CONSISTENCY OF MAXIMUM LIKELIHOOD FOR CONTINUOUS-SPACE NETWORK MODELS

BY COSMA ROHILLA SHALIZI AND DENA ASTA

Network analysis needs tools to infer distributions over graphs of arbitrary size from a single graph. Assuming the distribution is generated by a continuous latent space model which obeys certain natural symmetry and smoothness properties, we establish three levels of consistency for non-parametric maximum likelihood inference as the number of nodes grows: (i) the estimated locations of all nodes converge in probability on their true locations; (ii) the distribution over locations in the latent space converges on the true distribution; and (iii) the distribution over graphs of arbitrary size converges.

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 simulating 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 (Hoff, Raftery and Handcock, 2002), recent work has suggested that in many applications it would be better to take the space to non-Euclidean, specifically negatively curved or hyperbolic (Krioukov et al., 2010; Asta and Shalizi, 2015).

We provide three levels of consistency for all continuous latent space models which obey certain natural symmetry and smoothness assumptions, including standard Euclidean and hyperbolic models. The first level is the
consistency of 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. We prove that nonparametric maximum likelihood embedding is consistent, i.e., that the sum of distances between real and estimated locations tends to zero in probability [Theorem 7]. The second level is that if the node locations are drawn IIDly from a fixed but unknown distribution on the latent space, that distribution can be recovered by applying suitable nonparametric density estimators (Asta, 2015) to the locations obtained by embedding [Corollary 8]. The third level is that the distribution over new graphs produce by simulating from a model estimated in this way converges in probability on the true generating distribution [Corollary 9].

Section §2 reviews background on continuous latent space models of networks. Section §3 states our main results, along with certain technical assumptions. All proofs, and a number of subsidiary results and lemmas, are deferred to Section §4.

2. Background. In typical 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 where 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. General networks lack a natural, exogenous sense of distance for such decay of dependence.

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.

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 (Albert, DasGupta and Mobasher, 2014; Kennedy, Narayan and Saniee,
2013; Krioukov et al., 2010). Mathematically, this is because many real net-
works can be naturally embedded into such spaces. More substantively, many
real-world networks show highly skewed degree distributions, very short path
lengths, a division into a core and peripheries where short paths between pe-
ipheral 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 (Krioukov et al.,
2010).

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.

We therefore investigate three levels of estimation. The first and essential
one is that of estimating the nodes’ locations when we have observed the
graph. This is the problem of embedding the graph into the latent space.
We show that the maximum likelihood embedding converges in prob-
ability on the true locations, up to a global isometry. Here the crucial quantity
controlling the rate of convergence turns out to be the richness of the latent
space’s group of isometries, i.e., the complexity of its geometry.

The second level is that of estimating a distribution over the latent space
from the cloud of embedded points. Here the key issue is whether the space
has enough symmetry to allow for the definition of a generalization of the
Fourier transform, and with it of convolution and so of generalized kernel
density estimates, which are minimax-optimal. (We defer details to Asta
(2015).) We show that when the generalized KDE is well-defined, applying
it to the maximum-likelihood embedding consistently estimates the distri-
bution from which nodes are drawn.

The third level is that of distributions over graphs, i.e., does the distribu-
tion over new graphs implied by our estimated node density converge on the
true data-generating process? Here again the answer is in the affirmative.
This in turn opens the way to model-based bootstraps for assessing uncer-
tainty, and to two-sample tests for checking whether networks obtained un-
der different conditions could have come from the same probabilistic source
(Asta and Shalizi, 2015).
To the best of our knowledge, there are no results in the existing literature on the consistency of embedding for CLS models, much less for estimation of nodes densities or the distribution over graphs.

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 consistently estimates the true locations of the nodes (Theorem 7). This will allow us to consistently estimate the density of nodes (Corollary 8) and the distribution over graphs (Corollary 9) from a single observed graph.

All proofs are postponed to §4.

3.1. Setting and Conventions. 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 $\text{isom}(M)$ 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 $w_n(\text{dist}(x_i, x_j))$. By a latent space $(M, w_{1:\infty})$, we will mean the combination of $M$ and a sequence $w_{1:\infty}$ of link functions. When the latent space is understood, we write $\text{graph}_n(x_{1:n})$ for the distribution of a random graph on $n$ vertices located at $x_{1:n}$.

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

\begin{equation}
\text{graph}_n(x_{1:n}) \overset{d}{=} \text{graph}_n(\phi(x_{1:n}))
\end{equation}

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,

\begin{equation}
\text{dist}([x_{1:n}], [y_{1:n}]) = \inf_{\phi \in \text{isom}(M)} \sum_{i=1}^{n} \text{dist}(x_i, \phi(y_i)).
\end{equation}

We cannot hope to identify $x_{1:n}$ by observing the graph it leads to, but we can hope to estimate $[x_{1:n}]$. 
If we introduce a density $f$ on $M$, we can make the node locations themselves random, with $X_i$ drawn independently \(^1\) from $f$. Edges between the first $n$ nodes are then conditionally independent given $X_{1:n}$. A random graph $G$ having $n$ nodes is thus defined by the hierarchical procedure

\begin{align*}
X_i & \sim_{iid} f \\
G|X_{1:n} & \sim \text{graph}_n(X_{1:n})
\end{align*}

Which is to say, nodes are connected as before, but now node locations are random. Edges are conditionally independent given these random locations. The corresponding continuous latent space model (CLS model) will be written $(M, f, w_{1:∞})$. When the latent space is understood, we write $\text{graph}_n(f)$ for the distribution of $G$ above. Again,

\begin{equation}
\text{graph}_n(f) \overset{d}{=} \text{graph}_n(f \circ \phi),
\end{equation}

so the node density $f$ can only be identified up to a global isometry. Accordingly, we will use $[f]$ to indicate the equivalence class of densities $g$ on $M$, such that $g = f \circ \phi$ for some $\phi \in \text{isom}(M)$. We cannot hope to estimate a density $f$ by observing the graphs it generates, but we can hope to estimate $[f]$.

