

where $\mu_{p,d,r}$ is determined by setting the edge density in the graph to be equal to $p$, namely $\mathbb{E}[\mathbf{p}(\langle X_1,X_2\rangle)]=p$. They work in the dense regime by considering that $p\in(0,1)$ is independent of $n$. The parameter $r$ encodes the flatness of the connection function and is typically a function of $n$. The authors prove phase transitions of detecting geometry in this framework in terms of

the dimension of the underlying geometric space $d$ and the variance parameter $r$. The larger $r$, the smaller the dimension $d$ at which the phase transition occurs. When $r \underset{n\to\infty}{\to} 0$, the connection function becomes an indicator function and the transition appears at $d \asymp n^3$ (recovering the result from Theorem 1 established for RGGs on the Euclidean Sphere).

5. *Suppose that we know that the latent variables are embedded in a Eucliden Sphere, can we estimate the dimension $d$ from the observation of the graph?*
When $p = 1/2$, Bubeck et al. (2016) obtained a bound on the difference of the expected number of signed triangles between consecutive dimensions leading to Theorem 9.

   **Theorem 9.** *(Bubeck et al., 2016, Theorem 5)*

   *There exists a universal constant $C > 0$, such that for all integers $n$ and $d_1 < d_2$, one has*

$$\mathrm{TV}(G(n, 1/2, d_1), G(n, 1/2, d_2)) \geq 1 - C\left(\frac{d_1}{n}\right)^2.$$

   The bound provided by Theorem 9 is tight in the sense that when $d \gg n$, $G(n, 1/2, d)$ and $G(n, 1/2, d+1)$ are indistiguishable as proved in Eldan (2015). More recently, Araya Valdivia and De Castro (2019) proposed a method to infer the latent dimension of a Soft-RGG on the Euclidean Sphere in the low dimensional setting. Their approach is proved to correctly recover the dimension $d$ in the relatively sparse regime as soon as the connection function belongs to some Sobolev class and satisfies a spectral gap condition.

6. *Extension to hypergraphs and information-theoretic/computational gaps.*
Let us recall that a hypergraph is a generalization of a graph in which an edge can join any number of vertices. Extensions of RGGs to hypergraphs have already been proposed in the literature (see for example Lunagómez et al. (2017)). A nice research direction would consist in investigating the problem of geometry detection in these geometric representations of random hypergraphs. As already discussed, it has been conjectured that the problem of geometry detection in RGGs on $\mathbb{S}^{d-1}$ does not present a statistical-to-algorithmic gap meaning that whenever it is information theoretically possible to differ $G(n, p, d)$ from $G(n, p)$, we can do it with a computational complexity polynomial in $n$ (using the signed triangle statistic). Dealing with hypergraphs, one can legitimately think that statistical-to-algorithmic gaps could emerge. This intuition is based on the fact that most of the time, going from a matrix problem to a tensor problem brings extra challenges. One can take the example of principal component analysis of Gaussian $k$-tensors with a planted rank-one spike (cf. Ben Arous et al. (2020)). In this problem, we assume that we observe for any $l \in [n]$,

$$\mathbf{Y}^l = \lambda u^{\otimes k} + \mathbf{W}^l,$$

   where $u \in \mathbb{S}^{d-1}$ is deterministic, $\lambda \geq 0$ is the signal-to-noise ratio and where $(\mathbf{W}^l)_{l\in[n]}$ are independent Gaussian $k$-tensor (we refer to Ben Arous et al. (2020) for further details). The goal is to infer the "planted signal" or "spike", $u$. In the matrix case (i.e. when $k = 2$), whenever the problem is information theoretically solvable, we can also recover the spike with a polynomial time algorithm (using for example a spectral method). If we look at the tensor version of this problem where $k \geq 3$, there is a regime of signal-to-noise ratios for which it is information theoretically possible to recover the signal but for which there is no known algorithm to approximate it in polynomial time in $n$. This is a statistical-to-algorithmic gap and we refer to (Brennan and Bresler, 2020, Section 3.8) and references therein for more details.

7. *Can we describe the properties of high dimensional RGGs in the regimes where* $\mathrm{TV}(G(n, p), G(n, p, d)) \to 1$ *as* $n \to \infty$*?*
In the low dimensional case, RGGs have been extensively studied: their spectral or topological properties, chromatic number or clustering number are now well known (see e.g. Walters (2011); Penrose et al. (2003)). One of the first work studying the properties of high dimensional RGGs is Avrachenkov and Bobu (2020) where the authors are focused on the clique structure. These questions are essential to understand how good high dimensional RGGs are as models for the theory of network science.

8. *How to find a relevant latent space given a graph with an underlying geometric structure?* As stated in Racz and Bubeck (2016), "*Perhaps the ultimate goal is to find good representations of network data, and hence to faithfully embed the graph of interest into an appropriate metric space*". This task is known as *manifold learning* in the Machine learning community. Recently Smith et al. (2019) proved empirically that the eigenstructure of the Laplacian of the graph provides information on the curvature of the latent space. This is an interesting research direction to propose model selection procedure and infer a relevant latent space for a graph.

# 4 Non-parametric inference in RGGs

In this section, we are interested in non-parametric inference in TIRGGs (see Definition 3) on the Euclidean sphere $\mathbb{S}^{d-1}$. The methods presented rely mainly on spectral properties of such random graphs. Note that spectral aspects in (Soft-)RGGs have been investigated for a long time (see for example Rai (2004)) and it is now well-known that the spectra of RGGs are very different from the one of other random graph models since the appearance of particular subgraphs give rise to multiple repeated eigenvalues (see Nyberg et al. (2015) and Blackwell et al. (2007)). Recent works took advantage of the information captured by the spectrum of RGGs to study topological properties such as Aguilar-Sánchez et al. (2020). In this section, we will see that random matrix theory is a powerful and convenient tool to study the spectral properties of RGGs as already highlighted by Dettmann et al. (2017).

## 4.1 Description of the model and notations

We consider a Soft-RGG on the Euclidean Sphere $\mathbb{S}^{d-1}$ endowed with the geodesic distance $\rho$. We consider that the connection function $H$ is of the form $H : t \mapsto \mathbf{p}(\cos(t))$ where $\mathbf{p} : [-1,1] \to [0,1]$ is an unknown function that we want to estimate. This Soft-RGG belongs to the class of TIRGG has defined in Section 2 and corresponds to a graphon model where the graphon $W$ is given by

$$\forall x, y \in \mathbb{S}^{d-1}, \quad W(x,y) := \mathbf{p}(\langle x, y\rangle).$$

$W$ viewed as an integral operator on square-integrable functions, is a compact convolution (on the left) operator

$$\mathbb{T}_W : f \in L^2(\mathbb{S}^{d-1}) \mapsto \int_{\mathbb{S}^{d-1}} W(x,\cdot) f(x) \sigma(dx) \in L^2(\mathbb{S}^{d-1}), \tag{4}$$

where $\sigma$ is the Haar measure on $\mathbb{S}^{d-1}$. The operator $\mathbb{T}_W$ is Hilbert-Schmidt and it has a countable number of bounded real eigenvalues $\lambda_k^*$ with zero as only accumulation point. The eigenfunctions of $\mathbb{T}_W$ have the remarkable property that they do not depend on $p$ (see (Dai and Xu, 2013, Lemma 1.2.3)): they are given by the real Spherical Harmonics. We denote $\mathscr{H}_l$ the space of real Spherical Harmonics of degree $l$ with dimension $d_l$ and with orthonormal basis $(Y_{l,j})_{j\in[d_l]}$. We end up with the following spectral decomposition of the *envelope* function $\mathbf{p}$

$$\forall x, y \in \mathbb{S}^{d-1}, \quad \mathbf{p}(\langle x, y\rangle) = \sum_{l\geq 0} p_l^* \sum_{j=1}^{d_l} Y_{l,j}(x) Y_{l,j}(y) = \sum_{l\geq 0} p_l^* c_l G_l^\beta(\langle x, y\rangle), \tag{5}$$

where $\lambda^* := (p_0^*, p_1^*, \dots, p_1^*, \dots, p_l^*, \dots, p_l^*, \dots)$ meaning that each eigenvalue $p_l^*$ has multiplicity $d_l$ and $G_l^\beta$ is the Gegenbauer polynomial of degree $l$ with parameter $\beta := \frac{d-2}{2}$ and $c_l := \frac{2l+d-2}{d-2}$. $\mathbf{p}$ is assumed bounded and as a consequence $\mathbf{p} \in L^2((-1,1), w_\beta)$ where the weight function $w_\beta$ is defined by $w_\beta(t) := (1-t^2)^{\beta-1/2}$. Note that the decomposition (5) shows that it is enough to estimate the eigenvalues $(p_l^*)_l$ to recover the envelope function $\mathbf{p}$.

## 4.2 Estimating the matrix of probabilities

Let us denote $A$ the adjacency matrix of the Soft-RGG $G$ given by entries $A_{i,j} \in \{0,1\}$ where $A_{i,j} = 1$ if the nodes $i$ and $j$ are connected and $A_{i,j} = 0$ otherwise. We denote by $\Theta$ the $n \times n$ symmetric matrix with

entries $\Theta_{i,j} = \mathbf{p}\left(\langle X_i, X_j\rangle\right)$ for $1 \le i < j \le n$ and zero diagonal entries. We consider the scaled version of the matrices $A$ and $\Theta$ given by

$$\widehat{T}_n = \frac{1}{n}A \quad \text{and} \quad T_n = \frac{1}{n}\Theta.$$

Bandeira and van Handel (2016) proved a near optimal error bound for the operator norm of $\widehat{T}_n - T_n$. Coupling this result with the Weyl's perturbation Theorem gives a control on the difference between the eigenvalues of the matrices $\widehat{T}_n$ and $T_n$, namely with probability greater that $1-\exp(-n)$ it holds,

$$\forall k \in [n], \quad |\lambda_k(\widehat{T}_n) - \lambda_k(T_n)| \le \|\widehat{T}_n - T_n\| = O(1/\sqrt{n}), \tag{6}$$

where $\lambda_k(M)$ is the $k$-th largest eigenvalue of any symmetric matrix $M$. This result shows that the spectrum of the scaled adjacency matrix $\widehat{T}_n$ is a good approximation of the one of the scaled matrix of probabilities $T_n$.

### 4.3 Spectrum consistency of the matrix of probabilities

For any $R \ge 0$, we denote

$$\tilde{R} := \sum_{l=0}^{R} d_l, \tag{7}$$

which corresponds to the dimension of the space of Spherical Harmonics with degree at most $R$. Proposition 2 states that the spectrum of $T_n$ converges towards the one of the integral operator $\mathbb{T}_W$ in the $\delta_2$ metric which is defined as follows.

