Torus Graph Inference for Detection of Localized Activity

Elizabeth A. Beer, Carey E. Priebe, Edward R. Scheinerman

Abstract

Our goal is to detect localized regions of excessive activity in a network, distinguishing networks that contain such regions from networks whose activity is more homogeneous. We consider inference on random graphs from a latent position model whose latent position space is a torus, using edge density and maximum degree on as test statistics.

1 Torus Latent Position Models

By now, it goes without saying that random graph models are a popular and fruitful area of study, with applications to physics, biology, sociology, and other fields. We consider a random graph model whose edge probabilities depend on the latent positions of the graph's vertices; in our model, these latent positions lie on a torus.

1.1 Latent Position Models Latent position models, or latent space models, were introduced by Hoff, Raftery and Handcock [3]. In general, these random graph models propose a latent position (in some space of interest) for each vertex; the probability of the edge ij is a function of the distance between the latent positions of the vertices i and j.

To generate a random graph from a latent position model, we use a two-stage process: First, latent positions, $\ell_1, \ldots, \ell_n$, are drawn (i.i.d.) from a specified distribution on the latent position space. Given the latent positions, we generate a random graph $G(\ell_1, \ldots, \ell_n)$ by drawing $\binom{n}{2}$ Bernoulli random variables $Y_{12}, \ldots, Y_{n-1,n}$ for the edges. The edge variables are conditionally independent (conditioning on the positions $\ell_1, \ldots, \ell_n$), with Y_{ij} having success probability p_{ij}, some function of the distance between the latent positions ℓ_i and ℓ_j.

Formally, a *latent (fixed) position model* for graphs with n vertices is a sample space on $\mathcal{G}_n$;. The collection is defined by a metric space $(\mathcal{X}, d)$ (the space in which the vertices take their latent positions)[1] and a function $f : \mathbb{R} \to [0, 1]$ (used to convert distances between points into probabilities of vertex adjacency). Thus, we may call the model $LPM(\mathcal{X}, d, f)$. This model consists of the samples spaces $(\mathcal{G}_n, P_\ell)$, where $\ell : [n] \to \mathcal{X}$ assigns a latent position to each vertex. In a particular sample space $(\mathcal{G}_n, p_\ell) \in LPM(\mathcal{X}, d, f)$, the probability of a specific graph G is

$$P_\ell(G) = \prod_{u<v, u\sim v} f\left[d\left(\ell(u), \ell(v)\right)\right] \times \prod_{u<v, u\nsim v} \left[1 - f\left[d\left(\ell(u), \ell(v)\right)\right]\right]$$

A *latent random position model* (LRPM) for graphs with n vertices is also a collection of sample spaces on $\mathcal{G}_n$. Its construction is similar to that of a latent fixed position model; instead of the position function ℓ, however, the LRPM is indexed by μ, a probability measure on $\mathcal{X}$. The latent positions $\ell_1, \ldots, \ell_n$ are random variables drawn independently according to μ. (We may therefore think of ℓ, in this model, as a random function from $[n]$ to $\mathcal{X}$.) We write $LRPM(\mathcal{X}, d, f)$ for the model, and $(\mathcal{G}_n, P_\mu)$ for its sample spaces. The probability of a specific graph G in the sample space $(\mathcal{G}_n, P_\mu)$ takes just the same form as in a latent fixed position model:

$$P_\mu(G) = \prod_{u<v, u\sim v} f\left[d\left(\ell(u), \ell(v)\right)\right] \times \prod_{u<v, u\nsim v} \left[1 - f\left[d\left(\ell(u), \ell(v)\right)\right]\right].$$

The difference is that in this case the $\ell(i)$ are random variables.

Our torus model is a latent random position model whose latent space $\mathcal{X}$ is a k-torus, S^k, with d being Euclidean distance along the surface of the torus. A torus, like a sphere, is compact without having edges or corners; the k-torus has the additional advantage of being naturally representable as a rectangular region in $\mathbb{R}^k$ (variously named the "Asteroids," glued-square, or Poincaré polyhedron embedding). Notice that, although a square with opposite edges identified is *topologically* equivalent to any other torus, there is no isometry between the glued-square embedding and a donut-shaped embedding of S^k. Our model uses the glued-square embedding; any visualization involving donut-shaped embeddings must therefore be taken with a grain of salt. (For further explanation, see [2].)