We will carefully distinguish, below, between results which hold with fixed, or conditioned-on, node locations, and those which presume IID node locations.

**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.

**Definition 1.** A metric space $M$ is rigid when

1. For each $n$ and $x_{1:n}, y_{1:n} \in M^n$, $[x_{1:n}] = [y_{1:n}]$ whenever

\begin{equation}
\text{dist}(x_p, x_q) = \text{dist}(y_p, y_q), \quad p, q \in 1 : n
\end{equation}

2. $\text{isom}(M)$ has a finite number $B_M$ of connected components.

\(^1\)A closely related alternative is to distribute the $X_i$ according to a point process, rather than drawing from a density. Drawing from a density leads to dense graph sequences; point processes can lead to sparse graphs sequences.
By extension, the CLS model \((M, f, w_{1:\infty})\) is also rigid when \(M\) is rigid.

**Proposition 2.** The metric spaces \(\mathbb{R}^d\) and \(\mathbb{H}_2\) are rigid, with

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

The particular manner in which we define \(w_{1:\infty}\) determines how the random graph sequence changes.

**Definition 3.** A sequence of link functions \(w_{1:\infty}\) is *logit-bounded* with bounds \(v_n\) when for all \(n\) and \(x, y \in M\),

\[
-v_n \leq \log \frac{w_n(\text{dist}(x, y))}{1 - w_n(\text{dist}(x, y))} = \log \frac{1}{w_n(\text{dist}(x, y))} \leq v_n.
\]

and \(n/v_n \to \infty\) as \(n \to \infty\).

Two remarks are in order. First, demanding that \(v_n = o(n)\), rather than just finite, is done with an eye towards the needs of the proofs in §4. Second, if the range of \(w_n\) is the same proper sub-interval of \((0, 1)\) for all \(n\), then \(w_{1:\infty}\) is logit-bounded, but the converse is not true.

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

**Definition 4.** A latent space \((M, w_{1:\infty})\) is *regular* when:

1. \(M\) is a complete rigid Riemannian manifold; and
2. The function \(w_n\) is injective and smooth for each \(n\); and
3. The sequence \(w_{1:\infty}\) is logit-bounded.

By extension, CLS models are regular when \((M, w_{1:\infty})\) are.

**Theorem 5.** For regular CLS model

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

Theorem 5 lets us identify graph distributions of the form \(\text{graph}_n(x_{1:n})\) with isometry classes \([x_{1:n}]\). We can therefore define distances between such distributions by the distance between the isometry classes \((2)\).

**Theorem 6.** For regular CLS models \((M, f, w_{1:\infty})\) and \((M, g, w_{1:\infty})\),

\[
\text{graph}_n(f) \overset{d}{=} \text{graph}_n(g) \iff [f] = [g] \quad n = 1, 2, \ldots
\]
The theorem allows us to identify sequences of distributions of random graphs determined by regular CLS models \((M, f, w_{1:\infty})\) with isometry classes \([f]\) of densities \(f\) on \(M\). We can define distances between such sequences by

\[
\text{dist}([f],[g]) = \inf_{\phi \in \text{isom}(M)} \| f - g \circ \phi \|_2.
\]

3.3. An example in the literature. Latent spaces of the form \((\mathbb{H}_2, w_n)\).

where \(\mathbb{H}_2 = \{ z \in \mathbb{C} \mid \text{Im}(z) > 0 \}\) is the Poincaré halfplane with metric

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

were introduced (Krioukov et al., 2010) to model networks in nature with tree-like characteristics (e.g., the internet). The \(w_n\)’s considered can take one of the two following forms:

\[
w_n(t) = \begin{cases} 
1 & t \leq \ln n \\
0 & t \gt \ln n 
\end{cases} \hspace{1cm} w_n(t) = \frac{1}{1 + e^{\lambda(t-\ln n)}}
\]

With this latter choice of link functions, the latent space is regular. A variant of CLS models are then defined so that the node densities vary according to the number of nodes. Our notion of CLS models and hence our main results do not apply to such generative models, although we hope to address such density-varying models in future work.

3.4. Estimators. Given a latent space model \((M, w_{1:\infty})\) and an \(n\)-node graph \(G\), we define the log-likelihood of the location \(x_{1:n} \in M^n\) by:

\[
\ell(x_{1:n}; G) = \sum_{(p,q) \in G} \log w_n(\text{dist}(x_p, x_q)) + \sum_{(p,q) \notin G} \log(1 - w_n(\text{dist}(x_p, x_q)))
\]

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

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

\[
\hat{x}_{1:n} = \arg \max_{x_{1:n} \in M^n} \ell(x_{1:n}; G)
\]

The point \(\hat{x}_i \in M\) is, naturally, the estimated location in \(M\) of vertex \(i\). The MLE is consistent in the following sense:
**Theorem 7.** Fix a $x^*_{1:\infty} \in M^\infty$ and $w_{1:\infty}$, where the latent space $(M, w_{1:\infty})$ is regular. Suppose that for each $n$, $G_n \sim \text{graph}_n(x_{1:n})$. Then

$$\left[\hat{X}_{1:n}\right] \xrightarrow{P} [x^*_{1:n}].$$

The theorem, as stated, presumes the model is well-specified; the maximum likelihood embedding can be shown to converge for a broader class of random graphs where the model is mis-specified; details are given in the Appendix.