**Definition 6.** *Given two sequences $x, y$ of reals–completing finite sequences by zeros–such that $\sum_i x_i^2 + y_i^2 < \infty$, we define the $\ell_2$ rearrangement distance $\delta_2(x,y)$ as*

$$\delta_2^2(x,y) := \inf_{\sigma \in \mathfrak{S}_n} \sum_i (x_i - y_{\sigma(i)})^2,$$

*where $\mathfrak{S}_n$ is the set of permutations with finite support. This distance is useful to compare two spectra.*

**Proposition 2.** *(De Castro et al., 2020, Proposition 4)*
*There exists a universal constant $C > 0$ such that for all $\alpha \in (0, 1/3)$ and for all $n^3 \ge \tilde{R}\log(2\tilde{R}/\alpha)$, it holds*

$$\delta_2(\lambda(T_n), \lambda^*) \le 2\left[\sum_{l>R} d_l \left(p_l^*\right)^2\right]^{1/2} + C\sqrt{\tilde{R}\left(1 + \log(\tilde{R}/\alpha)\right)/n}, \tag{8}$$

*with probability at least $1 - 3\alpha$.*

Proposition 2 shows that the $\ell_2$ rearrangement distance between $\lambda^*$ and $\lambda(T_n)$ decomposes as the sum of a bias term and a variance term. The second term on the right hand side of (8) corresponds to the variance. The proof leading to this variance bound relies on the Hoffman-Wielandt inequality and borrows ideas from Koltchinskii and Giné (2000). It makes use of recent developments in random matrix concentration by applying a Bernstein-type concentration inequality (see Tropp (2015) for example) to control the operator norm of the sum of independent centered symmetric matrices given by

$$\sum_{i=1}^{n} \left(\mathbf{Y}(X_i)\mathbf{Y}(X_i)^\top - \mathbb{E}\left[\mathbf{Y}(X_i)\mathbf{Y}(X_i)^\top\right]\right), \tag{9}$$

with $\mathbf{Y}(x) = \left(Y_{0,0}(x), Y_{1,1}(x), \ldots, Y_{1,d_1}(x), Y_{2,1}(x), \ldots, Y_{2,d_2}(x), \ldots, Y_{R,1}(x), \ldots, Y_{R,d_R}(x)\right)^\top \in \mathbb{R}^{\tilde{R}}$ for all $x \in \mathbb{S}^{d-1}$. The proof of Proposition 2 also exploits concentration inequality for U-statistic dealing with a bounded, symmetric and $\sigma$-canonical kernel (see (De la Pena and Giné, 2012, Definition 3.5.1)). The first term on the right hand side of (8) is the bias arising from choosing a resolution level equal to $R$. Its behaviour as a function of $R$ can be analyzed by considering some regularity condition on the envelope $\mathbf{p}$.

Assuming that $\mathbf{p}$ belongs to the Sobolev class $Z^s_{w_\beta}((-1,1))$ (with regularity encoded by some parameter $s>0$) defined by

$$\left\{ g=\sum_{k\geq 0} g_k^* c_k G_k^\beta \in L^2((-1,1),w_\beta) \;\middle|\; \|g\|^*_{Z^s_{w_\beta}((-1,1))} := \left[\sum_{l=0}^{\infty} d_l |g_l^*|^2 \left(1+(l(l+2\beta))^s\right)\right]^{1/2} < \infty \right\},$$

and choosing the resolution level $R_{opt} = \lceil (n/\log n)^{\frac{1}{2s+d-1}} \rceil$ to balance the bias/variance tradeoff appearing on the right hand side of (8), we get that

$$\mathbb{E}\left[\delta_2^2(\lambda(T_n),\lambda^*)\right] \lesssim \left[\frac{n}{\log n}\right]^{-\frac{2s}{2s+(d-1)}}.$$

Thus we recover a classical nonparametric rate of convergence for estimating a function with smoothness $s$ in a space of dimension $d-1$. This is also the rate towards the probability matrix obtained by Xu (2018). Note that the choice of $R_{opt}$ requires the knowledge of the regularity parameter $s$. To overcome this issue, De Castro et al. (2020) proposed an adaptive procedure using the Goldenshluger-Lepski method.

## 4.4 Estimation of the envelope function

Let us denote $\lambda := \lambda(\widehat{T}_n)$. For a prescribed model size $R \in \mathbb{N}\backslash\{0\}$, De Castro et al. (2020) define the estimator $\widehat{\lambda}^R$ of the truncated spectrum $\lambda^{*R} := (p_0^*, p_1^*, \ldots, p_1^*, \ldots, p_R^*, \ldots, p_R^*)$ of $\lambda^*$ as

$$\widehat{\lambda}^R := (p_0^R(\hat{\sigma}), p_1^R(\hat{\sigma}), \ldots, p_1^R(\hat{\sigma}), \ldots, p_1^R(\hat{\sigma}), \ldots, p_R^R(\hat{\sigma}), \ldots, p_R^R(\hat{\sigma})),$$

with

$$\hat{\sigma} \in \operatorname*{arg\,min}_{\sigma\in\mathfrak{S}_n} \sum_{l=0}^{R} \sum_{k=\widetilde{l-1}}^{\widetilde{l}} \left(p_l^R(\sigma) - \lambda_{\sigma(k)}\right)^2 + \sum_{k=\bar{R}+1}^{n} \lambda_{\sigma(k)}^2 \quad \text{and} \quad p_l^R(\sigma) = \frac{1}{d_l} \sum_{k=\widetilde{l-1}}^{\widetilde{l}} \lambda_{\sigma(k)},$$

where $\mathfrak{S}_n$ is the set of permutations of $[n]$ and where we used the notation (7) with the convention $\widetilde{-1}=1$. Using the results of the two previous subsections namely (6) and Proposition 2, we obtain (De Castro et al., 2020, Theorem.6) which states that

$$\mathbb{E}\left[\delta_2^2\left(\widehat{\lambda}^{R_{opt}},\lambda^*\right)\right] \lesssim \left[\frac{n}{\log n}\right]^{-\frac{2s}{2s+(d-1)}}.$$

The envelope function $\mathbf{p}$ can then be approximated by the plug-in estimator $\widehat{\mathbf{p}} \equiv \sum_{l=0}^{R_{opt}} p_l^{R_{opt}}(\hat{\sigma}) c_l G_l^\beta$ based on the decomposition (5). One drawback of this approach is the exponential complexity in $R$ of the computation of $\widehat{\lambda}^R$. In the next section, we will describe an approach based on a Hierarchical Agglomerative Clustering algorithm to estimate the envelope function $\mathbf{p}$ efficiently.

## 4.5 Open problems and perspectives

The minimax rate of estimating a $s$-regular function on a space of (Riemannian) dimension $d-1$ such as $\mathbb{S}^{d-1}$ from $n$ observations is known to be of order $n^{-\frac{s}{2s+d-1}}$. In the framework of this section, even if the domain of the envelope function $\mathbf{p}$ is $[-1,1]$, inputs of $\mathbf{p}$ are the pairwise distances given by inner products of points embedded in $\mathbb{S}^{d-1}$. Hence it is still an open question to know if the dimension $d$ of the latent space appears in the minimax rate of convergence. Moreover, the number of observations in the estimation problem considered is $n^2$ since the full adjacency matrix is known. Nevertheless the problem suffers from the presence of unobserved latent variables. This all contributes to a non standard estimation problem and finding the optimal rate of convergence is an open problem.

# 5 Growth-model in RGGs

## 5.1 Description of the model

In Duchemin and De Castro (2022), a new growth model was introduced for RGGs. The so-called Markov Random Geometric Graph (MRGG) already presented in Definition 4 is a Soft-RGG where latent points are sampled with Markovian jumps. Namely, Duchemin and De Castro (2022) consider $n$ points $X_1, X_2, \dots, X_n$ sampled on the Euclidean sphere $\mathbb{S}^{d-1}$ using a Markovian dynamic. They start by sampling uniformly $X_1$ on $\mathbb{S}^{d-1}$. Then, for any $i \in \{2, \dots, n\}$, they sample

- a unit vector $Y_i \in \mathbb{S}^{d-1}$ uniformly, orthogonal to $X_{i-1}$,
- a real $r_i \in [-1, 1]$ encoding the distance between $X_{i-1}$ and $X_i$, see (11). $r_i$ is sampled from a distribution $f_{\mathscr{L}} : [-1, 1] \to [0, 1]$, called the *latitude function*,

then $X_i$ is defined by

$$X_i = r_i \times X_{i-1} + \sqrt{1 - r_i^2} \times Y_i \,. \tag{10}$$

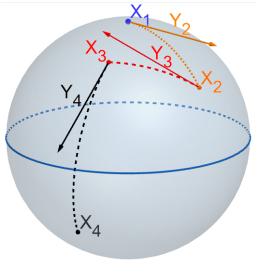


Figure 4: Visualization of the sampling scheme in $\mathbb{S}^2$.

This dynamic is illustrated with Figure 4 and can be understood as follows. Consider that $X_{i-1}$ is the north pole, then choose uniformly a direction (i.e. a longitude) and, in an independent manner, randomly move along the latitudes (the longitude being fixed by the previous step). The geodesic distance $\gamma_i$ drawn on the latitudes satisfies

$$\gamma_i = \arccos(r_i), \tag{11}$$

where random variable $r_i = \langle X_i, X_{i-1} \rangle$ has density $f_{\mathscr{L}}(r_i)$.

## 5.2 Spectral convergences

In this framework and keeping the notations of the previous section, one can show that if $\mathbf{p} \in Z^s_{w_\beta}((-1, 1))$ and if $f_{\mathscr{L}}$ satisfies the condition

$$(\mathscr{H}) \qquad \|f_{\mathscr{L}}\|_\infty := \sup_{t \in [-1,1]} |f_{\mathscr{L}}(t)| < \infty \quad \text{and} \quad f_{\mathscr{L}} \text{ is bounded away from zero,}$$

then

$$\mathbb{E}\left[\delta_2^2(\lambda(T_n), \lambda^*) \vee \delta_2^2(\lambda^{R_{opt}}(\widehat{T}_n), \lambda^*)\right] = \mathscr{O}\left(\left[\frac{n}{\log^2(n)}\right]^{-\frac{2s}{2s+d-1}}\right), \tag{12}$$

with $\lambda^{R_{opt}}(\hat{T}_n) = (\hat{\lambda}_1, \dots, \hat{\lambda}_{\tilde{R}_{opt}}, 0, 0, \dots)$ and $R_{opt} = \lfloor \left(n/\log^2(n)\right)^{\frac{1}{2s+d-1}} \rfloor$ where $\hat{\lambda}_1, \dots, \hat{\lambda}_n$ are the eigenvalues of $\widehat{T}_n$ sorted in decreasing order of magnitude. This result is the counterpart of Proposition 2 in this Markovian framework. The proof follows closely the steps of the one of the previous section but one needs to deal with the dependency of the latent positions. Results from Tropp (2015) are no longer suited to control the operator norm of (9) since $(X_i)_{i \geq 0}$ is a Markov chain. Nevertheless, this can be achieved by using concentration inequalities for sum of functions of Markov chains and by exploiting the rank one structure of the random matrices $\mathbf{Y}(X_i)\mathbf{Y}(X_i)^\top$ together with a covering set argument. Another difficulty induced by the latent dynamic is the control of a U-statistic of order 2 of the Markov chain $(X_i)_{i \geq 0}$ with a bounded kernel. Non-asymptotic results regarding the tail behaviour of U-statistics of a Markov chain have been so far very little touched. In a recent work, Duchemin et al. (2022) proved a concentration inequality for order 2 U-statistics with bounded kernels for uniformly ergodic Markov chain. Theorem 10 gives a simplified version of their main result. Assuming that the condition $(\mathscr{H})$ is fulfilled, the Markov chain $(X_i)_{i \geq 1}$ satisfies the assumptions of Theorem 10 and one can show that (12) holds true.

