Electronic Journal of Statistics
Vol. 18 (2024) 335–354
ISSN: 1935-7524
https://doi.org/10.1214/23-EJS2169

Consistency of maximum likelihood for continuous-space network models I

Cosma Shalizi

Departments of Statistics and of Machine Learning
Carnegie Mellon University
Pittsburgh, PA 15213
USA
Santa Fe Institute
1399 Hyde Park Road
Santa Fe, NM 87501
USA
e-mail: cshalizi@cmu.edu

and

Dena Asta

Department of Statistics
Ohio State University
Columbus, OH 43210
USA
e-mail: dasta@stat.osu.edu

Abstract: A very popular class of models for networks posits that each node is represented by a point in a continuous latent space, and that the probability of an edge between nodes is a decreasing function of the distance between them in this latent space. We study the *embedding problem* for these models, of recovering the latent positions from the observed graph. Assuming certain natural symmetry and smoothness properties, we establish the uniform convergence of the log-likelihood of latent positions as the number of nodes grows. A consequence is that the maximum likelihood embedding converges on the true positions in a certain information-theoretic sense. Extensions of these results, to recovering distributions in the latent space, and so distributions over arbitrarily large graphs, will be treated in the sequel.

MSC2020 subject classifications: 62G05, 91D30.
Keywords and phrases: Consistency, network models, latent space models, graph embedding, MLE.

Received June 2022.

Contents

arXiv: 1711.02123

1. Introduction

The statistical analysis of network data, like other sorts of statistical analysis, models the data we observe as the outcome of stochastic processes, and rests on inferring aspects of those processes from their results. It is essential that the methods of inference be consistent, that as they get more and more information, they should come closer and closer to the truth. In this paper, we address the consistency of non-parametric maximum likelihood estimation for a popular class of network models, those based on continuous latent spaces.

In these models, every node in the network corresponds to a point in a latent, continuous metric space, and the probability of an edge or tie between two nodes is a decreasing function of the distance between their points in the latent space. These models are popular because they are easily interpreted in very plausible ways, and often provide good fits to data. Moreover, they have extremely convenient mathematical and statistical properties: they lead to exchangeable, projectively-consistent distributions over graphs; the comparison of two networks reduces to comparing two clouds of points in the latent space, or even to comparing two densities therein; it is easy to simulate new networks from the estimated model for purposes of bootstrapping, etc. While the latent space has typically been taken to be a low-dimensional Euclidean space [11], recent work has suggested that in many applications it would be better to take the space to non-Euclidean, specifically negatively curved or hyperbolic [15, 5] and positively curved [17, 26].

We can estimate continuous latent space models in the sense of an *embedding*: given an observed graph, we wish to work backwards the locations of the nodes in the latent space, i.e., to "embed" the graph in the latent space. The most straightforward method of embedding is a *maximum likelihood estimator* (MLE), treating the latent position of each node as a parameter (or vector of parameters). While it is straightforward to *say* that a good embedding should converge on the true coordinates as the number of nodes $n \to \infty$, making this mathematically precise is somewhat tricky. (We would, for example, need to define a metric on the space of embeddings of graphs of different sizes.) Instead, we prove the next best thing: that for continuous latent space models of sufficient symmetry and tameness, the distribution over graphs implied by the MLE converges, in normalized Kullback-Leibler (KL) divergence, to the distribution implied by the true embedding. That is, the MLE becomes statistically indistinguishable, in its observable consequences, from the truth. This is a consequence of a result we establish along the way, about the uniform convergence of normalized log-likelihoods to their expectation values (Theorem 4); the rate of this uniform convergence upper-bounds the rate at which the MLE approaches the true embedding in KL divergence.

In the sequel in preparation, we combine our results about normalized log-likelihood with the construction of a specific class of metrics on growing sequences of embeddings, to establish a more conventional, coordinate-wise notion of consistency, and consistency for a subsequent estimator of the node density in the latent space.

Section 2 reviews background on continuous latent space models of networks. Section 3 states our main results, along with certain technical assumptions, and observes that these results generalize to mis-specified models. All proofs, and a number of subsidiary results and lemmas, are deferred to Section 5.

2. Background

In many, though not all, network data-analysis situations, we have only one network — perhaps not even all of that one network — from which we nonetheless want to draw inferences about the whole data-generating process. This clearly will require a law of large numbers or ergodic theorem to ensure that a single large sample is representative of the whole process. The network, however, is a single high-dimensional object whose every part is dependent on every other part. This is also true of time-series and spatial data, but there we can often use the fact that distant parts of the data should be nearly independent of each other. While general networks often exhibit such decay, networks in nature often lack a natural, exogenous sense of distance (in the technical, geometric sense) that explains such decay.

Continuous latent space (CLS) models are precisely generative models for networks which exhibit just such an exogenous sense of distance. Each node is represented as a location in a continuous metric space, the **latent space**. Conditional on the vector of all node locations, the probability of an edge between two nodes is a decreasing function of the distance between their locations, and all edges are independent. Generative models for networks for which the existence of different edges is conditionally independent with respect to some latent quantity μ are common; however CLS models, at least as taken in this paper, are distinguished by the particular geometric form that μ takes.

As mentioned above, the best-known CLS model for social networks is that of Hoff, Raftery and Handcock [11], where the metric space is taken to be Euclidean, and node locations are assumed to be drawn iidly from a Gaussian distribution. In random geometric graphs [18], the locations are drawn iidly from a distribution on a metric space possibly more general than Euclidean space and the probabilities of connecting edges are either 0 or 1 based on a threshold.

As also mentioned above, there is more recent work which indicates that for some applications it would be better to let the latent space be negatively curved, i.e. hyperbolic [1, 12, 15]. Remarkably, negatively curved spaces yield networks that simultaneously exhibit both local and global features found in real-world networks [15]. Many real-world networks show highly skewed degree distributions exhibiting power laws, very short path lengths, a division into a core and peripheries where short paths between peripheral nodes "bend back" towards the core, and a hierarchical organization of clustering. Thus if the latent space is chosen to be a certain hyperboloid, one naturally obtains graphs exhibiting all these properties [13, 15]. Inference of "true" coordinates in a negatively curved space leads to concrete applications, for example in routing (e.g. [9]) and in hypothesis testing for differences or changes in network structure salient to social network analaysis [5].

Consistent inference of node coordinates as networks grow in size provides a consistent method of inferring features of a large, random network from *partial observations*. Often in practice, one only observes a subnetwork of some impractically large or theoretically infinite network. Also often in practice, such as in the case of a large class of popular ERGM mopdels, inference on those subnetworks does not aymptotically agree with inference on the entire network [20]. Consistent inference of node coordinates in negatively curved spaces can thus be used to obtain consistent estimates on, say, hierarchical clusters, degree distributions and hypothesis test statistics (e.g. [5]) based on partial observations.