We can combine a consistent density estimator with an ML embedding to obtain a consistent node density estimator. An example of a consistent density estimator is a kernel density estimator on $\mathbb{R}^d$, or suitable generalizations (Asta, 2015). We say that a density estimator is continuous if the density it outputs is continuous on the data points it is given as input (with respect to $L^2$-distances between estimated densities and distances in $M^n$, respectively.)

**Corollary 8.** Suppose that for each $n$, the graph $G_n$ is drawn from a regular CLS model, $G_n \sim \text{graph}_n(f)$, and that $\hat{f}$ is a consistent, continuous density estimator on $M$. Then

$$\left[\hat{f}_{\hat{X}_{1:n}}\right] \xrightarrow{P} [f].$$

Here $\left[\hat{f}_{\hat{X}_{1:n}}\right] \xrightarrow{P} [f]$ means that for each $\epsilon > 0$,

$$\Pr\left(\text{dist}\left(\left[\hat{f}_{\hat{X}_{1:n}}\right], [f]\right) \geq \epsilon\right) \xrightarrow{n \to \infty} 0.$$

**Corollary 9.** Under the assumptions of Corollary 8,

$$\text{graph}_m(\hat{f}_{\hat{X}_{1:n}}) \xrightarrow{P} \text{graph}_m(f) \quad m = 1, 2, \ldots$$

**4. Proofs.** This section furnishes proofs of main results about networks, such as observations about the generative model and its assumptions (Proposition 2 and Theorem 6) and the consistency of the graph embeddings (Theorem 7). The consistency of the node density estimator (Corollary 8) and of the estimated graph distribution (Corollary 9) follow easily.

Since everything turns on the consistency of maximum likelihood embedding, and the argument is somewhat intricate, it is worth sketching the approach. We show that the expected log-likelihood achieves its maximum
precisely at the true coordinates up to isometry (Lemma 13). 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 14 and 15). We then extend that to a uniform convergence in probability, over all of \( M^n \) (Theorem 16). To do so, we need to bound the richness (pseudo-dimension) of the family of log-likelihood functions (Theorem 12), which involves the complexity of the latent space’s geometry, specifically of its isometry group \( \text{isom}(M) \). Having done this, we have shown that the MLE also has close to the maximum expected log-likelihood.

4.1. Notation. Before we dive into details, we first fix some additional notation for our proofs. We will use \( G \) for both a (random or deterministic) graph and its adjacency matrix. We assume that for each random graph \( G \) considered in this paper, there exists some random quantity \( \mu \) such that the edges of \( G \) are conditionally independent given \( \mu \). For the case where \( G \) is drawn from a CLS model, \( \mu \) can be taken to be the random latent coordinates of the nodes of \( G \).

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

\[
\lambda_n(x_p, x_q) = \log it w_n(\text{dist } (x_p, x_q)).
\]

As usual with binary observations, we can write the log-likelihood function for coordinates given a graph \( G \) of \( n \) nodes as

\[
\ell(x_{1:n}; G) = \frac{1}{n(n-1)} \sum_{p=1}^{n} \sum_{q \neq p} \log (1 - w_n(\text{dist } (x_p, x_q))) + G_{pq} \lambda_n(x_p, x_q).
\]

This brings out that the only data-dependent (and hence random) part of \( \ell \) is linear in the entries of the adjacency matrix, and in the logit transform of the link-probability function.

We write the class of log-likelihood functions as \( \mathcal{L}_n \). Taking expectations with respect to the actual graph distribution of a random graph \( G \) having \( n \) nodes, we define the expected log-likelihood (the \emph{cross-entropy}; Cover and Thomas 2006, ch. 2) by

\[
\mathcal{L}(x_{1:n}) = \mathbb{E}[\ell(x_{1:n}; G)].
\]

Setting \( \pi_{pq}(a) = Pr(G_{pq} = a \mid X_p, X_q) \) and \( \pi_{pq}^*(a) = Pr(G_{pq} = a \mid \mu) \),

\[
\mathcal{L}(x_{1:n}) = \frac{1}{n(n-1)} \sum_{p=1}^{n} \sum_{q \neq p} \sum_{a \in \{0,1\}} \pi_{pq}^*(a) \log \pi_{pq}(a).
\]
From information theory (Cover and Thomas, 2006, ch. 2), observe
\[- \sum_{a \in \{0,1\}} \pi^*_pq(a) \log \pi^*_pq(a) = H[\pi^*_pq] + D(\pi^*_pq \parallel \pi_{pq}),\]
as the left side is the cross-entropy of the distribution \(\pi_{pq}\) with respect to the distribution \(\pi^*_pq\) 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 (Cover and Thomas, 2006, ch. 2) like \(G_{pq}\), we have\(^2\), defining \(H[\pi^*]\) and \(D(\pi^* \parallel \pi)\) in the obvious ways,
\[(20) \quad -\overline{\ell}(x_{1:n}) = H[\pi^*] + D(\pi^* \parallel \pi)\]