**Theorem 10.** *(Duchemin et al., 2022, Theorem 2) Let us consider a Markov chain $(X_i)_{i\geq 1}$ on some measurable space $(E,\mathscr{E})$ (with $E$ Polish) with transition kernel $P : E\times E \to \mathbb{R}$ and a function $h : E\times E \to \mathbb{R}$. We assume that*

1. *$(X_i)_{i\geq 1}$ is a uniformly ergodic Markov chain with invariant distribution $\pi$,*
2. *$h$ is bounded and $\pi$-canonical, namely*

$$\forall x \in E, \quad \mathbb{E}_{X\sim\pi}[h(X,x)] = \mathbb{E}_{X\sim\pi}[h(x,X)] = 0,$$

3. *there exist $\delta > 0$ and some probability measure $\nu$ on $(E,\mathscr{E})$ such that*

$$\forall x \in E,\ \forall A \in \mathscr{E}, \quad P(x,A) \leq \delta\,\nu(A).$$

*Then there exist constants $\beta,\kappa > 0$ such that for any $u \geq 1$, it holds with probability at least $1-\beta e^{-u}\log n$,*

$$\frac{1}{n(n-1)} \sum_{1\leq i,j\leq n,\, i\neq j} h(X_i,X_j) \leq \kappa \|h\|_\infty \log n \left\{ \frac{u}{n} + \left[\frac{u}{n}\right]^2 \right\},$$

*where $\kappa$ and $\beta$ only depend on constants related to the Markov chain $(X_i)_{i\geq 1}$.*

**Remark.** Note that Theorem 10 holds for any initial distribution of the Markov chain. In their paper, Duchemin et al. (2022) go beyond the previous Hoeffding tail control by providing a Bernstein-type concentration inequality under the additional assumption that the chain is stationary. For the sake of simplicity we presented Theorem 10 for a single kernel $h$, but we point out that their results allow for the dependence of the kernels – say $h_{i,j}$ – on the indexes in the sums which brings technical difficulties since standard blocking techniques can no longer be applied. The interest for this concentration result goes beyond the scope of random graphs since U-statistics naturally arise in online learning Clémençon et al. (2008) or testing procedures Fromont and Laurent (2006).

## 5.3 Estimation procedure

Recalling the notation of the truncated spectrum $\lambda^{*R}$ (resp. $\lambda^R(\widehat{T}_n)$) of $\lambda^*$ (resp. $\lambda(\widehat{T}_n)$) from Section 4.4, Duchemin and De Castro (2022) introduce a new procedure (namely the SCCHEi algorithm) based on a Hierarchical Agglomerative Clustering that returns a partition $\mathscr{C}_{d_0},\dots,\mathscr{C}_{d_R},\Lambda$ of the $n$ eigenvalues of $\widehat{T}_n$ where for any $i\in\{0,\dots,R\}$, $|\mathscr{C}_{d_i}| = d_i$ (where we recall that $d_i$ is the dimension of the space of spherical Harmonics of degree $i$). The authors prove that for any fixed resolution level $R$, $n$ can be chosen large enough so that the clusters obtained in polynomial time from the SCCHEi algorithm satisfy

$$\delta_2^2(\lambda^{*R},\lambda^R(\widehat{T}_n)) = \sum_{k=0}^{R} \sum_{\hat{\lambda}\in\mathscr{C}_{d_k}} (\hat{\lambda}-p_k^*)^2. \tag{13}$$

The final estimate of the envelope function with resolution level $R$ is defined as

$$\widehat{\mathbf{p}} := \sum_{k=0}^{R} \widehat{p}_k G_k^\beta, \quad \text{where} \quad \forall k\in\mathbb{N}, \quad \widehat{p}_k = \begin{cases} \frac{1}{d_k}\sum_{\lambda\in\mathscr{C}_{d_k}} \lambda & \text{if } k\in\{0,\dots,R\} \\ 0 & \text{otherwise.} \end{cases} \tag{14}$$

Eq.(13) is not a sufficient condition to ensure that the $L^2$ error between the true envelope function and the plug-in estimator $\widehat{\mathbf{p}}$ (see Eq.(14)) goes to 0 has $n\to+\infty$. This is due to identifiability issues coming from the $\delta_2$ metric. In (Duchemin and De Castro, 2022, Theorem 3), the author obtain a theoretical guarantee on the $L^2$ error between the true envelope function and the plug-in estimate by considering additional assumptions on the eigenvalues $(p_k^*)_{k\geq 0}$. Let us finally mention that the optimal resolution level $R_{opt}$ is unknown in practice. To bypass this issue, the authors propose a model selection procedure based on the slope heuristic (see Arlot (2019)).

### 5.4 Non-parametric link prediction

We are now interesting in solving link prediction tasks. Namely, from the observation of the graph at time $n$, we want to estimate the probabilities of connection between the upcoming node $n+1$ and the nodes already present in the graph. Recalling the definition of the random variables $(Y_i)_{i\geq 2}$ from Section 5.1 and denoting further $\mathrm{proj}_{X_n^\perp}(\cdot)$ the orthogonal projection onto the orthogonal complement of $\mathrm{Span}(X_n)$, the decomposition

$$\langle X_i, X_{n+1}\rangle = \langle X_i, X_n\rangle\langle X_n, X_{n+1}\rangle + \sqrt{1-\langle X_n, X_{n+1}\rangle^2}\sqrt{1-\langle X_i, X_n\rangle^2}\,\Big\langle \frac{\mathrm{proj}_{X_n^\perp}(X_i)}{\|\mathrm{proj}_{X_n^\perp}(X_i)\|_2}, Y_{n+1}\Big\rangle, \tag{15}$$

shows that latent distances $\mathbf{D}_{1:n} = (\langle X_i, X_j\rangle)_{1\leq i,j\leq n} \in [-1,1]^{n\times n}$ are enough for link prediction. Indeed, it can be achieved by estimating the posterior probabilities defined for any $i \in [n]$ by

$$\begin{aligned}
\eta_i(\mathbf{D}_{1:n}) &= \mathbb{P}\left(A_{i,n+1} = 1 \mid \mathbf{D}_{1:n}\right) \\
\eta_i(\mathbf{D}_{1:n}) &= \int_{r,u\in(-1,1)} \mathbf{p}\Big(\langle X_i, X_n\rangle r + \sqrt{1-r^2}\sqrt{1-\langle X_i, X_n\rangle^2}u\Big) f_{\mathscr{L}}(r) w_{\frac{d-3}{2}}(u) \frac{\Gamma(\frac{d-1}{2})}{\Gamma(\frac{d-2}{2})\sqrt{\pi}} dr du,
\end{aligned} \tag{16}$$

where $A_{i,n+1} \in \{0,1\}$ is one if and only if node $n+1$ is connected to node $i$, $w_{\frac{d-3}{2}}(u) := (1-u^2)^{\frac{d-3}{2}-\frac{1}{2}}$ and where $\Gamma : a \in ]0,+\infty[ \mapsto \int_0^{+\infty} t^{a-1}e^{-t}dt$. Using an approach similar to Araya Valdivia and De Castro (2019), Duchemin and De Castro (2022) proved that one can get a consistent estimator $\widehat{G}$ of the Gram matrix of the latent positions $G = \left(\langle X_i, X_j\rangle\right)_{1\leq i,j\leq n}$ in Frobenius norm. Hence, one can use a traditional plug-in estimator for $\eta_i(\mathbf{D}_{1:n})$ by replacing in (16) $(i)$ the envelope function $\mathbf{p}$ by $\widehat{\mathbf{p}}$ from (14), $(ii)$ the pairwise distances by their estimates $\left(\widehat{G}_{i,j}\right)_{1\leq i,j\leq n}$ and $(iii)$ the latitude function $f_{\mathscr{L}}$ by a non-parametric kernel density estimator built from the latent distances between consecutive nodes $(\langle X_i, X_{i+1}\rangle)_{i\in[n-1]}$ estimated by $\left(\widehat{G}_{i,i+1}\right)_{i\in[n-1]}$.

Through the example of MRGG, one can easily grasp the interest of growth model for random graphs with a geometric structure. Modeling the time evolution of networks, one can hope to solve tasks such as link prediction or collaborative filtering. An interesting research direction would be to extend the previous work to an anisotropic Markov kernel.

## 6 Connections with community based models

We have already described open problems and interesting directions to pursue regarding the questions tackled in the Sections 3, 4 and 5. In this last section, we want to look at RGGs from a different lens by highlighting a recently born line of research that investigates the connections between RGGs and community based models. Without aiming at presenting in a comprehensive manner the literature on this question, we rather focus on a few recent works that could inspire the reader to contribute in this emerging field.

A plenty number of random graph models have been so far studied. However real world problems never match a particular model and most of the time present several internal structures. To take into account this complexity, a growing number of works have been trying to take the best of several known random graph models. Papadopoulos et al. (2012) introduced a growth model where new connections with the upcoming node are drawn taking into account both popularity and similarity of vertices. The motivation is to find a balance between two trends for new connections in social networks namely *homophily* and *popularity*. One can also mention Jordan and Wade (2015) who consider a growth model that interpolates between pure preferential attachment (essentially the well-known Barabasi–Albert model) and a purely geometric model (the online nearest-neighbour graph). As pointed out by (Barthélemy, 2011, Section II.B.3.a), " *it is clear that community detection in spatial networks is a very interesting problem which might receive a specific answer.*"

## 6.1 Extension of RGGs to take into account community structure

Galhotra et al. (2017) proposed a new random graph model that incorporates community membership in standard RGGs. More precisely, they introduce the Geometric Block Model which is defined as follows. Consider $V = V_1 \sqcup V_2 \sqcup \cdots \sqcup V_k$ a partition of $[n]$ in $k$ clusters, $(X_u)_{u\in[n]}$ independent and identical random vectors uniformly distributed on $\mathbb{S}^{d-1}$ and let $\left(r_{i,j}\right)_{1\leq i,j\leq k} \in [0,2]^{k\times k}$. The Geometric Block Model is a random graph with vertices $V$ and an edge exists between $v \in V_i$ and $u \in V_j$ if and only if $\|X_u - X_v\| \leq r_{i,j}$. Focusing on the case where $r_{i,i} = r_s, \forall i$ and $r_{i,j} = r_d,\ \forall i \neq j$, the authors want to recover the partition $V$ observing only the adjacency matrix of the graph. They proved that in the relatively sparse regime (i.e. when $r_s, r_d = \Omega_n\left(\frac{\log n}{n}\right)$), a simple motif-counting algorithm allows to detect communities in the Geometric Block Model and is near-optimal. The proposed greedy algorithm affects two nodes to the same community if the number of their common neighbours lies in a prescribed range whose bounds depend on $r_s$ and $r_d$ that are assumed to be known. The method is proved to recover the correct partition of the nodes with probability tending to 1 as $n$ goes to $+\infty$.