The CLS models we have mentioned so far have presumed that node locations follow tractable, parametric families in the latent space. This is mathematically inessential — many of the results carry over perfectly well to arbitrary densities — and scientifically unmotivated. Because CLS models may need very different spaces depending on applications, we investigate consistency of nonparametric estimation for them at a level of generality which abstracts away from many of the details of particular spaces and their metrics.

To the best of our knowledge, there are no results in the existing literature on the consistency of embedding for CLS models where edge probabilties vary continuously with distance.[1]

[1] Computationally-tractable and consistent embedding algorithms exist for some kinds of random geometric graph where edges are deterministically present between sufficiently-close nodes and otherwise deterministically absent [10], but they rely crucially on deterministic links, and their statistical efficiency is unknown. Uniform consistency for variants of these sorts of

3. Geometric network inference

Our goal is to show that when the continuous latent space model is sufficiently smooth, and the geometry of the latent space is itself sufficiently symmetric, then the maximum-likelihood embedding of a graph converges, in normalized Kullback-Leibler divergence, to the true locations of the nodes (Theorem 4). As an intermediate step, we show the uniform convergence of normalized log-likelihoods on their expectation values, at an explicit rate, which also gives us the rate of KL convergence of the MLE on the truth. All proofs are postponed to Section 5.

3.1. Setting and conventions

We consider only simple, undirected, unlabeled graphs; we will write a random graph as G, and will sometimes abuse notation to also write G for the adjacency matrix, so that $G_{pq} = G_{qp} = 1$ if there is an edge between nodes p and q, and $= 0$ otherwise.

All random graphs G in this paper have conditionally independent edges; that is, we assume for each G, there exists a random quantity μ such that $G \mid \mu$ has independently distributed edges. A *continuous latent space model* assumes that μ has a certain geometric nature, which will be defined in the succeeding paragraphs.

All the metrics of metric spaces will be denoted by *dist*; context will make clear which metric *dist* is describing. Our model for generating random graphs begins with a metric measure space M, a metric space equipped with a Borel measure, and the corresponding group $isom(M)$ of measure-preserving isometries $M \cong M$. Every node is located at (equivalently, "represented by" or "labeled with") a point in M, x_i for the i^{th} node; the location of the first n nodes is $x_{1:n} \in M^n$, and a countable sequence of locations will be $x_{1:\infty}$. For each n, there is a non-increasing **link function** $w_n : [0, \infty) \mapsto [0, 1]$, and nodes i and j are joined by an edge with probability $w_n(dist(x_i, x_j))$. By a *latent space* $(M, w_{1:\infty})$, we will mean the combination of M and a sequence $w_{1:\infty}$ of link functions $w_1, w_2, \ldots$. We simply write $\mathrm{graph}_n(x_{1:n})$ for the distribution of a random graph on n vertices located at $x_{1:n}$ when the latent space is understood from context. In other words, $\mathrm{graph}_n(x_{1:n})$ regarded as adjacency matrix is the random (undirected) symmetric (nxn) matrix with conditionally independent entries (conditioned on $x_{1:n}$) with $(\mathrm{graph}_n(x_{1:n}))_{ij}$ drawn from a Bernoulli random variable with parameter $w_n(dist(x_i, x_j))$ for distinct i, j and 0 for $i = j$. Thus in the particular case $G = \mathrm{graph}_n(x_{1:n})$, we have $\mu = x_{1:n}$.

It is clear that for any $\phi \in isom(M)$, we have for every n,

$$\mathrm{graph}_n(x_{1:n}) \overset{d}{=} \mathrm{graph}_n\big(\phi(x_{1:n})\big) \tag{3.1}$$

CLS models, such as *random dot product graphs* (RDPG) [25], have been well established (cf. [6]); the inherently linear algebraic methods used to develop estimators and consistency results in the RDPG setting do not seem portable in the metric setting.

Accordingly, we will use $[x_{1:n}]$ to indicate the equivalence class of n-tuples in M^n carried by isometries to $x_{1:n}$; the metric on M extends to these isometry classes in the natural way,

$$dist\big([x_{1:n}], [y_{1:n}]\big) = \inf_{\phi \in isom(M)} \sum_{i=1}^{n} dist\big(x_i, \phi(y_i)\big). \tag{3.2}$$

We cannot hope to find $x_{1:n}$ by observing the graph it leads to, but we can hope to identify $[x_{1:n}]$.

Conventions When n and m are integers, $n < m$, $n : m$ will be the set $\{n, n+1, \ldots m-1, m\}$. Unless otherwise specified, all limits will be taken as $n \to \infty$. All probabilities and expectations will be taken with respect to the actual generating distribution of G.

3.2. Axioms on the generative model

We recall that a metric space M is *k-homogeneous* if every isometry between finite submetric spaces each of size k of M extends to an isometry on M, an isometry $M \to M$. There we call a metric space M *∞-homogeneous* if every isometry between finite submetric spaces of M extends to an isometry on M. The literature takes *homogeneous* to usually mean 1-homogeneous but to sometimes to mean ∞-homogeneous. Motivating examples are Euclidean space $\mathbb{R}^d$ and the *Poincaré Halfplane* $\mathbb{H}_2$, described in Section 3.3. Almost any example of a metric space with a single "singularity" x_1, such as a "figure 8," is not 1-homogeneous; for a close enough point x_2, there are also points x_3, x_4 such that $dist(x_1, x_2) = dist(x_3, x_4)$, but intuitively there cannot be any isometry carrying a singularity to a non-singularity. An example of a metric space that is 1-homogeneous but not ∞-homogeneous is the orientable surface of infinite genus.

Identifiability of graph distributions determined by certain CLS models is possible. We define such CLS models below.

Definition 1. A laiiitent space $(M, w_{1:\infty})$ is **regular** when:

1. M is a complete ∞-homogeneous Riemannian manifold;
2. The group of isometries on M has only finitely-many connected components;
3. The function w_n is injective and smooth for each n; and
4. The sequence $w_{1:\infty}$ satisifes $-v_n \leqslant \sup_{x,y \in M} \operatorname{logit} w_n(dist(x, y)) \leqslant v_n$ for some $v_n \in o(\sqrt{n})$.

Proposition 2. The metric spaces $\mathbb{R}^d$ and $\mathbb{H}_2$ satisfy points (1) and (2) of Definition 1 with

$$B_{\mathbb{H}_2} = B_{\mathbb{R}^d} = 2.$$

where B_M denotes the number of connected components of the group of isometries on a metric space M.