[1] Note that Hoff, Raftery, and Handcock do not require $(\mathcal{X}, d)$ to be a metric space; indeed, one of their examples makes use of non-symmetric projections.

We embed S^k as the k-dimensional unit cube. The edge probability function is $f(d) = 1 - \frac{d}{2^{-k}} = 1 - 2^k d$. Since the maximum possible interpoint (torus) distance in the unit cube is 2^{-k}, this f allows edge probabilities to range from 0 (for vertices with latent positions as far apart as possible) to 1 (for vertices with the same latent position). The edge probability decreases linearly as the distance between the endpoints increases.

Using a torus as the latent space is one way of dealing with problems presented by more typical latent spaces.[3] A torus is both compact and edgeless. Furthermore, the glued-square representation is a preferred embedding for performing multidimensional scaling onto a torus (as noted by [2]), which may be useful for estimating latent positions if we observe only the resulting graph.

1.2 Uniform Models and Mixture Models To fully define our torus random graph model as a latent random position model, we must stipulate a distribution for the latent positions of the vertices. We consider latent positions distributed uniformly on S^k, as well as latent positions distributed as a mixture of two uniform distributions whose supports partition S^k. For the purposes of this paper, we consider mixture-model graphs of the form $\mathbb{T}_{n,w,p}$, where w is the side length (between 0 and 1) of a square region[2], and each latent position is drawn from this square region with probability p and from its complement with probability $1 - p$.

For example, to generate a random graph from the uniform-mixture 1-torus model, we draw $\ell_1, \ldots, \ell_n$ from a mixture of two uniform distributions on the 1-torus (embedded as the unit interval). With probability p, we draw ℓ_i from the interval $[0, w)$; with probability $1 - p$, we draw ℓ_i from the interval $[w, 1]$. Given ℓ_i, ℓ_j, the edge variable Y_{ij} is Bernoulli with success probability

$$p_{ij} = 1 - \frac{d(\ell_i, \ell_j)}{1/2},$$

where d is the 1-torus distance,

$$d(\ell_i, \ell_j) = \min(|\ell_i - \ell_j|, |(1 + \ell_i) - \ell_j|).$$

We generally assume w is small, so that vertices with latent positions inside the square are quite likely to be adjacent in the random graph (more so than in the larger remainder of the graph); thus, this model captures in a simple way the idea of a local subregion of excessive activity. In general, we could admit a mixture model with more components, each capturing a subgroup within the vertex set whose members are likely to communicate amongst themselves.

[2]Since the latent position space is a torus, and the positions are uniformly distributed, we may assume without loss of generality that this region has one corner at the origin.

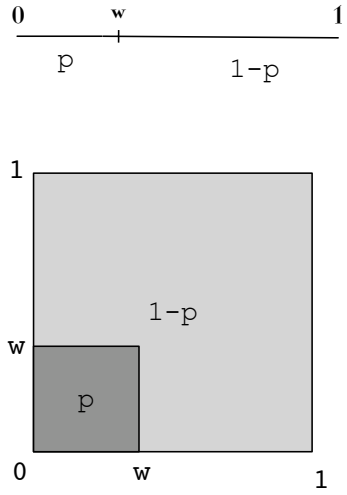


Figure 1: This figure illustrates the role of p and w in the mixture model: w defines the size of a square region, and each point is drawn from that square with probability p and from the rest of the torus with probability $1 - p$. At the top is a 1-torus; at the bottom is a 2-torus.

For simplicity, we write $\mathbb{T}_n$ for the n-vertex torus random graph model with uniformly distributed latent positions. We denote by $\mathbb{G}_{n,p}$ the standard Erdős-Rényi random graph with n vertices and edge probability p.

2 Motivation and Hypotheses

Suppose we observe a graph G, representing, for example, a social network. Our goal is to detect local subregions of excessive activity – regions of the network where the edges are unusually dense.[6] We therefore choose a homogeneous null hypothesis: for example, a torus random graph whose latent positions are distributed uniformly.

Note also that the model with uniformly distributed latent positions is *not* a standard Erdős-Rényi random graph, since the edges are not independent – for example, edges ij and ik both depend on the position of vertex i. Below, we consider each of these models as potential null hypothesis graphs.