In Sankararaman and Baccelli (2017), the previous work is extended by considering arbitrary connection function. The paper sheds light on interesting differences between the standard SBMs and community models that incorporates some geometric structure. We start by presenting their model before highlighting some interesting results. Their model is the Planted Partition Random Connection Model (PPCM) that relies on a Poisson Point Process on $\mathbb{R}^d$ with intensity $\lambda > 0$ $\varphi := \{X_1, X_2, \dots\}$ where it is assumed that the enumeration of the points $X_i$ is such that for all $i, j \in \mathbb{N}$, $i > j \implies \|X_i\|_\infty \geq \|X_j\|_\infty$. Each atom $i \in \mathbb{N}$ is marked with a random variable $Z_i \in \{-1, +1\}$. $\overline{\varphi}$ is the marked Poisson Point Process. The sequence $\{Z_i\}_{i\in\mathbb{N}}$ is i.i.d. with each element being uniformly distributed in $\{-1, +1\}$. The interpretation of this marked point process is that for any node $i \in \mathbb{N}$, its location label is $X_i$ and its community label is $Z_i$. Considering two connection functions $f_{in},\ f_{out} : \mathbb{R}_+ \to [0,1]$, they first construct an infinite graph $G$ with vertex set $\mathbb{N}$ and place an edge between any two nodes $i, j \in \mathbb{N}$ with probability $f_{in}(\|X_i - X_j\|)\mathbb{1}_{Z_i=Z_j} + f_{out}(\|X_i - X_j\|)\mathbb{1}_{Z_i\neq Z_j}$. The graph $G_n$ is then the induced subgraph of $G$ consisting of the nodes 1 through $N_n$ where $N_n := \sup\left\{i \geq 0 : X_i \in B_n := [-\frac{n^{1/d}}{2}, \frac{n^{1/d}}{2}]^d\right\}$.
Considering that the graph is observed and that the connections functions $f_{in}, f_{out}$ and the location labels $(X_i)_i$ are known, the authors investigate conditions on the parameters of their model allowing to extract information on the community structure from the observed data.

**Weak recovery** Weak Recovery is said to be solvable if for every $n \in \mathbb{N}\backslash\{0\}$, there exists some algorithm that - based on the observed data $G_n$ and $\varphi$ - provides a sequence of $\{-1, +1\}$ valued random variables $\{\tau_i^{(n)}\}_{i=1}^{N_n}$ such that there exists a constant $\gamma > 0$ such that the *overlap* between $\{\tau_i^{(n)}\}_{i=1}^{N_n}$ and $\{Z_i\}_{i=1}^{N_n}$ is asymptotically almost surely larger than $\gamma$, namely

$$\lim_{n\to\infty} \mathbb{P}\left(\frac{\sum_{i=1}^{N_n} \tau_i^{(n)} Z_i}{N_n} \geq \gamma\right) = 1.$$

The authors identify regimes where weak recovery can be solved or not. We summarize their results with Proposition 3.

**Proposition 3.** *(Sankararaman and Baccelli, 2017, Proposition 1 - Corollary 2 - Theorem 2)*
*For every $f_{in}(\cdot), f_{out}(\cdot)$ such that $\{r \in \mathbb{R}^+ : f_{in}(r) \neq f_{out}(r)\}$ has positive Lebesgue measure and any $d \geq 2$, there exists a $\lambda_c \in (0, \infty)$ such that*

- *for any $\lambda < \lambda_c$, weak recovery is not solvable.*
- *for any $\lambda > \lambda_c$, there exists an algorithm (which could possibly take exponential time) to solve weak recovery.*

*Moreover, there exists $\tilde{\lambda}_c < \infty$ (possibly larger than $\lambda_c$) depending on $f_{in}(\cdot), f_{out}(\cdot)$ and $d$, such that for all $\lambda > \tilde{\lambda}_c$ , weak recovery is solvable in polynomial time.*

The intrinsic nature of the problem of weak recovery is completely different in the PPCM model compared to the standard sparse SBM. Sparse SBMs are known to be locally tree-like with very few short

cycles. Efficient algorithms that solve weak recovery in the sparse SBM (such as message passing algorithm, convex relaxation or spectral methods) deeply rely on the local tree-like structure. On the contrary, PPCMs are locally dense even if their are globally sparse. This is due to the presence of a lot of short loops (such as triangles). As a consequence, the standard tools used for SBMs are not relevant to solve weak recovery in PPCMs. Nevertheless, the local density allows to design a polynomial time algorithm that solves weak recovery for $\lambda > \tilde{\lambda}_c$ (see Proposition 3) by simply considering the neighbours of each node. Proposition 3 lets open the question of the existence of a gap between information versus computation thresholds. Namely, is it always possible to solve weak recovery in polynomial time when $\lambda > \lambda_c$? In the sparse and symmetric SBM, it is known that there is no information-computation gap for $k = 2$ communities, while for $k \geq 4$ a non-polynomial algorithm is known to cross the Kesten-Stigum threshold which was conjectured by Decelle et al. (2011) to be the threshold at which weak recovery can be solved efficiently.

**Distinguishability** The distinguishability problem asks how well one can solve a hypothesis testing problem that consists in finding if a given graph has been sampled from the PPCM model or from the null, which is given by a plain random connection model with connection function $(f_{in}(\cdot) + f_{out}(\cdot))/2$ without communities but having the same average degree and distribution for spatial locations. Sankararaman and Baccelli (2017) prove that for every $\lambda > 0$, $d \in \mathbb{N}$ and connection functions $f_{in}(\cdot)$ and $f_{out}(\cdot)$ satisfying $1 \geq f_{in}(r) \geq f_{out}(r) \geq 0$ for all $r \geq 0$, and $\{r \geq 0 : f_{in}(r) \neq f_{out}(r)\}$ having positive Lebesgue measure, the probability distribution of the null and the alternative of the hypothesis test are mutually singular. As a consequence, there exists some regimes (such as $\lambda < \lambda_c$ and $d \geq 2$) where we can be very sure by observing the data that a partition exists, but cannot identify it better than at random. In these cases, it is out of reach to bring together the small partitions of nodes in different regions of the space into one coherent. Such behaviour does not exist in the sparse SBM with two communities as proved by Mossel et al. (2014) and was conjectured to hold also for $k \geq 3$ communities in Decelle et al. (2011).

## 6.2 Robustness of spectral methods for community detection with geometric perturbations

In another line of work, Péché and Perchet (2020) are studying robustness of spectral methods for community detection when connections between nodes are perturbed by some latent random geometric graph. They identify specific regimes in which spectral methods are still efficient to solve community detection problems despite geometric perturbations and we give an overview of their work in what follows. Let us consider some fixed parameter $\kappa \in [0,1]$ that drives the balance between strength of the community signal and the noise coming from the geometric perturbations. For sake of simplicity, they consider a model with two communities where each vertex $i$ in the network is characterized by some vector $X_i \in \mathbb{R}^2$ with distribution $\mathcal{N}(0, I_2)$. They consider $p_1, p_2 \in (0,1)$ that may depend on the number of nodes $n$ with $p_1 > p_2$ and $\sup_n p_1/p_2 < \infty$. Assuming for technical reason $\kappa + \max\{p_1, p_2\} \leq 1$, the probability of connection between $i$ and $j$ is

$$\mathbb{P}\left\{i \sim j \mid X_i, X_j\right\} = \kappa \exp\left(-\gamma \|X_i - X_j\|^2\right) + \begin{cases} p_1 & \text{if } i \text{ and } j \text{ belong to the same community} \\ p_2 & \text{otherwise.} \end{cases},$$

where the inverse width $\gamma > 0$ may depend on $n$. We denote by $\sigma \in \{\pm 1/\sqrt{n}\}^n$ the normalized community vector illustrating to which community each vertex belong ($\sigma_i = -1/\sqrt{n}$ if $i$ belongs to the first community and $\sigma_i = 1/\sqrt{n}$ otherwise). The matrix of probabilities of this model is given by $Q := P_0 + P_1$ where

$$P_0 := \begin{bmatrix} p_1 J & p_2 J \\ p_2 J & p_1 J \end{bmatrix} \quad \text{and} \quad P_1 := \kappa P = \kappa \left((1 - \delta_{i,j}) e^{-\gamma \|X_i - X_j\|^2}\right)_{1 \leq i,j \leq n}.$$

The adjacency matrix $A$ of the graph can thus be written as $A = P_0 + P_1 + A_c$ where $A_c$ is, conditionnally on the $X_i$'s, a random matrix with independent Bernoulli entries which are centered. Given the graph-adjacency matrix $A$, the objective is to output a normalized vector $x \in \{\pm 1/\sqrt{n}\}^n$ such that, for some $\varepsilon > 0$,

- Exact recovery: with probability tending to 1, $|\sigma^\top x| = 1$,
- Weak recovery (also called detection): with probability tending to 1, $|\sigma^\top x| > \varepsilon$.

Let us highlight that contrary to the previous section, the latent variables $(X_i)_i$ are not observed. When $\kappa = 0$, we recover the standard SBM: $Q = P_0$ has two non zero eigenvalues which are $\lambda_1 = n(p_1+p_2)/2$ with associated normalized eigenvector $v_1 = \frac{1}{\sqrt{n}}(1,1,\dots,1)^\top$ and $\lambda_2 = n(p_1-p_2)/2$ associated to $v_2 = \sigma = \frac{1}{\sqrt{n}}(1,\dots,1,-1,\dots,-1)^\top$. Spectral methods can thus be used to recover communities by computing the second eigenvector of the adjacency matrix $A$. To prove that spectral methods still work in the presence of geometric perturbations, one needs to identify regimes in which the eigenvalues of $A$ are well separated and the second eigenvector is approximately $v_2$.