Demanding that $v_n = o(\sqrt{n})$ is done with an eye towards the needs of the proofs in Section 5. Some common examples of link functions (cf. [15]) include the following two kinds:

$$w_n(t) = \begin{cases} 1 & t \leqslant \ln n \\ 0 & t > \ln n \end{cases} \qquad w_n(t) = \frac{1}{1 + e^{(T_n^{-1}/2)(t - \ln n)}} \tag{3.3}$$

The first sort defines a graph where edges are deterministically present between sufficiently-close nodes, and deterministically absent between more distant nodes. The second sort, in which the sequence of T_ns are fixed *temperature* parameters; the higher the temperature T_n, the closer the link function is to a constant probability $1/2$. The determinism of the first kind violates logit-boundedness. The second kind satisfies logit-boundedness when $T_n \in o((\ln n)n^{-1/2})$ and $t \leqslant \ln n$; in that case

$$\operatorname{logit} \frac{1}{1 + e^{(T_n^{-1}/2)(t - \ln n)}} = -\left(T_n^{-1}/2\right)(t - \ln n) \in o(\sqrt{n})$$

By extension, a CLS model is regular when $(M, w_{1:\infty})$ is. The proof of the following proposition, a straightforward consequence of ∞-homogeneity and injectivity of the link functions, is omitted.

Proposition 3. For regular CLS model

$$\operatorname{graph}_n(x_{1:n}) \overset{d}{=} \operatorname{graph}_n(y_{1:n}) \iff [x_{1:n}] = [y_{1:n}] \quad n = 1, 2, \ldots \tag{3.4}$$

Theorem 3 lets us identify graph distributions of the form $\operatorname{graph}_n(x_{1:n})$ with isometry classes $[x_{1:n}]$.

3.3. An example in the literature

Latent spaces of the form

$$(\mathbb{H}_2, w_n).$$

where $\mathbb{H}_2 = \{z = x + ij \in \mathbb{C} \mid x \in \mathbb{R},\ y \in [0, \infty)\}$ is the *Poincaré halfplane* with metric

$$dz = y^{-2}\, dx\, dy$$

were introduced [15] to model networks in nature with tree-like characteristics (e.g. the internet). With the first link function defined in (3.3), regularity is violated in multiple ways; the link functions are not logit-bounded as noted earlier, but also the link functions are neither smooth nor injective. With the second set of w_n's in (3.3), with temperature parameters $T_n \in o(\sqrt{n}^{-1})$, the CLS model is regular. Such CLS models have been shown to model salient for both large-scale and small-scale properties of various sorts of social networks [21], building on work of Krioukov et al. [15].

3.4. Estimation

Given a latent space model $(M, w_{1:\infty})$ and an n-node graph G, the likelihood $\mathcal{L}(x_{1:n}; G)$ of observing coordinates $x_{1:n} \in M^n$ is given as the product of edge probabilities:[2]

$$\mathcal{L}(x_{1:n}; G) \equiv \prod_{q=1}^{n} \prod_{p=1}^{q-1} w_n(dist\, x_p x_q)^{G_{pq}} \big(1 - w_n(dist\, x_p x_q)\big)^{1-G_{pq}}$$

A maximum likelihood (ML) *embedding* of an n-node graph G into M is

$$\hat{x}_{1:n} = \operatorname*{argmax}_{x_{1:n} \in M^n} \mathcal{L}(x_{1:n}; G) \tag{3.5}$$

Taking logs and dividing by the number of summands, we obtain the *normalized log-likelihood* $\ell(x_{1:n}; G)$ of observing coordinates $x_{1:n} \in M^m$ by:

$$\begin{aligned}\ell(x_{1:n}; G) = \frac{1}{n(n-1)} \Bigg(& \sum_{(p,q) \in G} \log \big(w_n\big(dist(x_p, x_q)\big)\big) \\ & + \sum_{(p,q) \notin G} \log \big(1 - w_n\big(dist(x_p, x_q)\big)\big) \Bigg)\end{aligned} \tag{3.6}$$

As usual, when there is no ambiguity about the graph G providing the data, we will suppress that as an argument, writing $\ell(x_{1:n})$.

Taking expectations with respect to the actual graph distribution of a random graph G having n nodes, we define the expected normalized log-likelihood by

$$\overline{\ell}(x_{1:n}) = \mathbb{E}_G \big[\ell(x_{1:n}; G)\big]. \tag{3.7}$$

As we review in Section 5.2, well-known results from information theory show that $-\overline{\ell}(x_{1:n})$ can always be decomposed into the sum of two non-negative terms, $-\overline{\ell}(x_{1:n}) = H + D(x_{1:n})$. For now, fix some true coordinates $x^*_{1:*}$ determining a CLS model. Here the first term, the "source entropy rate" H, captures the inherent stochasticity of the data source. The second term, the "divergence rate" $D(x_{1:n})$, measures the distance, or rather the normalized Kullback-Leibler divergence, between the "true" distribution $\text{graph}_n(x^*_{1:n})$, for some choice of true coordinates $x^*_{1:*}$, and $\text{graph}_n(x_{1:n})$. Among other properties, $D(x_{1:n})$ controls the power of any hypothesis test to distinguish $\text{graph}_n(x_{1:n})$ from the true distribution. This divergence is minimized by $x_{1:n} = x^*_{1:n}$; when the model is well-specified, $D(x^*_{1:n}) = 0$.

We are now in a position to state our main results.

[2] In this expression, every dyad appears twice, once as (p, q) and again as (q, p), but, since G is undirected, contributing the same factor to the likelihood each time. This is thus the square of another possible likelihood which counted each dyad only once. This will make no difference to the analysis, apart from needing to track a factor of 2 through our results.

Theorem 4. Suppose that the CLS model is regular. Then

$$\Gamma_n \equiv \sup_{x_{1:n}} |\ell(x_{1:n}) - \overline{\ell}(x_{1:n})| \overset{P}{\to} 0$$

where

Corollary 5. Suppose that the CLS model is regular, and $G \sim \mathrm{graph}_n(x^*_{1:n})$. Then

$$D(\hat{x}_{1:n}) - D\big(x^*_{1:n}\big) \leq 2\Gamma_n \overset{P}{\to} 0$$

4. Enrichments of the generative model

We briefly discuss a couple of natural enrichments on our generative model, reflecting structure and properties often observed in the real-world setting. These enrichments take the form of natural geometric structure on the latent space and suggest future areas of research.

4.1. Node covariates

Real-world networks often come equipped not just with the relational data of edges between nodes but, often, covariates on the nodes themselves. For example, social networks are comprised not just of the data of friendship relations between users but also demographic information for each user, such as age, height, and income. Assuming the covariates are sufficient to distinguish between the nodes, these covariates in effect embed the nodes in Euclidean space and therefore are modelled by the extra data of an embedding $M \hookrightarrow \mathbb{R}^e$ of the latent n-dimensional space M into Euclidean space $\mathbb{R}^e$ of some dimension $e \gg n$. If the embedding is assumed to pull back the Riemannian metric from $\mathbb{R}^e$ to M, then the complete geometry of an *unknown latent space* M can even be inferred from sample covariates by methods of non-parametric manifold learning [4]; the idea is that the sample covariates can be used to estimate the *Laplace-Beltrami operator* of the unknown latent space M, from which the complete geometry of M can be inferred by an application of Connes' Distance Formula [4]. Even if the embedding is more realistically assumed to satisfy some milder H older or Lipschitz constraints, then both some information about the geometry of M as well as the embedding of the nodes in M ought to be inferable from *partial* observations of node covariates and edge relationships. A consistent estimator for both the latent space geometry and all of the node covariates from partial observations of node covariates and edge relationships, adapting techniques developed in this paper, would find innumerable applications in social network analysis.