Our potential alternative hypothesis graphs are torus graphs whose latent positions may not be distributed uniformly. Here, we consider mixtures of uniform distributions for the latent positions; in particular, we consider graphs with a small square region of higher probability.

We have, then, two classes of null hypotheses

and one class of alternative hypotheses concerning our observed graph G:

$$H_0 : G \sim \mathbb{G}_{n,p}$$

$$H_0' : G \sim \mathbb{T}_n$$

$$H_A : G \sim \mathbb{T}_{n,w,p}$$

Notice that, in the 1-torus model, we can write $H_0' : G \sim \mathbb{T}_{n,p,p}$ (that is, as a mixture model with $p = w$). In general, for a k-torus model, we can write H_0' as $G \sim \mathbb{T}_{n,w,p}$ where $p = w^k$, so that the probability X_i comes from this cubical region with side length w is equal to the volume of the region.

We will consider inference in this framework.

3 Statistics

Since we assume that only the graph is observed (and not the latent positions), we must choose statistics that can be calculated from the graph alone.

3.1 Edge density Edge density (number of edges divided by number of possible edges) is one statistic that may be used to distinguish between a "quiet" null hypothesis graph and one containing a subregion of excessive activity. Note that this statistic is only useful if we postulate a specific edge density for our null-hypothesis graph – the edge density tells us nothing about the graph structure, but only tells us if the graph is more active (has more edges) than we expected. So, then, graph size is a useful test statistic for $H_0' : \mathbb{G} = \mathbb{G}_{n,p_0}$ versus $H_A : \mathbb{G} = \mathbb{T}_{n,w=1/4,p=1/2}$, since the alternative graph may be expected to have more edges than the null graph. It is not so useful for $H_0 : \mathbb{G} = \mathbb{G}_{n,p_0}$ versus $H_A' : \mathbb{G} = T_{n,w=1/2,p=1/4}$, since this alternative graph actually has uniformly distributed latent positions, and thus its expected edge density is 1/2, the same as for the Erdős-Rényi null.

3.1.1 (Edge Density) Null Distribution – Exact The expected edge density of an Erdős-Rényi random graph with edge probability p (our H_0) is p. In such a graph, the total number of edges is a sum of independent Bernoulli random variables, so it has a binomial distribution.

The expected edge density of a torus random graph with uniformly distributed latent positions (our H_0') is $\frac{1}{2}$, since $P(i \sim j) = \frac{1}{2}$ for arbitrary vertices i and j with unspecified latent positions. This probability is easily calculated by conditioning on the positions j – see Appendix. (If the distribution of latent positions is uniform, we may assume, without loss of generality, that the position of vertex i is 0; if it weren't, we could simply rotate the torus to make it so.)

The variance (under H_0') of the edge density of $\mathbb{T}_n$ is $\frac{1}{4\binom{n}{2}}$, so the edge density's second moment is $\frac{1}{4}\left(1 + \binom{n}{2}^{-1}\right)$. For calculation of the variance, see Appendix.

We do not yet have a closed form for the exact distribution of the edge density under H_0'; its first few moments are the same as those for the H_0 distribution, but Kolmogorov-Smirnov goodness-of-fit tests [3] suggest that the H_0' distribution of the number of edges is in fact not binomial.

3.1.2 (Edge Density) Null Distribution – Approximate For large n, the (normalized binomial) edge density of an Erdős-Rényi random graph is approximately normal.

Since the edge density of a torus random graph with uniformly distributed latent positions is a classical U-statistic, its distribution should also be asymptotically normal as n goes to infinity; for large n, then, we can approximate the edge density as Gaussian.

3.1.3 (Edge Density) Power Characteristics (Alternative Distribution) – Exact We don't yet know the exact distribution of the number of edges in a torus random graph with mixed-uniform latent position.

The expected edge density of such a graph is $P(i \sim j)$, since expected number of edges is simply $\binom{n}{2}P(i \sim j)$, and $P(i \sim j)$ can be calculated by conditioning on the latent positions of i and j. (In the mixed-uniform scenario, we cannot assume that i has latent position 0, since we don't know from which mixture component it is drawn.) The dependence of the edge density on both p and w is displayed in Figure 2.