In the regime where $\gamma \gg n/\log n$, the spectral radius $\rho(P_1)$ of $P_1$ vanishes and we asymptotically recover a standard SBM. Hence, they focus on the following regime

$$\gamma \underset{n\to\infty}{\to} \infty \quad \text{and} \quad \frac{1}{\gamma}\frac{n}{\ln n} \underset{n\to\infty}{\to} \infty. \tag{$A_1$}$$

Under Assumption ($A_1$), (Péché and Perchet, 2020, Proposition 2) states that with probability tending to one, $\rho(P_1)$ is of order $\frac{\kappa n}{2\gamma}$. Using (Benaych-Georges et al., 2020, Theorem 2.7) to get an asymptotic upper-bound on the spectral radius of $A_c$, basic perturbation arguments would prove that standard techniques for community detection work in the regime where

$$\frac{\kappa n}{2\gamma} \ll \sqrt{\frac{n(p_1+p_2)}{2}} = \sqrt{\lambda_1}.$$

Indeed, it is now well-known that weak recovery in the SBM can be solved efficiently as soon as $\lambda_2 > \sqrt{\lambda_1}$ (for example using the power iteration algorithm on the non-backtracking matrix from Bordenave et al. (2015)). Hence, the regime of interest correspond to the case where

$$\exists c, C > 0 \quad \text{s.t.} \quad \lambda_2^{-1}\frac{\kappa n}{2\gamma} \in [c, C], \quad \frac{\lambda_2}{\lambda_1} \in [c, C] \quad \text{and} \quad \lambda_2 \gg \sqrt{\lambda_1}, \tag{$A_2$}$$

which corresponds to the case where the noise induced by the latent random graph is of the same order of magnitude as the signal. Under ($A_2$), the problem of weak recovery can be tackled using spectral methods on the matrix $S = P_0 + P_1$: the goal is to reconstruct communities based on the second eigenvector of $S$. To prove that these methods work, the authors first find conditions ensuring that two eigenvalues of $S$ exit the support of the spectrum of $P_1$. Then, they provide an asymptotic lower bound for the level of correlation between $v_2 = \sigma$ and the second eigenvector $w_2$ of $S$, which leads to Theorem 11.

**Theorem 11.** *(Péché and Perchet, 2020, Theorem 10)*
*Suppose that Assumptions ($A_1$) and ($A_2$) hold and that $\lambda_1 > \lambda_2 + 2\frac{\kappa}{2\gamma}$. Then the correlation $|w_2^\top v_2|$ is uniformly bounded away from 0. Moreover, denoting $\mu_1$ the largest eigenvalue of $P_1$, if the ratio $\lambda_2/\mu_1$ goes to infinity then $|w_2^\top v_2|$ tends to 1, which gives weak (and even exact at the limit) recovery.*

## 6.3 Recovering latent positions

From another viewpoint, one can think RGGs as an extension of stochastic block models where the discrete community structure is replaced by an underlying geometry. With this mindset, it is natural to directly transport concepts and questions from clustered random graphs to RGGs. For instance, the task consisting in estimating the communities in SBMs may correspond to the estimation of latent point neighborhoods in RGGs. More precisely, community detection can be understood in RGGs as the problem of recovering the geometric representation of the nodes (e.g. through the Gram matrix of the latent positions). This question has been tackled by Eldan et al. (2020) and Araya Valdivia (2020). Both works consider random graphs sampled from the TIRGG model on the Euclidean sphere $\mathbb{S}^{d-1}$ with some envelope function $\mathbf{p}$ (see Definition 3), leading to a graphon model similar to the one presented in Section 4.1. While the result from Araya Valdivia (2020) holds in the dense and relatively sparse regimes, the one from Eldan et al. (2020) covers the sparse case. Thanks to harmonic properties of $\mathbb{S}^{d-1}$, the graphon eigenspace composed only with linear eigenfunctions (harmonic polynomials of degree one) directly relates to the pairwise distances of the latent positions. This allows Eldan et al. (2020) and Araya Valdivia (2020) to provide a consistent estimate of the Gram matrix of the latent positions in Frobenius norm using a spectral method. Their results hold under the following two key assumptions.

1. *An eigenvalue gap condition.* They assume that the $d$ eigenvalues of the integral operator $\mathbb{T}_W$ - associated with the graphon $W := \mathbf{p}(\langle \cdot, \cdot \rangle)$ (see (4)) - corresponding to the Spherical Harmonics of degree one is well-separated from the rest of the spectrum.

2. *A regularity condition.* They assume that the envelope function $\mathbf{p}$ belongs to some Weighted Sobolev space, meaning that the sequence of eigenvalues of $\mathbb{T}_W$ goes to zero fast enough.

In addition to similar assumptions, Eldan et al. (2020) and Araya Valdivia (2020) share the same proof structure. First they need to recover the $d$ eigenvectors from the adjacency matrix corresponding to the space of spherical Harmonics of degree one. Then the Davis-Kahan Theorem is used to prove that the estimate of the Gram matrix based on the previously selected eigenvectors is consistent in Frobenius norm. To do so, they require a concentration result ensuring that the adjacency matrix $A$ (or some proxy of it) converges in operator norm towards the matrix of probabilities $\Theta$ with entries $\Theta_{i,j} = \mathbf{p}\left(\langle X_i, X_j \rangle\right)$ for $1 \leq i \neq j \leq n$ and zero diagonal entries. Araya Valdivia (2020) relies on (Bandeira and van Handel, 2016, Corollary 3.12), already discussed in (6), that provides the convergence $\|A - \Theta\| \to 0$ as $n \to \infty$ in the dense and relatively sparse regimes. In the sparse regime, such concentration no longer holds. Indeed, in that case, degrees of some vertices are much higher than the expected degree, say deg. As a consequence, some rows of the adjacency matrix $A$ have Euclidean norms much larger than $\sqrt{\text{deg}}$, which implies that for $n$ large enough, it holds with high probability $\|A - \Theta\| \gg \sqrt{\text{deg}}$. To cope with this issue, Eldan et al. (2020) do not work directly on the adjacency matrix but rather on a slightly amended version of it - say $A'$ - where one reduces the weights of the edges incident to high degree vertices. In that way, all degrees of the new (weighted) network become bounded, and (Le et al., 2018, Theorem 5.1) ensures that $A'$ converges to $\Theta$ in spectral norm as $n$ goes to $+\infty$. Hence in the sparse regime the adjacency matrix converges towards its expectation *after regularization*. The proof of this random matrix theory tool is based on a famous result in functional analysis known as the Grothendieck-Pietsch factorization.
Let us finally mention that this change of behaviour of the extreme eigenvalues of the adjacency matrix according to the maximal mean degree has been studied in details for inhomogeneous Erdös-Rényi graphs in Benaych-Georges et al. (2020) and Benaych-Georges et al. (2019).

## 6.4 Some perspectives

The paper Sankararaman and Baccelli (2017) makes the strong assumption that the locations labels $(X_i)_{i \geq 1}$ are known. Hence it should be considered as an initial work calling for future theoretical and practical investigations. Keeping the same model, it would be of great interest to design algorithms able to deal with unobserved latent variables to allow real-data applications. A first step in this direction was made by Avrachenkov et al. (2021) where the authors propose a spectral method to recover hidden clusters in the Soft Geometric Block Model where latent positions are not observed. On the theoretical side, Sankararaman and Baccelli (2017) describe at the end of their paper several open problems. Their suggestions for future works include $i$) the extension of their work to a larger number of communities, $ii$) the estimation from the data of the parameters of their model (namely $f_{in}$ and $f_{out}$ that they assumed to be known), and $iii$) the existence of a possible gap between information versus computation thresholds, namely, they wonder if there is a regime where community detection is solvable, but without any polynomial (in $n$) time and space algorithms.
Another possible research direction is the extension of the work from Section 6.2 to study the same kind of robustness results for more than 2 communities and especially in the sparse regime where $\frac{1}{\gamma} \sim p_i \sim \frac{1}{n}$. As highlighted by Péché and Perchet (2020), the sparse case may bring additional difficulties since " *standard spectral techniques in this regime involve the non-backtracking matrix (see Bordenave et al. (2015)), and its concentration properties are quite challenging to establish.*" Regarding Section 6.3, for some applications it may be interesting to go beyond the recovery of the pairwise distances by embedding the graph in the latent space while preserving the Gram structure. Such question has been tackled for example by Perry et al. (2020) but only for the Euclidean sphere in small dimensions.

**Acknowledgements** The authors are in debt to Tselil Schramm who gave a great talk at the S.S.Wilks Memorial Seminar in Statistics (at Princeton University) providing insightful comments on the problem of geometry detection or more specifically on her paper Liu et al. (2021).

# A Outline of the proofs of Theorems 6 and 7

The proofs of Theorems 6 and 7 (cf. Section 3.3) are quite complex and giving their formal descriptions would require heavy technical considerations. In the following, we provide an overview of the proofs highlighting the nice mathematical tools used by Liu et al. (2021) and their innovative combination while putting under the rug some technical aspects.

Step 1. Relate the TV distance of the whole graphs to single vertex neighbourhood.

$$\begin{aligned} 2\mathrm{TV}(G(n,p,d),G(n,p))^2 &\leq \mathrm{KL}(G(n,p,d)||G(n,p)) \qquad \text{from Pinsker's inequality} \\ &\leq n \times \mathbb{E}_{G_{n-1}\sim G(n-1,p,d)}\left[\mathrm{KL}\left(\nu_n(\cdot|G_{n-1}),\mathrm{Bern}(p)^{\otimes(n-1)}\right)\right] \qquad \text{from Lemma 2} \\ &= \mathbb{E}_{G_{n-1}\sim G(n-1,p,d)}\mathbb{E}_{S\sim\nu_n(\cdot|G_{n-1})}\log\left(\frac{\nu_n(S|G_{n-1})}{p^{|S|}(1-p)^{n-1-|S|}}\right), \end{aligned} \tag{17}$$

where $\nu_n(\cdot|G_{n-1})$ denotes the distribution of the neighbourhood of vertex $n$ when the graph is sampled from $G(n,p,d)$ conditional on the knowledge of the connections between pairs of nodes in $[n-1]$ given by $G_{n-1}$. Hence, the main difference with Brennan et al. (2020) is that the tensorization argument from Lemma 2 is used node-wise (and not edge-wise). We are reduced to understand how a vertex incorporates a given graph of size $n-1$ sampled from the distribution $G(n-1,p,d)$. At a high level, the authors show that if one can prove that for some $\varepsilon>0$, with high probability over $G_{n-1}\sim G(n-1,p,d)$, it holds

$$\forall S\subseteq[n-1],\quad \nu_n(S|G_{n-1}) = \mathbb{P}_{G\sim G(n,p,d)}(N_G(n)=S\,|\,G_{n-1}) = (1\pm\varepsilon)p^{|S|}(1-p)^{n-1-|S|}, \tag{18}$$

where $N_G(n)$ denotes the set of nodes connected to node $n$ in the graph $G$, then

$$\mathrm{TV}(G(n,p,d),G(n,p)) = o_n(n\varepsilon^2). \tag{19}$$

Step 2. Geometric interpretation of neighbourhood probabilities from Eq.(18).
For $G\sim G(n,p,d)$, if vertex $i$ is associated to a (random) vector $X_i$, and $(i,j)$ is an edge, we consequently know that $\langle X_i,X_j\rangle\geq t_{p,d}$. On the sphere $\mathbb{S}^{d-1}$, the locus of points where $X_j$ can be, conditioned on $(i,j)$ being an edge, is a sphere cap centered at $X_i$ with a $p$ fraction of the sphere's surface area, which we denote by $\mathrm{cap}(X_i)$. Similarly, if we know that $i$ and $j$ are not adjacent, the locus of points where $X_j$ can fall is the complement of a sphere cap with measure $1-p$ namely $\overline{\mathrm{cap}(X_i)}$, which we call an "anti-cap". Let us denote $\sigma$ is the normalized Lebesgue measure on $\mathbb{S}^{d-1}$ so that $\sigma(\mathbb{S}^{d-1})=1$. Equipped with this geometric picture, we can view the probability that vertex $n$'s neighborhood is exactly equal to $S\subseteq[n-1]$ as $\sigma(L_S)$, where $L_S\subseteq\mathbb{S}^{d-1}$ is a random set defined by

$$L_S := \Big(\bigcap_{i\in S}\mathrm{cap}(X_i)\Big)\cap\Big(\bigcap_{j\notin S}\overline{\mathrm{cap}(X_j)}\Big).$$

To show that the TV distance between $G(n,p,d)$ and $G(n,p)$ is small, we need to prove that $\sigma(L_S)$ concentrates around $p^{|S|}(1-p)^{n-1-|S|}$ as suggested by Eqs.(18) and (19).