4.2. Volume

It is often desirable to control the occurrences of cliques of various orders in a generative model. Riemannian metrics, infinitesimal distances, govern the formation of edges, 2-cliques. The occurrence of higher order cliques follows from the properties of infinitesimal notions of compatible higher order distance, such as area forms in surfaces and more general volume forms. For example, the area form on the Poincaré Halfplane, a uniformly negatively curved space, governs the exponential rate at which the area of disks grows as a function of radius; this exponential growth severely constraints the growth of higher order cliques in a network sampled from a CLS model having the Poincaré Halfplane as a latent space [15]. Extending the detailed statistical analysis of local and global properties of such hyperbolic networks [12, 14, 15] for other latent spaces with different volume forms would find immediate applications in CLS model selection (cf. Smith, Asta and Calder [21] Hoff, Raftery and Handcock [11]).

5. Proofs

This section furnishes proofs of main results about networks. We can sketch the general approach as follows. We show that the expected log-likelihood achives its maximum precisely at the true coordinates up to isometry (Lemma 6). We then show that (in large graphs) the log-likelihood $\ell(x_{1:n})$ is, with arbitrarily high probability, arbitrarily close to its expectation value for each $x_{1:n}$ (Lemmas 9 and 10). We then extend that to a uniform convergence in probability, over all of M^n (Theorem 4). To do so, we need to bound the richness (*pseudo-dimension* [3, §11], a continuous generalization of VC dimension) of the family of log-likelihood functions (Theorem 8), which involves the complexity of the latent space's geometry, specifically of its isometry group $isom(M)$. Having done this, we have shown that the MLE also has close to the maximum expected log-likelihood. We emphasize this because the expected log-likelihood has a natural information-theoretic interpretation in terms of divergence from the truth (Eq. (5.3) below).

5.1. Notation

Before we dive into details, we first introduce some additional notation for our proofs. We will use G for both a (random or deterministic) graph and its adjacency matrix.

We fix the latent space as $(M, w_{1:\infty})$. For brevity, define

$$\lambda_n(x_p, x_q) \equiv \operatorname{logit} w_n\big(dist(x_p, x_q)\big). \tag{5.1}$$

As usual with binary observations, we can rewrite (3.6) so that the sum is taken over all pairs of distinct (p, q) and then replace each summand by $\log(1 - w_n(dist(x_p, x_q))) + G_{pq}\lambda_n(x_p, x_q)$. This brings out that the only data-dependent (and hence random) part of ℓ is linear in the entries of the adjacency matrix, and in the logit transform of the link-probability function. As usual, when there is no ambiguity about the graph G providing the data, we will suppress that as an argument, writing $\ell(x_{1:n})$. We write the class of log-likelihood functions as $\mathcal{L}_n$.

5.2. Information theory

Recall the definition of expected normalized log-likelihood from Eq. (3.7):

Taking expectations with respect to the actual graph distribution of a random graph G having n nodes, we define the expected normalized log-likelihood (the *cross-entropy*; Cover and Thomas 8, ch. 2) by

$$\overline{\ell}(x_{1:n}) = \mathbb{E}_G\big[\ell(x_{1:n}; G)\big], \tag{5.2}$$

where the expectation is taken with respect to the random graph G (and not the random graph G conditioned on some random equantity μ making the edges independent). For notational convenience, set

$$\pi_{pq}(a) = Pr(G_{pq} = a \mid x_p, x_q)$$
$$\pi^*_{pq}(a) = Pr\big(G_{pq} = a \mid x^*_p, x^*_q\big)$$

(so that $\pi_{pq}(1) = w_n(x_p, x_q)$ and $\pi^*_{pq}(1) = w_n(x^*_p, x^*_q)$). Then

$$\overline{\ell}(x_{1:n}) = \frac{1}{n(n-1)} \sum_{1\le p<q\le n} \sum_{a\in\{0,1\}} \pi^*_{pq}(a)\log \pi_{pq}(a).$$

In information theory [8, ch. 2], this quantity is known as the (normalized) **cross-entropy**, and we know that

$$-\sum_{a\in\{0,1\}} \pi^*_{pq}(a)\log\pi_{pq}(a) = H\big[\pi^*_{pq}\big] + D\big(\pi^*_{pq}\|\pi_{pq}\big),$$

as the left side is the cross-entropy of the distribution π_{pq} with respect to the distribution π^*_{pq} and the right side is the sum of ordinary entropy H with the Kullback-Leibler divergence D. Since both entropy and KL divergence are additive over independent random variables [8, ch. 2] like G_{pq}, we have,[3] defining $H[\pi^*]$ and $D(\pi^*\|\pi)$ in the obvious ways,

$$-\overline{\ell}(x_{1:n}) = H\big[\pi^*\big] + D\big(\pi^*\|\pi\big) \tag{5.3}$$

Unsurprisingly, $\overline{\ell}$ achieves a maximum at the (isometry class of) the true coordinates.[4]

Lemma 6. For ∞-homogeneous M and $G \sim \text{graph}_n(x^*_{1:n})$,

$$\big[x^*_{1:n}\big] = \underset{x_{1:n}\in M^n}{\operatorname{argmax}}\, \overline{\ell}_{norm}(x_{1:n}).$$

[3] The decomposition of expected log-likelihood into a entropy term which only involves the true distribution of the data, plus a KL divergence, goes back to at least Kullback [16].

[4] The statement and proof of the following lemma presume that the model is well-specified. If the model is mis-specified, then $\inf_{x_{1:n}} D(\pi^*\|\pi)$ is still well-defined, and still defines the value of the supremum for $\overline{\ell}$. The pseudo-true parameter value would be one which actually attained the infimum of the divergence [24]. This, in turn, would be the projection of π^* on to the manifold of distributions generated by the model [2]. All later invocations of Lemma 6 could be replaced by the assumption merely that this pseudo-truth is well-defined.