3.1.4 (Edge Density) Power Characteristics – Monte Carlo The Monte Carlo power of the edge density statistic for our problem, when $\alpha = 0.05$, $H_0' : G \sim \mathbb{T}_n$, and $H_A : G \sim \mathbb{T}_{n,w,p}$, varies from 0.05 (when $p = w$ and so $H_A = H_0'$) to approximately 1 (when ($|p - w|$ is greater than about 0.3). See Figure 3.

3.2 Maximum degree The maximum degree might be expected to have greater power for detecting local regions of excessive activity; the homogeneous main part of the graph will wash out some of the edge-density signal, whereas maximum degree, being a local statistic, contains less noise to prevent it from detecting excessively active small regions. As we shall see, if our null and alternative hypotheses are, respectively, uniform

[3] For justification of the use of Kolmogorov-Smirnov test in this case, see [4]; for the test itself, see [5].

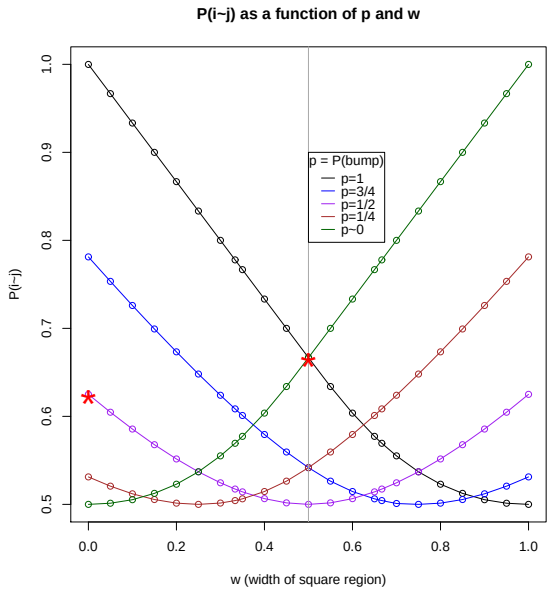


Figure 2: $P(i \sim j)$ as a function of p and w in a mixed-uniform 1-dimensional torus random graph. Note that $P(i \sim j)$ is *not* symmetric in p and w. For example, the starred points mark, respectively, $p = 1/2, w = 0$ and $p = 0, w = 1/2$, but they do not have the same value for $P(i \sim j)$. Since $P(i \sim j)$ is never less than $1/2$, scaling is always required to create sparser graphs. Scaling is likewise required to create *very* dense graphs, except when $p = 0$ or $p = 1$.

and mixed-uniform torus random graphs, maximum degree is very rarely a more powerful statistic than edge density; however, this statistic is useful for selecting an appropriate null hypothesis (choosing between Erdős-Rényi and uniform-torus random graphs for a quiet or homogeneous null hypothesis).

3.2.1 (Maximum Degree) Null Distribution – Exact The exact distribution of the maximum degree is not known in closed form for Erdős-Rényi or torus random graphs.

3.2.2 (Maximum Degree) Null Distribution – Approximate For large n, the maximum in an Erdős-Rényi random graph $\mathbb{G}_{n,p}$ is approximately Gumbel. In particular, as n increases to infinity, the maximum degree in such a graph has a limiting Gumbel distribution with shape and location parameters (calculated by Bollobás, [1])

$$a_0 = pn + \sqrt{2pqn \log n}\left(1 - \frac{\log\log n}{4\log n} - \frac{\log\left(2\sqrt{\pi}\right)}{2\log n}\right)$$

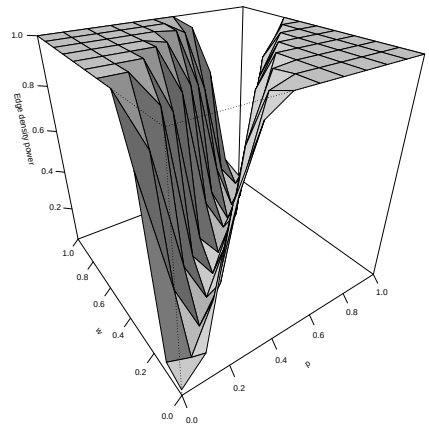


Figure 3: Monte Carlo power of edge density statistic. Note that this power is not symmetric in p and w, though it may appear to be.

$$b_0 = \frac{\sqrt{2pqn \log n}}{2\log n}$$

3.2.3 (Maximum Degree) Null Distribution – Monte Carlo Monte Carlo simulations make it clear that the maximum degree of $\mathbb{T}_n$ does not have the same distribution as the maximum degree of $\mathbb{G}_{n,1/2}$ – the Gumbel distribution with Bollobás's parameters is a very poor fit.