Step 3. Concentration of measure of intersections of sets in $\mathbb{S}^{d-1}$ with random spherical caps.
An essential contribution of Liu et al. (2021) is a novel concentration inequality for the area of the intersection of a random spherical cap with any subset $L\subseteq\mathbb{S}^{d-1}$.

**Lemma 3.** *(see Liu et al., 2021, Corollary 4.10)* **Set-cap intersection concentration Lemma.** *Suppose $L\subseteq\mathbb{S}^{d-1}$ and let us denote by $\sigma$ the uniform probability measure on $\mathbb{S}^{d-1}$. Then with high probability over $z\sim\sigma$ it holds*

$$\Big|\frac{\sigma(L\cap\mathrm{cap}(z))}{p\sigma(L)}-1\Big| = \mathcal{O}_n\big(\delta_n(L)\big) \quad \textit{and} \quad \Big|\frac{\sigma(L\cap\overline{\mathrm{cap}(z)})}{(1-p)\sigma(L)}-1\Big| = \mathcal{O}_n\Big(\frac{p}{1-p}\delta_n(L)\Big),$$

*where* $\delta_n(L) = \sqrt{\frac{\log\frac{1}{p}+\log\frac{1}{\sigma(L)}}{\sqrt{d}}}\,\mathrm{polylog}(n)$.

*Sketch of proof of Lemma 3.* We give an overview of the proof of Lemma 3, highlighting its interesting connection with optimal transport. Let us consider some probability distribution $\nu$ on $\mathbb{S}^{d-1}$. Let us denote $\mathscr{D}$ the optimal coupling between the measures $\nu$ and $\sigma$, i.e. $\mathscr{D}$ is a probability measure on $\mathbb{S}^{d-1}\times\mathbb{S}^{d-1}$ with marginals $\nu$ and $\sigma$ such that

$$W_2(\nu,\sigma)^2 = \int \|x-y\|_2^2 d\mathscr{D}(x,y),$$

where $W_2(\nu,\sigma)$ is the Wasserstein 2-distance between the measures $\sigma$ and $\nu$. Then for any $z\in\mathbb{S}^{d-1}$ it holds

$$\begin{aligned}\mathbb{P}_{x\sim\nu}(\langle z,x\rangle > t_{p,d}) =& \mathbb{P}_{(x,y)\sim\mathscr{D}}(\langle z,y\rangle > t_{p,d} - \langle z,x-y\rangle)\\ \leq& \mathbb{P}_{y\sim\sigma}(\langle z,y\rangle > t_{p,d}-u(p,d)) + \mathbb{P}_{(x,y)\sim\mathscr{D}}(|\langle z,x-y\rangle| > u(p,d)),\end{aligned} \tag{20}$$

for some well chosen threshold $u(p,d)$ depending on $p$ and $d$. The first term in the right hand side of Eq.(20) can be proven to concentrate around $p$ with high probability over $z\sim\sigma$ with standard arguments. The second term in Eq.(20) quantifies how often a randomly chosen transport vector $x-y$ with $(x,y)\sim\mathscr{D}$ has a large projection in the direction $z$. One can prove that the optimal transport map $\mathscr{D}$ between $x\sim\nu$ and $y\sim\sigma$ has bounded length with high probability, and then translate this into a tail bound for the inner product $\langle z,x-y\rangle$ for a random vector $z\sim\sigma$. As a consequence, one can bound with high probability over $z\sim\sigma$ the fluctuations of $\left|\mathbb{P}_{x\sim\nu}(\langle z,x\rangle > t_{p,d})-p\right|$ which gives Lemma 3 if we take for $\nu$ the uniform measure on the set $L\subseteq\mathbb{S}^{d-1}$. □

Applying Lemma 3 inductively and using a martingale argument, the authors prove that intersecting $j$ random caps and $(k-j)$ random anticaps, we get a multiplicative fluctuation for $\sigma(L_S)$ around $p^{|S|}(1-p)^{n-1-|S|}$ that is of the order of $(1\pm\sqrt{j}\delta+\sqrt{k-j}\frac{p}{1-p}\delta)$. Going back to Eq.(19), this approach is sufficient to prove that

$$\mathrm{TV}(G(n,p,d),G(n,p)) = o_n\Big(\frac{n^3p^2}{d}\Big),$$

leading to the first statement of Theorem 7.

Step 4. The sparse case and the use of the cavity method.
To get down to a polylogarithmic threshold in the sparse regime, the authors change of paradigm. Previously, they were bounding the quantity

$$\mathbb{P}_{G\sim G(n,p,d)}(N_G(n)=S\,|\,G_{n-1}) = \mathbb{E}_{X_1,\dots,X_{n-1}\,|\,G_{n-1}}\mathbb{E}_{X_n\sim\sigma}\big[\mathbb{1}_{N_G(n)=S}\big] = \mathbb{E}_{X_1,\dots,X_{n-1}\,|\,G_{n-1}}\big[\sigma(L_S)\big], \tag{21}$$

by fixing a specific realization of latent positions $X_1,\dots X_{n-1}$ and then analyzing the probability that the node $n$ connects to some $S\subseteq[n-1]$. The probability that vertex $n$ is adjacent to all vertices in $S\subseteq[n-1]$ is exactly equal to the measure of the set-caps intersection, which appears to be tight. At a high level, this is a "*worst case approach*" to upper bound Eq.(21) in the sense that the bound obtained from this analysis may be due to an unlikely latent configuration conditioned on $X_1,\dots,X_{n-1}$ producing $G_{n-1}$. To obtain a polylogarithmic threshold in the sparse case, one needs to analyze the concentration of $\sigma(L_S)$ on average over vector embeddings of $G_{n-1}$. To do so, the authors rely on the so-called *cavity method* borrowed from the field of statistical physics. The cavity method allows to understand the distribution of $(X_i)_{i\in S}$ conditional on forming $G_{n-1}$ for any $S\subseteq[n-1]$ with size of the order $pn=\Theta(1)$. We provide further details on this approach in the following.

**A simplification using tight concentration for intersections involving anti-caps.** Liu et al. (2021) first prove that due to tight concentration for the measure of the intersection of random anticaps with sets of lower bounded measure, one can get high-probability estimates for $\nu_n(S \mid G_{n-1})$ by studying the probability that $S \subseteq N_G(n)$, namely

$$\mathbb{P}(S \subseteq N_G(n) \mid G_{n-1}) = \mathbb{P}\big(\forall i \in S,\ \langle X_i, X_n\rangle \geq t_{p,d} \mid G_{n-1}\big) = \mathop{\mathbb{E}}_{\substack{X_n \sim \sigma \\ (X_i)_{i\in[n-1]} \sim \sigma^{G_{n-1}}}} \prod_{i\in S} \mathbb{1}_{\langle X_i, X_n\rangle \geq t_{p,d}}, \tag{22}$$

where $\sigma^{G_{n-1}} := \big[\sigma^{\otimes(n-1)} \mid G_{n-1}\big]$. If $(X_i)_{i\in S}$ in Eq.(22) was a collection of independent random vectors distributed uniformly on the sphere then Eq.(22) would be exactly equal to $p^{|S|}$. In the following, we explain how the authors prove that both of these properties are approximately true.

**The cavity-method.** To bound the fluctuation of Eq.(22) around $p^{|S|}$, Liu et al. (2021) use the cavity-method. Let us consider $S \subseteq [n-1]$, $G_{n-1}$ sampled from $G(n-1,p,d)$ and its corresponding latent vectors. Let us denote by $\mathscr{B}_{G_{n-1}}(i,\ell)$ the ball of radius-$\ell$ around a vertex $i \in [n-1]$ in the graph $G_{n-1}$. Fixing all vectors except those in $K := \bigcup_{i\in S} \mathscr{B}_{G_{n-1}}(i,\ell-1)$, the cavity method aims at computing the joint distribution of $(X_i)_{i\in S}$ conditional to $(X_i)_{i\notin K}$ and $G_{n-1}$. Informally speaking, we "carve out" a cavity of depth $\ell$ around each vertex $i \in S$ and we fix all latent vectors outside of these cavities as presented with Figure 5. The choice of the depth $\ell$ results from the following tradeoff:

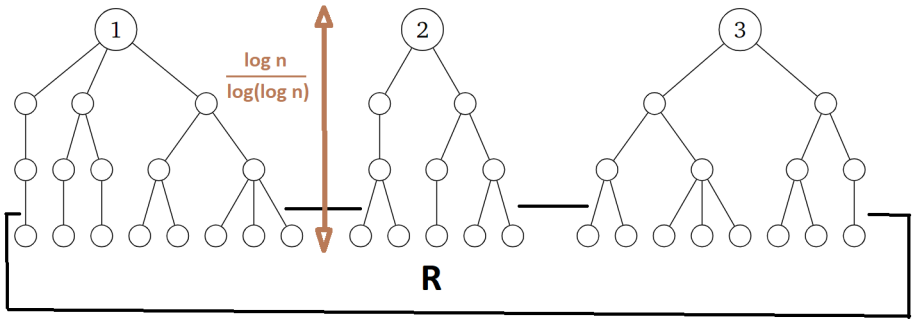


Figure 5: Illustration of the cavity method to bound the fluctuation of Eq.(22) around $p^{|S|}$ *i.e.,* to bound the deviation of the random variable $\sigma(L_S)$ conditioned on $X_1,\dots,X_{n-1}$ producing $G_{n-1}$. With high probability, the neighbourhood until depth $\ell = \frac{\log n}{\log\log n}$ of vertices in $S$ are disjoint trees. We fix the latent representation of vertices in the set $R := [n-1]\backslash K$. Using the Belief Propagation algorithm, one can compute the distribution of $(X_i)_{i\in S} \mid (X_j)_{j\in R}$ where the latent positions $(X_j)_{j\in[n-1]}$ are sampled according to $\sigma^{G_{n-1}}$. This allows to bound the fluctuation of Eq.(22) around $p^{|S|}$.