4.2. Generative-Model Properties. We first show that Euclidean space and the hyperboloid are rigid.

**Proof of Proposition 2.** We consider first the case \(M = \mathbb{H}_2\). We wish to show that for all \(n\), if two ordered sets of \(n\) points \(x_{1:n}\) and \(y_{1:n}\) have the same inter-point distances, then there is a isometry \(\phi \in \text{isom}(\mathbb{H}_2)\) taking \(x_i\) to \(y_i\) for each \(i \in 1:n\). This is trivially true when \(n = 1\). It is also easily seen to be true when \(n = 2\), because there is always an isometry \(\phi_1\) taking \(x_1\) to the origin (in the Poincaré disk representation of \(\mathbb{H}_2\)), and a second isometry \(\phi_2\) doing the same to \(y_1\). Since \(\phi_1(x_2)\) and \(\phi_2(y_2)\) are points equidistant from the origin, there is a third isometry \(\phi_3\) taking \(\phi_1(x_2)\) to \(\phi_2(y_2)\) while leaving the origin fixed. Hence the composition \(\phi_2^{-1} \circ \phi_3 \circ \phi_1\) is an isometry taking each \(x_i\) to \(y_i\). Almost the same argument works when \(n = 3\): \(\phi_1\) and \(\phi_2\) again take \(x_1\) and \(y_1\) (respectively) to the origin, so that \(\phi_1(x_2)\) and \(\phi_1(x_3)\) sit on the same circles around the origin as \(\phi_2(y_2)\) and \(\phi_2(y_3)\) (respectively). Then \(\phi_3\) can take \(\phi_1(x_i)\) to \(\phi_2(y_i)\) by some combination of rotation and reflection, leaving the origin fixed, and the again the composition \(\phi_2^{-1} \circ \phi_3 \circ \phi_1\) takes every \(x_i\) to \(y_i\).

Cases where \(n > 3\) are handled by induction. Suppose that the result holds for \(n - 1\); we now show that it also holds for \(n\). By the inductive hypothesis, there exists at least one \(\phi\) where \(y_i = \phi(x_i)\) for \(i \in 1:(n - 1)\). Because \(\phi\) is a global isometry, \(\phi(x_n)\) is well-defined. We will be done if \(\phi(x_n)\) can be shown to equal \(y_n\). Now, an arbitrary point \(y \in \mathbb{H}_2\) can be uniquely identified by its distance from three points \(y_1, y_2, y_3\) in general position. This is because circles in the Poincaré disk are also Euclidean circles, and three circles with

---

\(^2\)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 (1968).
non-collinear centers will have a unique intersection, if they intersect at all. So \( y_n \) is uniquely determined by \( \text{dist}(y_1, y_n), \text{dist}(y_2, y_n) \) and \( \text{dist}(y_3, y_n) \). But since \( \phi \) is an isometry,

\[
(21) \quad \text{dist}(x_j, x_n) = \text{dist}(\phi(x_j), \phi(x_n)) = \text{dist}(y_j, \phi(x_n)) \quad , \quad j \in 1:3
\]

where the second equality holds by the inductive hypothesis. Hence \( \phi(x_n) \) must equal \( y_n \). But this means that \( \phi \) takes all \( n \) of the \( x_i \) to the corresponding \( y_i \). Hence \( \mathbb{H}_2 \) satisfies condition (1) in the definition of a rigid space (Definition 1).

According to Katok (1992, Theorem 1.4.1, p. 10), the group of isometries of \( \mathbb{H}_2 \) consists of the semi-direct product of the continuous group of orientation-preserving isometries, \( PSL(2, \mathbb{R}) \) (generated from translations, rotations and dilations), with the two-element group consisting of the identity and the orientation-reversing reflection around (in the upper-half-plane representation) the imaginary axis. Moreover, the \( PSL(2, \mathbb{R}) \) subgroup has index 2 in the full group of isometries. Thus, the number of connected components is 2. Hence (2) is satisfied in the definition of rigidity.

As for the case \( M = \mathbb{R}^d \) with the Euclidean metric, the proof is entirely parallel. Cases up through \( n = d + 1 \) can be handled due to the degrees of freedom of the isometry group, while when \( n > d + 1 \) rely on induction, since an arbitrary point is uniquely determined by its distance to \( d + 1 \) points in general position. This takes care of condition (1) in the definition of rigidity. As for condition (2), \( isom(\mathbb{R}^d) \) can be divided into the orientation-preserving and orientation-reversing isometries. The former form a sub-group. Every orientation-reversing isometry is the composition of a single reflection with an orientation-preserving isometry, so the orientation-preserving isometries have index 2 in \( isom(\mathbb{R}^d) \). Thus, the number of connected components is 2, independent of the dimension \( d \).

We next show that regular CLS models are identifiable.

**Proof of Theorem 6.** For densities \( h \) and \( g \) on \( M \), let the \( X_i \) be IID with distribution \( h \), and \( X'_i \) be IID with distribution \( g \), and let \( \Delta_{h,n} \) and \( \Delta_{g,n} \) respectively represent the random matrices

\[
\text{dist}(X_i, X_j)_{i,j} , \quad \text{dist}(X'_i, X'_j)_{i,j}.
\]

First, we assume that \( h = g \circ \phi \) for some isometry \( \phi \in isom(M) \), and show that the resulting graph distributions are equal.