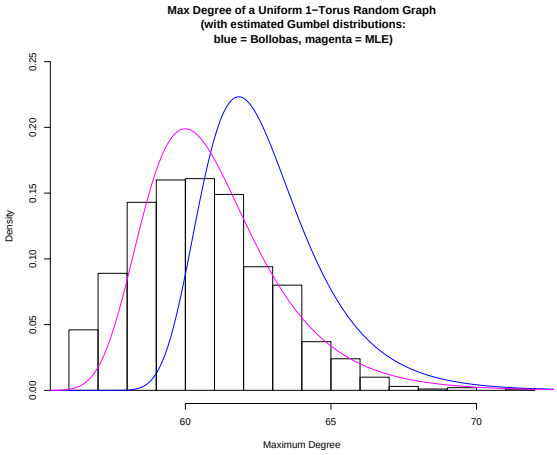


Figure 4: Maximum degree for a 1-torus random graph with uniformly distributed latent positions.

Note that this difference opens up the possibility of doing inference, not just between H_0 and H_A, but between H_0 and H_0'. Unlike the edge density, the

maximum degree may be able to distinguish between our two null hypotheses – see Section 3.2.5 for details.

3.2.4 (Maximum Degree) Power Characteristics (Alternative Distribution) Even if the alternative-hypothesis random graph has the same overall edge density as the null-hypothesis graph, vertices in its more active subregion can be expected to have higher degree than vertices in the more homogeneous null-hypothesis random graph. This intuition does not necessarily hold; see Figure 6. Indeed, when $|p - w| \approx 0.3$, the edge density has greater power.

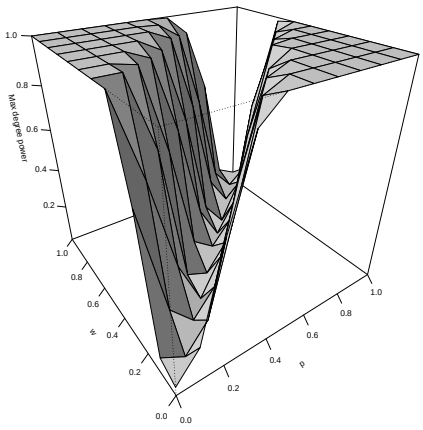


Figure 5: Monte Carlo power of maximum degree statistic.

3.2.5 Other tests using maximum degree The maximum degree, then, is no more useful than edge density for distinguishing H_0' (uniform torus graph) from H_A (mixture torus graph). However, it may be useful for distinguishing H_0 (Erdős-Rényi graph) from H_0'. As Figure 7 suggests, the expected maximum degree of a uniform torus random graph is somewhat lower than that of an Erdős-Rényi random graph.

4 Science News Data

We apply two kinds of hypothesis tests to a network of Science News articles. We embed a set of 579 Science News articles in term-document space, reduce that space to the span of its 573 principal components (the principal components matrix was not full rank because of underflow), and construct a simple distance-based graph: two points are adjacent exactly when they lie within a certain distance of each other in this latent semantic space. We examine the induced subgraph of

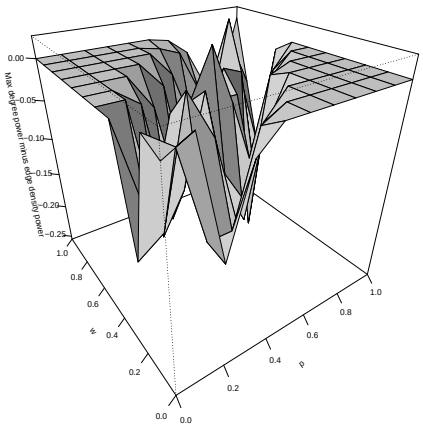


Figure 6: This plot shows the excess power achieved by the maximum degree over that achieved by the edge density.