Proof. Letting H and D respectively denote entropy and KL divergence as in (5.3), $D(\pi^* \| \pi) \geq 0$, with equality if and only if $\pi^* = \pi$. Therefore we have that the divergence-minimizing π must be the distribution over graphs generated by some $x_{1:n} \in [x^*_{1:n}]$, and conversely that any parameter vector in that isometry class will minimize the divergence. The lemma follows from (5.3). □

5.3. *Geometric complexity of continuous spaces*

For various adjacency matrices G^1, G^2, etc., let us abbreviate $\ell(x_{1:n}; G^i)$ as $\ell^i(x_{1:n})$ (following Anthony and Bartlett 3, p. 91). Let us pick r different adjacency matrices $G^1, \ldots, G^r$, and set $\psi(x_{1:n}) = (\ell^1(x_{1:n}), \ldots, \ell^r(x_{1:n}))$. We will be concerned with the geometry of the level sets of ψ, i.e., the sets defined by $\psi^{-1}(c)$ for $c \in \mathbb{R}^r$. We say that a function $\psi : M^n \to \mathbb{R}^r$ has **has fibers with uniform bound B on the number of path-components** if $\psi^{-1}(x)$ has at most B path-components, equivalence classes of points where two points are equivalent if there is a path in $\psi^{-1}(x)$ connecting them, for each $x \in \mathbb{R}^r$.

Proposition 7. Suppose that all functions in $\mathcal{L}_n$ are jointly continuous in their d parameters almost everywhere, and that $\mathcal{L}_n$ has fibers with uniform bound B on the number of path-components. Then the growth function of $\mathcal{L}_n$, i.e., the maximum number of ways that $m \geq d$ data points $G^1, \ldots G^m$ could be dichotomized by thresholded functions from $\mathcal{L}_n$, is at most

$$\Pi(m) \leq B\left(\frac{em}{d}\right)^d \tag{5.4}$$

Thus the pseudo-dimension of $\mathcal{L}_n$ is at most $2\log_2 B + 2d\log_2 2/\ln 2$.

Proof. The inequality (5.4) is a simplification of Theorem 7.6 of Anthony and Bartlett [3, p. 91], which allows for sets to be defined by k-term Boolean combinations of thresholded functions from $\mathcal{L}_n$. (That is, the quoted bound is that of the theorem with $k = 1$.) Moreover, while Theorem 7.6 of Anthony and Bartlett [3] assumes that all functions in $\mathcal{L}_n$ are C^d, the proof (*op. cit.*, sec. 7.4) only requires continuity in the simplified setting $k = 1$.

For any class of sets with VC dimension $v < \infty$, the growth function is polynomial in m, $\Pi(m) \leq (em/v)^v$ [3, Theorem 3.7, p. 40], and, conversely, if $\Pi(m) < 2^m$ for any m, then the class of sets has VC dimension at most m. Since Eq. (5.4) shows that $\Pi(m)$ grows only polynomially in m, the VC dimension must be finite. Comparing the $O((m/d)^d)$ rate of Eq. (5.4) to the $O((m/v)^v)$ generic VC rate suggests $v = O(d)$, but it is desirable, for later purposes, to find a more exact result.

To do so, we find the least m where Eq. (5.4) is strictly below 2^m, and take the logarithm:

$$B\left(\frac{em}{d}\right)^d < 2^m \tag{5.5}$$

$$\log_2 B + d \log_2 \frac{e}{d} + d \log_2 m < m \tag{5.6}$$

Now, one can show that $\log_2 m \leq \frac{m}{2d} + \log_2 \frac{2d}{e \ln 2}$ [3, p. 91], so that

$$\log_2 B + d \log_2 \frac{e}{d} + d \log_2 m \leq \log_2 B + d \log_2 \frac{e}{d} + \frac{m}{2} + d \log_2 \frac{2d}{e \ln 2} \tag{5.7}$$

and it will be sufficient for the right-hand side to be $< m$. This in turn is implied by

$$2 \log_2 B + 2d \log_2 \frac{2}{\ln 2} < m \tag{5.8}$$

so this is an upper bound on the VC dimension of the subgraphs of $\mathcal{L}_n$, and so on the pseudo-dimension of $\mathcal{L}_n$. □

Next we bound the complexity of log-likelihoods for certain latent spaces.

Theorem 8. If $(M, w_{1:\infty})$ is regular, the pseudo-dimension of $\mathcal{L}_n$ is at most

$$2 \log_2 B_M + 2n \dim M \log_2 2/\ln 2, \tag{5.9}$$

where B_M is the number of path-components of the space of isometries on M. $isom(M)$.

Proof. By the fact that $(M, w_{1:\infty})$ is smooth, $\mathcal{L}_n$ is C^∞ in all its $n \dim M$ continuous parameters, so in applying Proposition 7, we may set $d = n \dim M$. Define $\phi(x_{1:n}; G)$ to be the function $M^n \to \mathbb{R}^{n^2}$ sending a tuple $x_{1:n}$ to the vector whose (pq)th coordinate, for $1 \leq p, q \leq n$, is $dist(x_p, x_q)$. Define $T : \mathbb{R}^{n^2} \to \mathbb{R}^n$ by the rule

$$T(y_1, y_2, \ldots, y_{n^2}) = \frac{1}{n(n-1)} \sum_{1 \leq p \leq q \leq n} y_{pq}.$$

Note each $\ell_{norm}(-; G) \in \mathcal{L}_n$ satisfies $\ell_{norm} = T\phi(-; G)$. The preimage $T^{-1}(c)$ of a point under T, a linear transformation, is either empty or a (connected and convex) linear subspace of $\mathbb{R}^{n^2}$. The function $\phi(-; G)$ has bounded connected components with bound B_M because $\phi(x_{1:n}) = \phi(y_{1:n})$ if and only if $[x_{1:n}] = [y_{1:n}]$ by ∞-homogeneity. Each $x_{1:n} \in M^n$ has a neighborhood U (e.g. a product of normal convex neighborhoods of $x_1, x_2, \ldots, x_n$ in M) such that $\phi(U; G)$ is convex in $\mathbb{R}^{n^2}$. It is then straightforward to show that every path in $\phi(M^n; G)$ starting from a point $\phi(x_{1:n}; G)$ lifts under $\phi(-; G)$ to a path in M^n from $x_{1:n}$. Thus $CC(\phi(-; G)^{-1}(d)) = CC((\phi(-; G)^{-1}(T^{-1}(c)))$ for each $d \in T^{-1}(c)$. Thus

$$\begin{aligned} CC\big(\phi(-; G)^{-1}\big(T^{-1}(c)\big)\big) &\leq CC\big(\phi(-; G)^{-1}(d)\big) \quad d \in T^{-1}(c) \\ &\leq B_M, \end{aligned}$$

where $CC(X)$ denotes the number of path-components of a space X. Thus each $\ell(-; G) \in \mathcal{L}_n$ has bounded connected components with bound B_M. The hypotheses of Proposition 7 being satisfied, (5.9) follows from Proposition 7. □

5.4. Pointwise convergence of log-likelihoods

Lemma 9. Suppose that all of the edges in G are conditionally independent given some random variable μ. Then for any $\epsilon > 0$,

$$Pr\Big(|\ell(x_{1:n}) - \overline{\ell}(x_{1:n})| > \epsilon\Big) \leq 2e^{(-2\frac{n^2(n-1)^2\epsilon^2}{\sum_{p=1}^{n}\sum_{q>p} 2\lambda_n^2(x_p,x_q)})} \tag{5.10}$$

In particular, this holds when $G \sim \text{graph}_n(x^*_{1:n})$ or $G \sim \text{graph}_n(f)$.