Consider an isometry \( \phi : M \cong M \) and assume \( \hat{h} = g \circ \phi \). Then,

\[
(22) \quad \Delta_{h,n} \overset{d}{=} \Delta_{g \circ \phi, n} \overset{d}{=} \Delta_{g, n} ;
\]
the first equality holding by assumption, the second because \( \phi \) is an isometry. Since edge probabilities are function of distances alone, the distribution over graphs is a function of the distribution of distances, it follows that for all \( n \),

\[
\text{graph}_n(g) \overset{d}{=} \text{graph}_n(h).
\]

Now suppose \( h \) and \( g \) generate the same distribution over random graphs for all sizes \( n \). By the fact that \( w_n \) is injective for each \( n \), it follows that \( \Delta_{h,n} = \Delta_{g,n} \). Then there is a version of \( \Delta_{h,n} \) and \( \Delta_{g,n} \) on a common probability space such that \( \Delta_{h,n} = \Delta_{g,n} \) a.s. by Theorem 3.30 of Kallenberg (2002). Thus the metric subspaces \( \{X_1, X_2 \ldots X_n\}, \{X'_1, X'_2 \ldots X'_n\} \) of \( M \) are isometric a.s. Hence there exists an isometry \( \phi : M \cong M \) such that \( X_i = \phi(X'_i) \) a.s. for \( i = 1, 2, \ldots n \) by Condition 1 of Definition 1. It follows \( h = g \circ \phi \).

4.3. Geometric Complexity of Continuous Spaces. We consider the normalized log-likelihood functions of latent space models. 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 1999, 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 \( \psi \), i.e., the sets defined by \( \psi^{-1}(c) \) for \( c \in \mathbb{R}^r \).

**Definition 10.** A function \( \psi : M^n \to \mathbb{R}^r \) has bounded connected components (with bound \( B \)) when, for any \( r \leq n \), and for almost all \( c \in \mathbb{R}^r \),

\[
CC(\psi^{-1}(c)) \leq B
\]

where \( CC(S) \) counts the number of connected components of the set \( S \).

**Proposition 11.** Suppose that all functions in \( \mathcal{L}_n \) are \( C^d \) in their \( d \) parameters almost everywhere, and that \( \mathcal{L}_n \) has bounded connected components with bound \( B \). Then the growth function of \( \mathcal{L}_n \), i.e., the maximum number of ways that \( m \) data points \( G^1, \ldots G^m \) could be dichotomized by thresholded functions from \( \mathcal{L}_n \), is at most

\[
\Pi(m) \leq B \left( \frac{em}{d} \right)^d
\]

Thus the pseudo-dimension of \( \mathcal{L}_n \) is at most \( 2 \log_2 B + 2d \log_2 e \).
Proof. The inequality (24) is a simplification of Theorem 7.6 of Anthony and Bartlett (1999, p. 91), which allows for sets to be defined by \( k \)-term Boolean combinations of thresholded functions from \( L_n \). (That is, the quoted bound is that of the theorem with \( k = 1 \).) Moreover, while Theorem 7.6 of Anthony and Bartlett (1999, p. 91) assumes that all functions in \( L_n \) are \( C^d \), the proof only requires this to be true almost everywhere. The last statement follows from Lemma 4.6(3) of Vidyasagar (2003, p. 134), which asserts that

\[
\frac{m}{\gamma + \alpha \log_2 (\beta m)} \Rightarrow m < 2 \gamma + \alpha \log_2 (\alpha \beta)
\]

We apply this to bound on the growth function from Proposition 11: if \( m \leq \nu \), the VC dimension, then \( 2^m \leq \Pi(m) \). Taking the \( \log_2 \) of both sides,

\[
m \leq \log_2 B + d \log_2 (em) - d \log_2 d
\]

Setting \( \alpha = d \), \( \beta = e \), and \( \gamma = \log_2 B - d \log_2 d \), the inequality gives us

\[
m < 2 \log_2 B - 2d \log_2 d + 2d \log_2 de = 2 \log_2 B + 2d \log_2 e
\]

which is the upper bound on the VC dimension of the subgraphs of \( L_n \), and so on the pseudo-dimension of \( L_n \). \( \square \)

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

Theorem 12. Suppose that \((M, w_{1:\infty})\) is regular. The pseudo-dimension of \( L_n \) is at most

\[
2 \log_2 B_M + 2n \dim M \log_2 e,
\]

where \( B_M \) is the number of connected components of \( isom(M) \).

Proof. By the fact that \((M, w_{1:\infty})\) is smooth, \( L_n \) is \( C^\infty \) in all its \( n \dim M \) continuous parameters almost everywhere, so in applying Proposition 11, we may set \( d = n \dim M \). By condition 1 of Definition 1, two parameter vectors \( x_{1:n} \) and \( y_{1:n} \) give the same distribution over graphs if and only if \([x_{1:n}] = [x_{1:n}]\). Thus, by appropriate choices of \( G^1, \ldots, G^r \), we can ensure that \( \psi(x_{1:n}) = \psi(y_{1:n}) \) only if \( x_{1:n} \) and \( y_{1:n} \) are isometric, since each distance shows up as a distinct term in the log-likelihood sum. The number of connected components of \( \psi^{-1}(c) \) is thus (at most) \( B_M < \infty \). The hypotheses of Proposition 11 being satisfied, (28) follows from Proposition 11. \( \square \)

3Note Vidyasagar writes “lg” for “log_2”.
Unsurprisingly, $\ell$ achieves a maximum at the (isometry class of) the true coordinates$^4$.