- We want to choose the depth $\ell$ small enough so that the balls $\mathscr{B}_{G_{n-1}}(i,\ell)$ for $i \in S$ are all trees and are pairwise disjoint with high probability.
- We want to choose $\ell$ as large as possible in order to get a bound on the fluctuations of Eq.(22) around $p^{|S|}$ as small as possible.

To formally analyze the distribution of the unfixed vectors upon resampling them, the authors set up a constraint satisfaction problem instance over a continuous alphabet that encodes the edges of $G_{n-1}$ within the trees around $S$: each node has a vector-valued variable in $\mathbb{S}^{d-1}$, and the constraints are that nodes joined by an edge must have vectors with inner product at least $t_{p,d}$. The marginal of the latent vectors $X_i$ for $i \in S$ can be obtained using the Belief-Propagation algorithm. Let us recall that Belief-Propagation computes marginal distributions over labels of constraints satisfaction problems when the constraints graph is a tree.

**A simple analysis of Belief-Propagation.** To ease the reasoning, let us suppose that $1 \in S$ is such that $\mathscr{B}_{G_{n-1}}(1,\ell-1)$ is a path. Without loss of generality, we consider that the path is given by

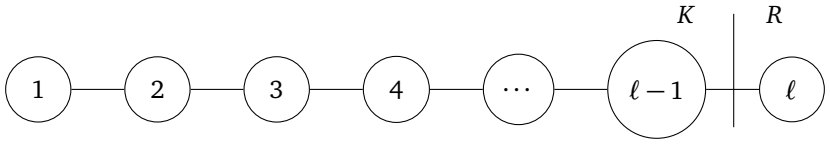


Figure 6: Simple analysis of the Belief Propagation algorithm when the neighbourhood of vertex $1 \in S$ at depth $\ell$ is a path.

Figure 6. Every vector is passing to its parent along the path a convolution of its own measure (corresponding to its "message") with a cap of measure $p$. Denoting by $P$ the linear operator defined so that for any function $h : \mathbb{S}^{d-1} \to \mathbb{R}$,

$$Ph(x) = \frac{1}{p} \int_{\mathrm{cap}(x)} h(y) d\sigma(y),$$

the authors prove that for some $a > 0$, for any probability measure $\mu$ on $\mathbb{S}^{d-1}$ with density $h$ with respect to $\sigma$,

$$\mathrm{TV}(Ph, \sigma) \leq \mathcal{O}_n\Big(\frac{\log^a n}{\sqrt{d}}\Big)\mathrm{TV}(\mu, \sigma), \tag{23}$$

which is a contraction result. Since at every step of the Belief Propagation algorithm, a vertex sends to its parent the image by the operator $P$ of its own measure, we deduce from Eq.(23) that the parent receives a measure which is getting closer to the uniform distribution by a multiplicative factor equal to $\frac{1}{\sqrt{d}}$. The proof of Eq.(23) relies on the set-cap intersection concentration result (see Lemma 3). To get an intuition of this connection, let us consider that $h$ is the density of the uniform probability measure $\mu$ on some set $L \subseteq \mathbb{S}^{d-1}$, then

$$Ph(x) = \frac{1}{p}\mathbb{P}_{Y\sim\mu}(Y \in \mathrm{cap}(x)) = \frac{1}{p}\frac{\sigma(L \cap \mathrm{cap}(x))}{\sigma(L)},$$

and we can conclude using Lemma 3 that ensures that with high probability over $x \sim \sigma$, $\sigma(L \cap \mathrm{cap}(x)) = (1 \pm \mathcal{O}_n(\frac{\log^a n}{\sqrt{d}}))p\sigma(L)$. Applying Eq.(23) $\ell = \frac{\log n}{\log\log n}$ times for $d$ being some power of $\log n$, one can show that,

$$\mathrm{TV}(P^\ell \mu, \sigma) = \mathcal{O}_n\Big[\Big(\frac{\log^a n}{\sqrt{d}}\Big)^\ell\Big] = o_n\Big(\frac{1}{\sqrt{n}}\Big).$$

With this approach, one can prove that the distribution of $(X_i)_{i\in S}$ is approximately $\sigma^{\otimes|S|}$. This allows to bound the fluctuations of Eq.(22) around $p^{|S|}$ which leads to Theorem 6 using Eqs.(18) and (19).

As a concluding remark, we mention that Liu et al. (2021) demonstrate a coupling of $G_- \sim G(n, p - o_n(p))$, $G \sim G(n, p, d)$, and $G_+ \sim G(n, p + o_n(p))$ that satisfies $G_- \subseteq G \subseteq G_+$ with high probability. This sandwich-type result holds for a proper choice of the latent dimension and allows to transfer known properties of Erdös-Renyi random graphs to RGGs in the studied regime. For example, the authors use this coupling result to upper bound the probability that the depth-$\ell$ neighborhood of some $i \in [n]$ forms a tree under $G(n, p, d)$ in the sparse regime with $d = \mathrm{polylog}(n)$.

## References


Abbe, E. (2018). Community detection and Stochastic Block Models. Foundations and Trends in Communications and Information Theory, 14(1-2):1–162.

Aguilar-Sánchez, R., Méndez-Bermúdez, J. A., Rodrigues, F. A., and Sigarreta, J. M. (2020). Topological versus spectral properties of Random Geometric Graphs. Phys. Rev. E, 102:042306.

Allen-Perkins, A. (2018). Random Spherical Graphs. Physical Review E, 98(3):032310.

Araya Valdivia, E. (2020). Random Geometric Graphs on Euclidean Balls. arXiv e-prints, pages arXiv–2010.

Araya Valdivia, E. and De Castro, Y. (2019). Latent distance estimation for Random Geometric Graphs. In Wallach, H., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E., and Garnett, R., editors, Advances in Neural Information Processing Systems, volume 32, pages 8724–8734. Curran Associates, Inc.

Arcones, M. A. and Gine, E. (1993). Limit theorems for U-Processes. Ann. Probab., 21(3):1494–1542.

Arlot, S. (2019). Minimal penalties and the slope heuristics: a survey. Journal de la Société Française de Statistique, 160(3):1–106.

Avrachenkov, K. and Bobu, A. (2020). Cliques in high-dimensional Random Geometric Graphs. In Cherifi, H., Gaito, S., Mendes, J. F., Moro, E., and Rocha, L. M., editors, Complex Networks and Their Applications VIII, pages 591–600, Cham. Springer International Publishing.

Avrachenkov, K., Bobu, A., and Dreveton, M. (2021). Higher-order spectral clustering for geometric graphs. Journal of Fourier Analysis and Applications, 27(2).

Bandeira, A. S. and van Handel, R. (2016). Sharp nonasymptotic bounds on the norm of random matrices with independent entries. The Annals of Probability, 44(4):2479–2506.

Barabási, A.-L. (2009). Scale-free networks: a decade and beyond. science, 325(5939):412–413.

Barthélemy, M. (2011). Spatial networks. Physics Reports, 499(1-3):1–101.

Ben Arous, G., Gheissari, R., and Jagannath, A. (2020). Algorithmic thresholds for tensor PCA. Annals of Probability, 48:2052–2087.

Benaych-Georges, F., Bordenave, C., Knowles, A., et al. (2019). Largest eigenvalues of sparse inhomogeneous Erdős–Rényi graphs. Annals of Probability, 47(3):1653–1676.

Benaych-Georges, F., Bordenave, C., Knowles, A., et al. (2020). Spectral radii of sparse random matrices. In Annales de l'Institut Henri Poincaré, Probabilités et Statistiques, volume 56, pages 2141–2161. Institut Henri Poincaré.

Blackwell, P., Edmondson-Jones, M., and Jordan, J. (2007). Spectra of adjacency matrices of Random Geometric Graphs. University of Sheffield. Department of Probability and Statistics.

Bollobás, B. (2001). Random Graphs. Cambridge Studies in Advanced Mathematics. Cambridge University Press, 2 edition.

Bordenave, C., Lelarge, M., and Massoulié, L. (2015). Non-backtracking spectrum of random graphs: community detection and non-regular Ramanujan graphs. In 2015 IEEE 56th Annual Symposium on Foundations of Computer Science, pages 1347–1357. IEEE.

Brennan, M. and Bresler, G. (2019). Optimal average-case reductions to sparse PCA: From weak assumptions to strong hardness. arXiv preprint arXiv:1902.07380.

Brennan, M. and Bresler, G. (2020). Reducibility and statistical-computational gaps from secret leakage. In Conference on Learning Theory, pages 648–847. PMLR.

Brennan, M., Bresler, G., and Huang, B. (2021). De finetti-style results for Wishart matrices: Combinatorial structure and phase transitions.

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

Bresler, G. and Nagaraj, D. (2018). Optimal single sample tests for structured versus unstructured network data. In Bubeck, S., Perchet, V., and Rigollet, P., editors, Proceedings of the 31st Conference On Learning Theory, volume 75 of Proceedings of Machine Learning Research, pages 1657–1690. PMLR.

Breu, H. and Kirkpatrick, D. G. (1998). Unit disk graph recognition is NP-hard. Computational Geometry, 9(1):3 – 24. Special Issue on Geometric Representations of Graphs.

Bubeck, S., Ding, J., Eldan, R., and Racz, M. Z. (2016). Testing for high-dimensional geometry in random graphs. Random Structures & Algorithms, 49:503–532.

Bubeck, S. and Ganguly, S. (2015). Entropic CLT and phase transition in high-dimensional Wishart matrices. CoRR, abs/1509.03258.

Channarond, A. (2015). Random graph models: an overview of modeling approaches. Journal de la Société Française de Statistique, 156(3):56–94.

Chatterjee, S. (2015). Matrix estimation by universal singular value thresholding. The Annals of Statistics, 43(1).

Clémençon, S., Lugosi, G., and Vayatis, N. (2008). Ranking and Empirical Minimization of U-statistics. The Annals of Statistics, 36(2):844 – 874.

Dai, F. and Xu, Y. (2013). Approximation theory and harmonic analysis on spheres and balls, volume 23. Springer.

Dall, J. and Christensen, M. (2002). Random Geometric Graphs. Phys. Rev. E, 66:016121.

De Castro, Y., Lacour, C., and Ngoc, T. M. P. (2020). Adaptive estimation of nonparametric Geometric Graphs.