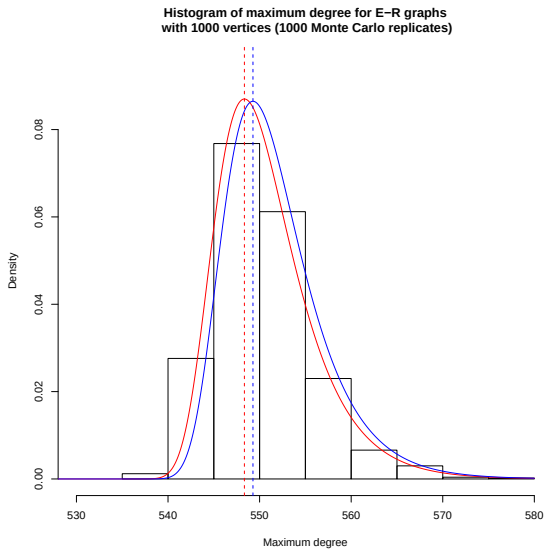
this network whose vertices are articles on mathematics and physics. In this case, the adjacency distance was selected so that this induced subgraph would have edge density approximately 1/2. Edge density is therefore not an appropriate test statistic in this case; we restrict our attention to tests using maximum degree.

First, we select an appropriate null hypothesis: we test H_0 (Erdős-Rényi graph) versus H_0' (uniform torus graph). Monte Carlo simulations suggest that, for $\alpha = 0.05$, a critical value of 104 is appropriate for the maximum degree statistic. The math and physics article network has maximum degree 145. Recall that, for H_0 v. H_0', we reject for *small* values of maximum degree; thus, we have no cause to reject the simpler Erdős-Rényi null hypothesis.

Next, we test our homogeneous null hypothesis against a model allowing for subregions of excessive activity: that is, we test H_0 (Erdős-Rényi) versus H_A (mixed-uniform torus graph). In this case, we reject for *large* values of the maximum degree. Monte Carlo simulations suggest that, in this case, an appropriate critical value (for $\alpha = 0.05$) is 113. Since the math and physics article network has maximum degree 145, we reject the null in this case; this network appears more likely to be a mixed-uniform torus graph than an Erdős-Rényi graph.

5 Higher Dimensions

If the latent positions of the vertices lie on a torus of dimension greater than 1, there is no longer just one natural choice of metric to associate with the latent



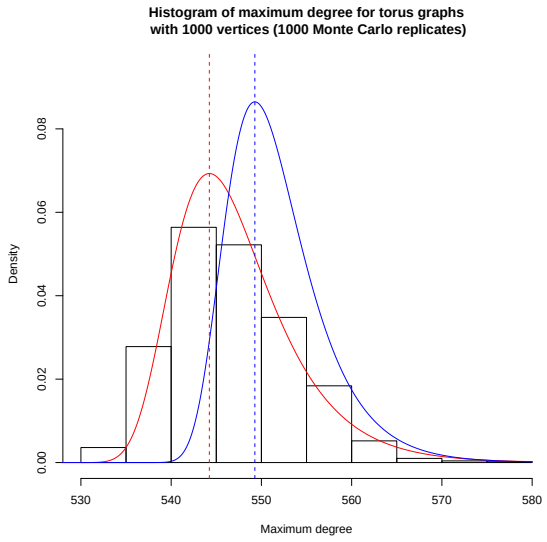


Figure 7: These plots illustrate the difference between H_0 and H_0' in terms of the distribution of the maximum degree. The red curves are maximum-likelihood Gumbels; the blue curves show the approximate Gumbel distribution for the maximum degree of an Erdős-Rényi random graph, as determined by Bollobas [1].