**Lemma 13.** For $M$ satisfying (1) in Definition 1 and $G \sim \text{graph}_n(x^*_n)$,

$$[x^*_{1:n}] = \underset{x_{1:n} \in M^n}{\text{argmax}} \ell(x_{1:n}).$$

**Proof.** Letting $H$ and $D$ respectively denote entropy and KL divergence, $D(\pi^*\parallel \pi) \geq 0$, with equality if and only if $\pi^* = \pi$. Therefore we have that the divergence-minimizing $\pi$ 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 (20). □

4.4. **Pointwise Convergence of Log-Likelihoods.**

**Lemma 14.** Suppose that all of the edges in $G$ are conditionally independent given some random variable $\mu$. Then for any $\epsilon > 0$,

$$Pr \left( |\ell(x_{1:n}) - \bar{\ell}(x_{1:n})| > \epsilon \right) \leq 2 \exp \left\{ -\frac{n^2(n-1)^2 \epsilon^2}{2 \sum_{p=1}^{n} \sum_{q>p} \lambda_n^2(x_p, x_q)} \right\}$$

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 $\frac{2}{n(n-1)} \lambda_n(x_p, x_q)$. (The factor of 2 arises because $G_{pq} = G_{qp}$ in an undirected graph.) The $G_{pq}$ are all independent given $\mu$. We may thus appeal to the bounded difference (McDiarmid) inequality (Boucheron, Lugosi and Massart, 2013, Theorem 6.2, p. 171): if $\ell$ is a function of independent random variables, and changing the $k^{th}$ variable changes $\ell$ by at most $c_k$, then

$$Pr \left( |\ell - \mathbb{E}[\ell]| > \epsilon \right) \leq 2 \exp \left\{ -\frac{\epsilon^2}{2c^2 k} \right\}$$

$^4$The statement and proof of the following lemma presume that the model is well-specified. If the model is mis-specified, then $\inf_{x_{1:n}} D(\pi^*\parallel \pi)$ is still well-defined, and still defines the value of the supremum for $\bar{\ell}$. The pseudo-true parameter value would be one which actually attained the infimum of the divergence (White, 1994). This, in turn, would be the projection of $\pi^*$ on the manifold of distributions generated by the model (Amari et al., 1987). All later invocations of Lemma 13 could be replaced by the assumption merely that this pseudo-truth is well-defined.
where $\nu = \frac{1}{4} \sum c_k^2$. In the present case, we have $c_{pq} = \frac{2}{n(n-1)} \lambda_n(x_p, x_q)$. Thus,

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

and so

$$\Pr \left( |\ell(x_{1:n}) - \overline{\ell}(x_{1:n})| > \epsilon \mid \mu \right) \leq 2 \exp \left\{ \frac{-n^2(n-1)^2 \epsilon^2}{2 \sum_{p=1}^{n} \sum_{q>p} \lambda_n^2(x_p, x_q)} \right\}$$

Since the unconditional deviation probability $\Pr \left( |\ell(x_{1:n}) - \overline{\ell}(x_{1:n})| > \epsilon \right)$ is just the expected value of the conditional probability, which has the same upper bound regardless of $\mu$, the result follows (cf. Shalizi and Kontorovich 2013, Theorem 2).

Finally, note that all edges in graph$_n(x^*_1:n)$ are unconditionally independent, while those in graph$_n(f)$ are conditionally independent given $X_1:n$, which plays the role of $\mu$.

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 15.** Assume $w_n$ is logit-bounded. Then for any $x_{1:n}$ and $\epsilon > 0$,

$$\Pr \left( |\ell(x_{1:n}) - \overline{\ell}(x_{1:n})| > \epsilon \right) \leq 2 \exp \left\{ \frac{-n(n-1)^2 \epsilon^2}{v_n^2} \right\}$$

**Proof.** By assumption, $\lambda_n^2(x_p, x_q) \leq v_n^2$. Thus $\sum_{p=1}^{n} \sum_{q>p}^{n} \lambda_n^2(x_p, x_q) \leq \binom{n}{2} v_n^2$, and the result follows from Lemma 14.

4.5. **Uniform Convergence of Log-Likelihoods.** Lemmas 14 and 15 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

$$\sup_{x_{1:n}} \frac{1}{n} |\ell(x_{1:n}) - \overline{\ell}(x_{1:n})| \overset{P}{\to} 0$$
This is a stronger conclusion than even that of Lemma 15: 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 15 with a union bound to deduce (34). 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 function class at scale $\epsilon$ 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 that, for any choice of points $a_1, \ldots, a_m$, $\sum_{i=1}^{m} |f(a_i) - f_j(a_i)| \leq \epsilon$ for some $f_j$. Typically, as in Anthony and Bartlett (1999, Theorem 17.1, p. 241), a uniform concentration inequality takes the form of

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

where the individual deviation inequality is

\[
Pr (|f - \mathbb{E}[f]| \geq \epsilon) \leq c_0 \exp \{ -\epsilon^2 r(m) \}
\]

In turn, Anthony and Bartlett (1999, Theorem 18.4, p. 251) shows that the $L_1$ covering number of a class of functions $F$ with finite pseudo-dimension $v$ at scale $\epsilon$ and $m$ observations is bounded:

\[
\mathcal{N}_1(\epsilon, F, m) \leq e(v + 1) \left( \frac{2e}{\epsilon} \right)^v
\]

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 $\mathbb{H}_2$. We bracket this matter in §4.3.