Proof. Changing a single G_{pq}, but leaving the rest the same, changes $\ell(x_{1:n}; G)$ by $\frac{1}{n(n-1)}\lambda_n(x_p, x_q)$. The G_{pq}, for $p < q$, are all independent given μ. We may thus appeal to the bounded difference (McDiarmid) inequality [7, Theorem 6.2, p. 171]: if f is a function of independent random variables, and changing the k^{th} variable changes ℓ by at most c_k, then

$$Pr\Big(|f - \mathbb{E}[f]| > \epsilon\Big) \leq 2e^{(-\frac{\epsilon^2}{2\nu})} \tag{5.11}$$

where $\nu = \frac{1}{4}\sum c_k^2$. In the present case, $c_{pq} = \lambda_n(x_p, x_q)$. Thus,

$$\nu = \frac{1}{4}\sum_{p=1}^{n}\sum_{q>p} n^{-2}(n-1)^{-2}\lambda_n^2(x_p, x_q) = \frac{1}{4n^2(n-1)^2}\sum_{p=1}^{n}\sum_{q>p}\lambda_n^2(x_p, x_q) \tag{5.12}$$

and so $Pr(|\ell(x_{1:n}) - \overline{\ell}(x_{1:n})| > \epsilon \mid \mu)$ is bounded from above by

$$2e^{(-\frac{2n^2(n-1)^2\epsilon^2}{\sum_{p=1}^{n}\sum_{q>p}\lambda_n^2(x_p,x_q)})} \tag{5.13}$$

Since the unconditional deviation probability

$$Pr\Big(|\ell(x_{1:n}) - \overline{\ell}(x_{1:n})| > \epsilon\Big)$$

is just the expected value of the conditional probability, which has the same upper bound regardless of μ, the result follows (cf. Shalizi and Kontorovich 19, Theorem 2).

Finally, note that all edges in $\text{graph}_n(x^*_{1:n})$ are unconditionally independent, while those in $\text{graph}_n(f)$ are conditionally independent given $X_{1:n}$, which plays the role of μ. □

This lemma appears to give exponential concentration at an $O(n^4)$ rate, but of course the denominator of the rate itself contains $\binom{n}{2} = O(n^2)$ terms, so the over-all rate is only $O(n^2)$. Of course, there must be some control over the elements in the denominator.

Lemma 10. If $-v_n \leq \lambda_n(x_p, x_q) \leq v_n$, then for any $x_{1:n}$ and $\epsilon > 0$,

$$Pr\Big(|\ell(x_{1:n}) - \overline{\ell}(x_{1:n})| > \epsilon\Big) \leq 2e^{(-2\frac{n(n-1)\epsilon^2}{v_n^2})} \tag{5.14}$$

Proof. By assumption, $\lambda_n^2(x_p, x_q) \leq v_n^2$. Thus $\sum_{p=1}^{n}\sum_{q>p}\lambda_n^2(x_p, x_q) \leq \binom{n}{2}v_n^2$, and the result follows from Lemma 9. □

5.5. Uniform convergence of log-likelihoods

Lemmas 9 and 10 show that, with high probability, $\ell(x_{1:n})$ is close to its expectation value $\overline{\ell}(x_{1:n})$ for any given parameter vector $x_{1:n}$. However, we need to show that the MLE $\hat{X}_{1:n}$ has an *expected* log-likelihood close to the optimal value. We shall do this by showing that, *uniformly* over M^n, $\ell(x_{1:n})$ is close to $\overline{\ell}(x_{1:n})$ with high probability. That is, we will show that

$$\sup_{x_{1:n}} \left|\ell(x_{1:n}) - \overline{\ell}(x_{1:n})\right| \overset{P}{\to} 0 \tag{5.15}$$

This is a stronger conclusion than even that of Lemma 10: since M is a continuous space, even if each parameter vector has a likelihood which is exponentially close to its expected value, there are an uncountable infinity of parameter vectors. Thus, for all we know right now, an uncountable infinity of them might be simultaneously showing large deviations, and continue to do so no matter how much data we have. We will thus need to show that likelihood at different parameter values are *not* allowed to fluctuate independently, but rather are mutually constraining, and so eventually force uniform convergence.

If there were only a finite number of allowed parameter vectors, we could combine Lemma 10 with a union bound to deduce (5.15). With an infinite space, we need to bound the covering number of $\mathcal{L}_n$. To recall,[5] the L_1 covering number of a class F of functions at scale ϵ and m points, $\mathcal{N}_1(\epsilon, F, m)$, is the cardinality of the smallest set of functions $f_j \in F$ which will guarantee that, for any choice of points $a_1, \dots a_m$, $\sup_{1,\dots,a_m} \frac{1}{m}\sum_{i=1}^{m} |f(a_i) f_j(a_i)| \leq \epsilon$ for some f_j (this definition can be straightforwardly shown to be equivalent to that of Anthony and Bartlett [3]). Typically, as in Anthony and Bartlett [3, Theorem 17.1, p. 241], a uniform concentration inequality takes the form of

$$Pr\Big(\sup_{f\in F} \left|f - \mathbb{E}[f]\right| \geq \epsilon\Big) \leq c_0 c_1 \mathcal{N}_1(\epsilon c_2, F, c_3 m) e^{(-c_4 \epsilon^2 r(m))} \tag{5.16}$$

where the individual deviation inequality is

$$Pr\big(|f - \mathbb{E}[f]| \geq \epsilon\big) \leq c_0 e^{-(-\epsilon^2 r(m))}. \tag{5.17}$$

In turn, Anthony and Bartlett [3, Theorem 18.4, p. 251] shows that the L_1 covering number $\mathcal{N}_1(\epsilon, F, m)$ of a class F of functions with finite pseudo-dimension v at scale ϵ and m observations is bounded:

$$\mathcal{N}_1(\epsilon, F, m) \leq e(v+1)\left(\frac{2e}{\epsilon}\right)^v. \tag{5.18}$$

In our setting, we have $m = 1$. (That is, we observe *one* high-dimensional sample; notice that the bound is independent of m so this hardly matters.)

It thus remains to bound the pseudo-dimension of $\mathcal{L}_n$. This involves a rather technical geometric argument, ultimately revolving on the group structure of the isometries of $(M, dist)$. This may be summed up in the existence of a constant B_M, which is 2 for any Euclidean space, and (as it happens) also 2 for $\mathbb{H}_2$. This matter was handled in Section 5.3.