De la Pena, V. and Giné, E. (2012). Decoupling: from dependence to independence. Springer Science & Business Media.

Decelle, A., Krzakala, F., Moore, C., and Zdeborová, L. (2011). Asymptotic analysis of the Stochastic Block Model for modular networks and its algorithmic applications. Phys. Rev. E, 84:066106.

Dettmann, C. P. and Georgiou, O. (2016). Random Geometric Graphs with general connection functions. Physical Review E, 93(3).

Dettmann, C. P., Georgiou, O., and Knight, G. (2017). Spectral statistics of Random Geometric Graphs. EPL (Europhysics Letters), 118(1):18003.

Devroye, L., György, A., Lugosi, G., and Udina, F. (2011). High-dimensional Random Geometric Graphs and their clique number. Electron. J. Probab., 16:2481–2508.

Diaconis, P. and Janson, S. (2007). Graph limits and exchangeable random graphs. arXiv preprint arXiv:0712.2749.

Duchemin, Q. (2022). Reliable Time Prediction in the Markov Stochastic Block Model. working paper or preprint.

Duchemin, Q. and De Castro, Y. (2022). Markov Random Geometric Graph (MRGG): A Growth Model for Temporal Dynamic Networks. Electronic Journal of Statistics, 16(1):671 – 699.

Duchemin, Q., De Castro, Y., and Lacour, C. (2022). Concentration inequality for U-statistics of order two for uniformly ergodic Markov chains. Bernoulli.

Eldan, R. (2015). An efficiency upper bound for inverse covariance estimation. Israel Journal of Mathematics, 207(1):1–9.

Eldan, R. and Mikulincer, D. (2020). Information and dimensionality of anisotropic Random Geometric Graphs. In Geometric Aspects of Functional Analysis, pages 273–324. Springer.

Eldan, R., Mikulincer, D., and Pieters, H. (2020). Community detection and percolation of information in a geometric setting. arXiv preprint arXiv:2006.15574.

Erba, V., Ariosto, S., Gherardi, M., and Rotondo, P. (2020). Random Geometric Graphs in high dimension. arXiv preprint arXiv:2002.12272.

Estrada, E. and Sheerin, M. (2016). Consensus dynamics on Random Rectangular Graphs. Physica D: Nonlinear Phenomena, 323-324:20 – 26. Nonlinear Dynamics on Interconnected Networks.

Fromont, M. and Laurent, B. (2006). Adaptive goodness-of-fit tests in a density model. Ann. Statist., 34(2):680–720.

Galhotra, S., Mazumdar, A., Pal, S., and Saha, B. (2017). The Geometric Block Model. arXiv preprint arXiv:1709.05510.

Gao, C. and Lafferty, J. (2017). Testing network structure using relations between small subgraph probabilities. arXiv preprint arXiv:1704.06742.

Ghoshdastidar, D., Gutzeit, M., Carpentier, A., Von Luxburg, U., et al. (2020). Two-sample hypothesis testing for inhomogeneous random graphs. Annals of Statistics, 48(4):2208–2229.

Gilbert, E. N. (1961). Random plane networks. Journal of the Society for Industrial and Applied Mathematics, 9(4):533–543.

Giné, E., Latała, R., and Zinn, J. (2000). Exponential and moment inequalities for U-Statistics. High Dimensional Probability II, page 13–38.

Goel, A., Rai, S., and Krishnamachari, B. (2005). Monotone properties of Random Geometric Graphs have sharp thresholds. Ann. Appl. Probab., 15(4):2535–2552.

Grygierek, J. and Thäle, C. (2020). Gaussian fluctuations for edge counts in high-dimensional Random Geometric Graphs. Statistics & Probability Letters, 158:108674.

Haenggi, M., Andrews, J. G., Baccelli, F., Dousse, O., and Franceschetti, M. (2009). Stochastic geometry and random graphs for the analysis and design of wireless networks. IEEE Journal on Selected Areas in Communications, 27(7):1029–1046.

Higham, D., Rasajski, M., and Przulj, N. (2008). Fitting a geometric graph to a protein-protein interaction network. Bioinformatics (Oxford, England), 24:1093–9.

Houdré, C. and Reynaud-Bouret, P. (2002). Exponential inequalities for U-statistics of order two with constants. Stochastic Inequalities and Applications. Progress in Probability, 56.

Issartel, Y., Giraud, C., and Verzelen, N. (2021). Optimal embedding on the sphere in non-parametric latent space models.

Jin, J., Ke, Z. T., and Luo, S. (2019). Optimal adaptivity of signed-polygon statistics for network testing.

Joly, E. and Lugosi, G. (2016). Robust estimation of U-statistics. Stochastic Processes and their Applications, In Memoriam: Evarist Giné:3760–3773.

Jordan, J. and Wade, A. R. (2015). Phase Transitions for Random Geometric Preferential Attachment Graphs. Advances in Applied Probability, 47(2):565–588.

Klopp, O. and Verzelen, N. (2019). Optimal graphon estimation in cut distance. Probability Theory and Related Fields, 174(3):1033–1090.

Koltchinskii, V. and Giné, E. (2000). Random Matrix Approximation of Spectra of Integral Operators. Bernoulli, 6.

Kontorovich, A. and Raginsky, M. (2017). Concentration of Measure Without Independence: A Unified Approach Via the Martingale Method, pages 183–210. Springer New York.

Krioukov, D., Papadopoulos, F., Kitsak, M., Vahdat, A., and Boguná, M. (2010). Hyperbolic geometry of complex networks. Physical Review E, 82(3):036106.

Le, C. M., Levina, E., and Vershynin, R. (2018). Concentration of random graphs and application to community detection. World Scientific.

Liu, S., Mohanty, S., Schramm, T., and Yang, E. (2021). Testing thresholds for high-dimensional sparse random geometric graphs. ArXiv, abs/2111.11316.

Liu, S. and Racz, M. Z. (2021a). Phase transition in noisy high-dimensional Random Geometric Graphs.

Liu, S. and Racz, M. Z. (2021b). A probabilistic view of latent space graphs and phase transitions.

Lovász, L. (2012). Large networks and graph limits, volume 60. American Mathematical Soc.

Lunagómez, S., Mukherjee, S., Wolpert, R. L., and Airoldi, E. M. (2017). Geometric Representations of Random Hypergraphs. Journal of the American Statistical Association, 112(517):363–383.

Mao, G. and Anderson, B. (2012). Connectivity of large wireless networks under a general connection model. IEEE Transactions on Information Theory, 59.

Mossel, E., Neeman, J., and Sly, A. (2014). Reconstruction and estimation in the planted partition model. Probability Theory and Related Fields, 162.

Méliot, P.-L. (2019). Asymptotic representation theory and the spectrum of a Random Geometric Graph on a compact Lie group. Electron. J. Probab., 24:85 pp.

Müller, T. and Prałat, P. (2015). The acquaintance time of (percolated) Random Geometric Graphs. European Journal of Combinatorics, 48:198 – 214. Selected Papers of EuroComb'13.

Nyberg, A., Gross, T., and Bassler, K. E. (2015). Mesoscopic structures and the Laplacian spectra of Random Geometric Graphs. Journal of Complex Networks, 3(4):543–551.

Ostilli, M. and Bianconi, G. (2015). Statistical mechanics of Random Geometric Graphs: Geometry-induced first-order phase transition. Physical Review E, 91(4).

Papadopoulos, F., Kitsak, M., Serrano, M., Boguna, M., and Krioukov, D. (2012). Popularity versus similarity in growing networks. Nature, 489(7417):537–540.

Péché, S. and Perchet, V. (2020). Robustness of community detection to random geometric perturbations. Advances in Neural Information Processing Systems, 33.

Penrose, M. et al. (2003). Random Geometric Graphs, volume 5. Oxford university press.

Penrose, M. D. (2016). Connectivity of Soft Random Geometric Graphs. The Annals of Applied Probability, 26(2):986–1028.

Pereda, M. and Estrada, E. (2019). Visualization and machine learning analysis of complex networks in hyperspherical space. Pattern Recognit., 86:320–331.

Perry, S., Yin, M. S., Gray, K., and Kobourov, S. (2020). Drawing graphs on the sphere. In Proceedings of the International Conference on Advanced Visual Interfaces, pages 1–9.

Preciado, V. and Jadbabaie, A. (2009). Spectral analysis of virus spreading in random geometric networks. Proceedings of the 48h IEEE Conference on Decision and Control (CDC) held jointly with 2009 28th Chinese Control Conference, pages 4802–4807.

Racz, M. and Bubeck, S. (2016). Basic models and questions in statistical network analysis. Statistics Surveys, 11.

Rai, S. (2004). The spectrum of a Random Geometric Graph is concentrated. Journal of Theoretical Probability, 20.

Sankararaman, A. and Baccelli, F. (2017). Community detection on Euclidean random graphs. In 2017 55th Annual Allerton Conference on Communication, Control, and Computing (Allerton), pages 510–517.

Smith, A. L., Asta, D. M., and Calder, C. A. (2019). The geometry of continuous latent space models for network data. Statist. Sci., 34(3):428–453.

Solovey, K., Salzman, O., and Halperin, D. (2018). New perspective on sampling-based motion planning via Random Geometric Graphs. The International Journal of Robotics Research, 37(10):1117–1133.

Tropp, J. A. (2015). An introduction to matrix concentration inequalities. Found. Trends Mach. Learn., 8(1–2).

Walters, M. (2011). Random geometric graphs, page 365–402. London Mathematical Society Lecture Note Series. Cambridge University Press.

Wang, G. and Lin, Z. (2014). On the performance of multi-message algebraic gossip algorithms in dynamic Random Geometric Graphs. IEEE Communications Letters, PP:1–1.

Watts, D. J. and Strogatz, S. H. (1998). Collective dynamics of 'small-world' networks. nature, 393(6684):440–442.

Wolfe, P. J. and Olhede, S. C. (2013). Nonparametric graphon estimation. arXiv e-prints, page arXiv:1309.5936.

Xie, Z., Ouyang, Z., Liu, Q., and Li, J. (2016). A geometric graph model for citation networks of exponentially growing scientific papers. Physica A: Statistical Mechanics and its Applications, 456:167–175.

Xie, Z. and Rogers, T. (2016). Scale-invariant geometric random graphs. Physical Review E, 93(3).

Xie, Z., Zhu, J., Kong, D., and Li, J. (2015). A Random Geometric Graph built on a time-varying Riemannian manifold. Physica A: Statistical Mechanics and its Applications, 436:492 – 498.

Xu, J. (2018). Rates of convergence of spectral methods for graphon estimation. In International Conference on Machine Learning, pages 5433–5442. PMLR.

Zuev, K., Boguna, M., Bianconi, G., and Krioukov, D. (2015). Emergence of soft communities from Geometric Preferential Attachment. Scientific reports, 5.