---

\(^5\)See, e.g., Anthony and Bartlett (1999) or Vidyasagar (2003).
Theorem 16. Suppose that the CLS model is regular and logit-bounded. Then
\[
\sup_{x_{1:n}} |\ell(x_{1:n}) - \overline{\ell}(x_{1:n})| \xrightarrow{P} 0
\]
with the probability of a size \(\epsilon\) deviation shrinking exponentially in \(\epsilon^2 n^2/v_n^2\).

Proof. Presume for the moment that we know the \(L_1\) covering number of \(L_n\) is at most \(\mathcal{N}_1(L_n, \epsilon)\). Then
\[
\Pr \left( \sup_{x_{1:n}} |\ell(x_{1:n}) - \overline{\ell}(x_{1:n})| \geq \epsilon \right) \leq 4\mathcal{N}_1(L_n, \epsilon/16) \exp \left\{ -\frac{n(n-1)\epsilon^2}{16v_n^2} \right\}
\]
The proof is entirely parallel to that of Theorem 17.1 in Anthony and Bartlett (1999, p. 241), except for using Lemma 15 in place of Hoeffding’s inequality, and so omitted.

Now, by Theorem 12 and Proposition 2, the pseudo-dimension of \(L_n\) is at most \(2\log_2 B_M + 2n \dim M \log_2 e\). The \(L_1\) covering number of \(L_n\) is thus exponentially bounded in \(O(n \log 1/\epsilon)\), specifically (Anthony and Bartlett, 1999, Theorem 18.4, p. 251):
\[
\mathcal{N}_1(L_n, \epsilon) \leq e(1 + 2\log_2 B_M + 2n \dim M \log_2 e) \left( \frac{2\epsilon}{e} \right)^{2\log_2 B_M + 2n \dim M \log_2 e}
\]
(38) grows exponentially in \(O(n \log 1/\epsilon)\), while (33) shrinks exponentially in \(O(\epsilon^2 n^2/v_n^2)\). For fixed \(\epsilon\), then, the uniform deviation probability over all of \(L_n\) in (38) is therefore exponentially small in \(\epsilon^2 n^2/v_n^2\), hence we have convergence in probability to zero.

Remark 1: In applying the theorems from Anthony and Bartlett (1999), 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 the uniform deviation probabilities are exponentially small in \(n^2/v_n^2\). Thus 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. Shalizi and Rinaldo 2013).
4.6. The Main Convergence Results.

**Proof of Theorem 7.** For each $n$, define

$$f_n(x_{1:*}) = \sum_{i=1}^{n} 2^{-i} \ell(x_{1:i}), \quad f(x_{1:*}) = \sum_{i=1}^{\infty} 2^{-i} \ell(x_{1:i}).$$

For each $n$, $f_n : M^\infty \to \mathbb{R}$ is continuous because $\ell$ is continuous. The function $f : M^\infty \to \mathbb{R}$ is continuous because the defining series above converges uniformly. In an abuse of notation, let $[M^n]$ denote the quotient of $M^n$ by the natural action of the isometry group of $M$ for $n = 0, 1, 2, \ldots, \infty$. Note that $f_0, f_1, \ldots, f$ induce well-defined continuous functions $\bar{f}_1, \bar{f}_2, \ldots, \bar{f} : [M^n] \to \mathbb{R}$ for $n = 0, 1, 2, \ldots, \infty$.

First note that $\sup_{x_{1:*}} |f_n(x_{1:*}) - f(x_{1:*})| = \cdots$

$$= \sup_{x_{1:*}} \left| \sum_{i=1}^{n} 2^{-i} (\ell(x_{1:i}) - \ell(x_{1:i})) + \sum_{i=n+1}^{\infty} 2^{-i} \ell(x_{1:i}) \right|$$

$$\leq \sup_{x_{1:*}} \left| \sum_{i=1}^{n} 2^{-i} (\ell(x_{1:i}) - \ell(x_{1:i})) \right| + O(n^{-1})$$

because the $w_i$'s are logit-bounded. Each summand in the last line converges in probability to 0 by Theorem 16. It follows that

$$\sup_{x_{1:*}} |f_n([x_{1:*}]) - \bar{f}([x_{1:*}])| \xrightarrow{P} 0.$$

It suffices to show that $\bar{f}$ has a well-separated maximum. Then the desired convergence would follow (van der Vaart, 1998, Theorem 5.7, p. 45).

Let $B_n$ be the finite union of $n$ closed balls, each with radius

$$R = \max_{i,j} w_n^{-1} \left( \frac{w_n \left( \text{dist} \left( x_i^*, x_j^* \right) \right)}{\sum_{i',j'} \log w_n \text{dist} \left( x_i^*, x_j^* \right)} \right),$$

about each of the points $x_1^*, x_2^*, \ldots, x_n^*$. Then $B_n$ is compact because $M$ is complete and closed balls in complete Riemannian manifolds are compact. Therefore $[B_n^n]$ is compact for each $n$ because quotients and products of compact spaces are compact. Moreover, $[B_n^n]$ contains $[x_{1:n}]$. Therefore the restriction of $f_n$ to $[B_n^n]$ has $[x_{1:n}]$ as its well-separated maximum because $\bar{f}_n$ is continuous and has compact domain.
Consider \([x_{1:n}] \in [M^n] - [B^n]_n\). We can take \(x_1 = x_1^\ast\) because \(M\) is homogeneous and \(x_2 \notin B_n\) because \([x_{1:n}] \notin [B^n]_n\). Therefore

\[
\log w_n(dist (x_1^\ast, x_2^\ast)) - \log w_n(dist (x_1, x_2)) > \log w_n(dist (x_1^\ast, x_2^\ast)) - \log w_n(R) > \sum_{i,j} \log w_n dist (x_i^\ast, x_j^\ast)
\]

Therefore \(\ell(x_{1:n}) < \ell(x_{1:n}^\ast) - \epsilon_n\). Thus \(\bar{f}_n\) has a well-separated maximum on all of \([M^n]\) for each \(n\). It then follows that \(f\) has a well-separated maximum.