[5] See, e.g., Anthony and Bartlett [3] or Vidyasagar [23].

5.6. Proof of Theorem 4

By assumption, there exists a sequence $\nu_1, \nu_2, \ldots$ of non-negative reals such that $|\lambda_n(x_p, x_q)| \leq v_n$ for each n and p, q with $\nu_n \in o(\sqrt{n})$.

Presume for the moment that we know the L_1 covering number of $\mathcal{L}_n$ is at most $\mathcal{N}_1(\mathcal{L}_n, \epsilon, 1)$. Then

$$Pr\left(\sup_{x_{1:n}} |\ell(x_{1:n}) - \overline{\ell}(x_{1:n})| \geq \epsilon\right) \leq 4\mathcal{N}_1(\mathcal{L}_n, \epsilon/16, 2)e^{(-\frac{\epsilon^2 n(n-1)}{8v_n^2})} \tag{5.19}$$

The proof is entirely parallel to that of Theorem 17.1 in Anthony and Bartlett [3, p. 241], except for using Lemma 10 in place of Hoeffding's inequality, and so omitted.

Now, by Proposition 2 $B_M = 2$ and therefore by Theorem 8, the pseudo-dimension of $\mathcal{L}_n$ is at most $2\log_2 B_M + 2n \dim M \log_2 2/\ln 2$. The L_1 covering number of $\mathcal{L}_n$ is thus exponentially bounded in $O(n \log 1/\epsilon)$, specifically [3, Theorem 18.4, p. 251]: $\mathcal{N}_1(\mathcal{L}_n, \epsilon, 2)$ is bounded above by

$$e(1 + 2\log_2 B_M + 2n \dim M \log_2 2/\ln 2)\left(\frac{2e}{\epsilon}\right)^{2\log_2 B_M + 2n \dim M \log_2 2/\ln 2} \tag{5.20}$$

(5.20) grows exponentially in $O(n \log 1/\epsilon)$, while the rightmost factor in the upper bound of (5.19) shrinks exponentially in $O(\epsilon^2 n^2/v_n^2)$ and hence $O(n\epsilon^2)$ by our regularity assumption. For fixed ϵ, then, the uniform deviation probability over all of $\mathcal{L}_n$ in (5.19) is therefore exponentially small, hence we have convergence in probability to zero. □

Remark 1. In applying the theorems from Anthony and Bartlett [3], remember that we have only one sample ($m = 1$), which is however of growing ($O(n^2)$) dimensions, with a more-slowly growing ($O(n)$) number of parameters.

Remark 2. From the proof of the theorem, we see that if v_n^2 grows slowly enough, the sum of the deviation probabilities tends to a finite limit. Convergence in probability would then be converted to almost-sure convergence by means of the Borel-Cantelli lemma, *if* the graphs at different n can all be placed into a common probability space. Doing so however raises some subtle issues we prefer not to address here (cf. [20]).

5.7. Proof of Corollary 5

We adapt a very standard pattern of argument used to prove oracle inequalities in learning theory. This begins with Lemma 6, that $\overline{\ell}(x^*_{1:n}) \geq \overline{\ell}(\hat{x}_{1:n})$. This implies that $|\overline{\ell}(\hat{x}_{1:n}) - \overline{\ell}(x^*_{1:n})| = \overline{\ell}(x^*_{1:n}) - \overline{\ell}(\hat{x}_{1:n})$. Now add and subtract log-likelihoods:

$$\begin{aligned} 0 \leq \overline{\ell}\big(x^*_{1:n}\big) - \overline{\ell}(\hat{x}_{1:n}) &= \overline{\ell}\big(x^*_{1:n}\big) - \ell(\hat{x}_{1:n}) + \ell(\hat{x}_{1:n}) - \overline{\ell}(\hat{x}_{1:n}) &(5.21)\\ &\leq \overline{\ell}\big(x^*_{1:n}\big) - \ell\big(x^*_{1:n}\big) + \ell(\hat{x}_{1:n}) - \overline{\ell}(\hat{x}_{1:n}) &(5.22)\end{aligned}$$

$$\leq |\overline{\ell}\big(x^*_{1:n}\big) - \ell\big(x^*_{1:n}\big)| + |\ell(\hat{x}_{1:n}) - \overline{\ell}(\hat{x}_{1:n})| \tag{5.23}$$

$$\leq 2 \sup_{x_{1:n}} |\ell(x_{1:n}) - \overline{\ell}(x_{1:n})| \overset{P}{\to} 0 \tag{5.24}$$

where in Eq. (5.22) we use the trivial fact that since $\hat{x}_{1:n}$ maximizes the likelihood, $\ell(x^*_{1:n}) \leq \ell(\hat{x}_{1:n})$, and the last line invokes Theorem 4. □

6. Conclusion

We have formulated and proven a notion of convergence for non-parametric likelihood estimators of graphs generated from continuous latent space models, under some mild assumptions on the generative models. Traditional convergence results for statistical estimators are a kind of ergodicity, or long-term mixing, for multiple, independent samples. The size of a single sample network here plays the role of the number of samples in traditional formulations of consistency. These main results hold even when our generative models are mis-specified, i.e. when we fix a latent space but the generating graph distributions are not defined in terms of the space, under some additional assumptions [Appendix A]. Continuous latent space models turn out to provide the necessary ergodicity through conditional independence. A consequent notion of consistency, which we save for future work, requires some formalization of what we mean by convergence of estimates, i.e. sequences of coordinates, of varying sizes. And a proof of such a consistency result will likely require some adaptation of standard technical tools for concluding convergence of extremal estimators from convergence in random objective functions (e.g. [22]).

Appendix A: Mis-specified models

Our consistency results extend from specified to certain mis-specified models. We still assume the existence of a latent space $(M, w_{1:\infty})$ as before, but assume that sample graphs are sampled not by a distribution of the form $\mathrm{graph}_n(x_{1:n})$ but in fact by some arbitrary distribution of graphs having n nodes. The only assumption we make about such random graphs G in this section, as before, is that there exists some random variable μ such that the edges of G are conditionally independent given μ. For the case where G is drawn from a CLS model, μ can be taken to be the random latent coordinates of the nodes of G. We call a sequence $G_1, G_2, \ldots$ of random graphs *almost-specified* if there exists $x^*_{1:\infty} \in M^\infty$ such that, for all sufficiently large n, $\overline{\ell}(x_{1:n})$ achieves a maximum uniquely exactly for $x_{1:n} \in [x^*_{1:n}]$. For such an almost-specified model, $x^*_{1:\infty}$ plays the role of the true coordinates and the assumption of being almost specified plays the role of Lemma 6 (e.g. in all proofs); we call such $x^*_{1:\infty}$ the *pseudo-coordinates* of the almost-specified model. Consequently, we can restate our main results at the following level of generality.