**Proof of Corollary 8.** Define \(X_{1:n}^\ast\) so that

\(X_1^\ast, X_2^\ast, \ldots, X_n^\ast \sim f\).

Since the density estimator is consistent, \(\hat{f}_{X_{1:n}^\ast} \xrightarrow{P} f\) and so

\[
[\hat{f}_{X_{1:n}^\ast}] = \hat{f} [X_{1:n}^\ast] \xrightarrow{P} [f].
\]

Also by Theorem 7,

\[
[\hat{X}_{1:n}] \xrightarrow{P} [X_{1:n}^\ast].
\]

Since the density estimator is continuous in its inputs, an application of the continuous mapping principle (Mann and Wald, 1943) yields the result.

**Proof of Corollary 9.** For every \(m\), graph\(_m\)(\(f\)) is continuous in \(f\). Combining Corollary 8 with another application of the continuous mapping principle thus proves the desired convergence.

5. Conclusion. Thus we have formulated and proven a notion of consistency for non-parametric likelihood estimators of graphs generated from continuous latent space models, under some mild assumptions on the generative models. Traditional consistency 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. Continuous latent space models turn out to provide the necessary ergodicity through conditional independence. Consequently, we obtain three levels of consistency: consistency
of MLE graph embeddings, consistency of non-parametric node density estimators, and consistency of non-parametric random graph estimators, all as the size of the graphs increases. 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].

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 \((M_{\infty}, w_{1:\infty})\) as before, but assume that sample graphs are sampled not by a distribution of the form \(\operatorname{graph}_{n}(x_{1:n})\) but in fact by some arbitrary distribution of graphs having \(n\) nodes. We call a sequence \(G_{1}, G_{2}, \ldots\) of random graphs almost-specified if there exists \(x_{1:n}^{*} \in M_{\infty}\) such that, for all sufficiently large \(n\), \(\ell(x_{1:n})\) achieves a maximum uniquely exactly for \(x_{1:n} \in [x_{1:n}^{*}]\). For such an almost-specified model, \(x_{1:n}^{*}\) plays the role of the true coordinates and the assumption of being almost specified plays the role of Lemma 13 (e.g. in all proofs); we call such \(x_{1:n}^{*}\) the pseudo-coordinates of the almost-specified model. Consequently, we can restate our main results at the following level of generality.

**Theorem 17.** For an almost specified model with pseudo-coordinates \(x_{1:n}^{*}\), and a compact, regular latent space \((M_{\infty}, w_{1:\infty})\),

\[
\left[ \hat{X}_{1:n} \right] \overset{\mathcal{P}}{\to} [x_{1:n}^{*}].
\]

ACKNOWLEDGEMENTS

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. 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 and UW-Seattle.

REFERENCES

Albert, R., DasGupta, B. and Mobasher, N. (2014). Topological Implications of Negative Curvature for Biological and Social Networks. *Physical Review E* **89** 032811.
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.

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

Asta, D. (2015). Kernel Density Estimation on Symmetric Spaces. In Geometric Science of Information (F. Nielsen and F. Barbaresco, eds.) 779–787. Springer, Berlin.

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.

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

Cover, T. M. and Thomas, J. A. (2006). Elements of Information Theory, Second ed. John Wiley, New York.

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.

Kallenberg, O. (2002). Foundations of Modern Probability, Second ed. Springer-Verlag, New York.

Katok, S. (1992). Fuchsian Groups. University of Chicago Press, Chicago.

Kennedy, W. S., Narayan, O. and Sanjee, I. (2013). On the Hyperbolicity of Large-Scale Networks. arxiv:1307.0031.

Krioukov, D., Papadopoulos, F., Kitsak, M., Vahdat, A. and Boguñá, M. (2010). Hyperbolic Geometry of Complex Networks. Physical Review E 82 036106.

Kullback, S. (1968). Information Theory and Statistics, 2nd ed. Dover Books, New York.

Mann, H. B. and Wald, A. (1943). On stochastic limit and order relationships. Annals of Mathematical Statistics 14 217–226.

Penrose, M. (2003). Random Geometric Graphs. Oxford University Press, Oxford.

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.

Shalizi, C. R. and Rinaldo, A. (2013). Consistency Under Sampling of Exponential Random Graph Models. Annals of Statistics 41 508–535.

van der Vaart, A. W. (1998). Asymptotic Statistics. Cambridge University Press, Cambridge, England.

Vidyasagar, M. (2003). Learning and Generalization: With Applications to Neural Networks, Second ed. Springer-Verlag, Berlin.

White, H. (1994). Estimation, Inference and Specification Analysis. Cambridge University Press, Cambridge, England.

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

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