Theorem 11. For an almost specified model with pseudo-coordinates $x^*_{1:*}$ and a compact, regular latent space $(M, w_{1:\infty})$,

$$\sup_{x_{1:n}} |\ell(x_{1:n}) - \overline{\ell}(x_{1:n})| \xrightarrow{P} 0$$

In particular, a reordering of the nodes is irrelevant for these results. This is essentially due to logit-boundedness and the fact that we are only claiming convergence in log-likelihoods, not convergence in coordinates. In our setup, we are *given* an ordering, in the sense that we are given for each n an observed graph G_n of the *first* n nodes (determined by the first n true coordinates $x^*_{1:n}$). The fact that the MLE's converge *in log-likelihoods*, in fact at a certain rate irrespective of node ordering, follows from the logit-boundedness condition.

If we want to refine our results to obtain a convergence of the MLEs to the true coordinates *in some sort of metric space of coordinates*, the particular choice of ordering of the nodes might cause some problems. In order to go from log-likelihood convergence to convergence of MLEs in standard settings, one requires some assumption that the MLE is a *well-separated maximum*. To go from our convergence results to convergence of MLEs in our setting is trickier because the MLEs themselves live in different metric spaces as $n \to \infty$, so one needs to refine the notion of well-separated maximum to take this into account. This refined notion of well-separatedness is sensitive to node ordering, and further assumptions need to be made on the generative model to get the desired refined notion of convergence.

Acknowledgments

We are grateful for valuable discussions with Carl Bergstrom, Elizabeth Casman, David Choi, Aaron Clauset, Steve Fienberg, Christopher Genovese, Aryeh Kontorovich, Dmitri Krioukov, Cris Moore, Alessandro Rinaldo, Mitch Small, Neil Spencer, Andrew Thomas, Larry Wasserman, and Chris Wiggins, and for feedback from seminar audiences at CMU, UCLA, UW-Seattle, and SFI.

Funding

Our work was supported by NSF grant DMS-1418124; DA also received support from NSF Graduate Research Fellowship under grant DGE-1252522, and CRS from NSF grant DMS-1207759.

References

[1] Albert, R., DasGupta, B. and Mobasheri, N. (2014). Topological implications of negative curvature for biological and social networks. *Physical Review E* **89** 032811. https://doi.org/10.1103/PhysRevE.89.032811

[2] Amari, S.-i., Barndorff-Nielsen, O. E., Kass, R. E., Lauritzen, S. L. and Rao, C. R. (1987). *Differential Geometry in Statistical Inference. Institute of Mathematical Statistics Lecture Notes-Monographs Series* **10**. Institute of Mathematical Statistics, Hayward, California. MR0932246

[3] Anthony, M. and Bartlett, P. L. (1999). *Neural Network Learning: Theoretical Foundations*. Cambridge University Press, Cambridge, England. MR1741038

[4] Asta, D. M. (2022). Non-parametric manifold learning. Submitted.

[5] Asta, D. and Shalizi, C. R. (2015). Geometric network comparison. In *31st Conference on Uncertainty in Artificial Intelligence [UAI 2015]* (M. Meila and T. Heskes, eds.) 102–110. AUAI Press, Corvallis, Oregon.

[6] Athreya, A., Fishkind, D. E., Tang, M., Priebe, C. E., Park, Y., Vogelstein, J. T., Levin, K., Lyzinski, V. and Qin, Y. (2017). Statistical inference on random dot product graphs: a survey. *The Journal of Machine Learning Research* **18** 8393–8484. MR3827114

[7] Boucheron, S., Lugosi, G. and Massart, P. (2013). *Concentration Inequalities: A Nonasymptotic Theory of Independence*. Oxford University Press, Oxford. MR3185193

[8] Cover, T. M. and Thomas, J. A. (2006). *Elements of Information Theory*, 2nd ed. John Wiley, New York. MR2239987

[9] Cvetkovski, A. and Crovella, M. (2009). Hyperbolic embedding and routing for dynamic graphs. In *IEEE INFOCOM 2009* 1647–1655. IEEE.

[10] Dani, V., Diaz, J., Hayes, T. P. and Moore, C. (2021). Improved reconstruction of random geometric graphs. 2107.14323. MR4473353

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

[12] Kennedy, W. S., Narayan, O. and Saniee, I. (2013). On the hyperbolicity of large-scale networks. arXiv:1307.0031.

[13] Kovács, B. and Palla, G. (2021). The inherent community structure of hyperbolic networks. *Scientific Reports* **11** 16050.

[14] Krioukov, D., Papadopoulos, F., Vahdat, A. and Boguñá, M. (2009). Curvature and temperature of complex networks. *Physical Review E* **80** 035101. `https://doi.org/10.1103/PhysRevE.80.035101`

[15] Krioukov, D., Papadopoulos, F., Kitsak, M., Vahdat, A. and Boguñá, M. (2010). Hyperbolic geometry of complex networks. *Physical Review E* **82** 036106. `https://doi.org/10.1103/PhysRevE.82.036106`. MR2787998

[16] Kullback, S. (1968). *Information Theory and Statistics*, 2nd ed. Dover Books, New York. MR1461541

[17] McCormick, T. H. and Zheng, T. (2015). Latent surface models for networks using aggregated relational data. *Journal of the American Statistical Association* **110** 1684–1695. MR3449064

[18] Penrose, M. (2003). *Random Geometric Graphs*. Oxford University Press,

Oxford. MR1986198

[19] Shalizi, C. R. and Kontorovich, A. L. (2013). Predictive PAC learning and process decompositions. In *Advances in Neural Information Processing Systems 26 [NIPS 2013]* (C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani and K. Q. Weinberger, eds.) 1619–1627. MIT Press, Cambridge, Massachusetts.

[20] Shalizi, C. R. and Rinaldo, A. (2013). Consistency under sampling of exponential random graph models. *Annals of Statistics* **41** 508–535. https://doi.org/10.1214/12-AOS1044. MR3099112

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

[22] van der Vaart, A. W. (1998). *Asymptotic Statistics*. Cambridge University Press, Cambridge, England. MR1652247

[23] Vidyasagar, M. (2003). *Learning and Generalization: With Applications to Neural Networks*, 2nd ed. Springer-Verlag, Berlin. MR1938842

[24] White, H. (1994). *Estimation, Inference and Specification Analysis*. Cambridge University Press, Cambridge, England. MR1292251

[25] Young, S. J. and Scheinerman, E. R. (2007). Random dot product graph models for social networks. In *International Workshop on Algorithms and Models for the Web-Graph* 138–149. Springer. MR2504912

[26] Yu, X. and Rodríguez, A. (2021). Spatial voting models in circular spaces: A case study of the US House of Representatives. *The Annals of Applied Statistics* **15** 1897–1922. MR4355081