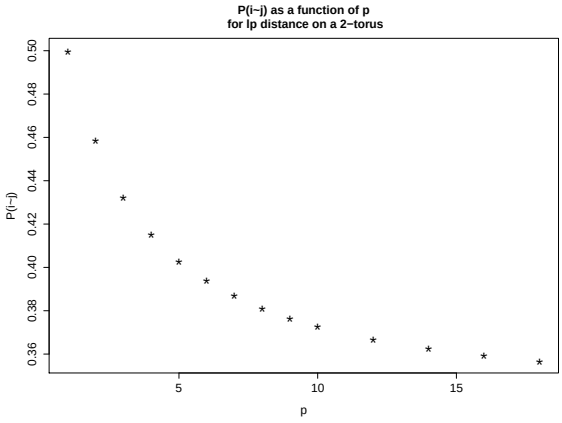


Figure 8: Estimates of $P(i \sim j)$ for uniformly distributed latent position on S^2 with l_p distance as the associated metric. When $p = \infty$, $P(i \sim j) = 1/3$.

space. On S^1, all l_p distances are equivalent to l_1 or angular distance. For any l_p distance on S^1, and for l_1 distance on any S^k, $P(i \sim j) = \frac{1}{2}$. On S^2, the choice of metric can greatly affect the structure of the resulting graph. Figure 8 shows the results of Monte Carlo estimation of $P(i \sim j)$ when the latent positions are uniformly distributed on S^2 and the associated metric is the l_p metric on the torus, for several values of p. Under the l_∞ norm, $P(i \sim j) = \frac{1}{3}$.

A Calculation of $P(i \sim j)$ in a simple case

If the latent positions are uniformly distributed on S^k and the metric associated with the latent space is l_1 distance, then $P(i \sim j) = \frac{1}{2}$ The calculation proceeds as follows:

$$P(i \sim j) = \int_0^1 \dots \int_0^1 \left(1 - \frac{d_1(\vec{x}, \vec{0})}{\sqrt{k}}\right) \frac{1}{(2)^k} dx_1 \dots dx_k$$

$$= \frac{1}{(2)^k} \int_0^1 \dots \int_0^1 \left(1 - \frac{\sum_{r=1}^k |x_r|}{k}\right) dx_1 \dots dx_k$$

$$= \frac{2^k}{(2)^k} \int_0^{1/2} \dots \int_0^{1/2} \left(1 - \frac{\sum_{r=1}^k x_r}{mk}\right) dx_1 \dots dx_k$$

$$= \int_0^{1/2} \dots \int_0^{1/2} 1 dx_1 \dots dx_k$$

$$-\frac{1}{k} \int_0^{1/2} \dots \int_0^{1/2} \sum_{r=1}^k x_r dx_1 \dots dx_k$$

$$= -\frac{1}{k} \int_0^{1/2} \dots \int_0^{1/2} \sum_{r=1}^k x_r dx_1 \dots dx_k$$

$$= 1 - \frac{1}{k}\int_0^{1/2} \dots \int_0^{1/2} \sum_{r=1}^{k} x_r dx_1 \dots dx_k$$

$$= 1 - \frac{1}{k}\frac{k}{2}$$

$$= 1 - \frac{1}{2}$$

$$= \frac{1}{2}$$

B Calculation of Edge-Density Variance for $\mathbb{T}_n$

Let X_n be the number of edges in $\mathbb{T}_n$, and let I_{ij} be an indicator random variable whose value is 1 if the edge ij is present and 0 otherwise. Then, clearly, we have

$$X_n = \sum_{i<j} I_{ij},$$

so calculating the variance of X_n is straightforward:

$$Var[X_n] = Var\left[\sum_{i<j} I_{ij}\right]$$

$$= \sum_{i<j} Var[I_{ij}] + \sum_{i<j,k<l,(i,j)\neq(k,l)} Cov[I_{ij}, I_{kl}]$$

Since edge probabilities depend only on the position of the edge's own endpoints, $Cov[I_{ij}, I_{kl}] = 0$ when i, j, k, l are all distinct. So we have

$$Var[X_n] = \dots = \sum_{i<j} Var[I_{ij}] + \sum_{i\neq j\neq k, i<k} Cov[I_{ij}, I_{jk}].$$

Now, I_{ij} is a Bernoulli random variable, and we established above that $p := P[I_{ij} = 1] = \frac{1}{2}$; therefore, $Var[I_{ij}] = p(1-p) = \frac{1}{4}$. Conditioning on the positions of i (which may be assumed to be 0), j (which may be assumed to be between 0 and $\frac{1}{2}$), and k (which may only be assumed to be somewhere on the torus), we calculate $Cov[I_{ij}, I_{jk} = 0$ (for $i \neq j \neq k$, $i < k$). So now we have

$$Var[X_n] = \dots = \sum_{i<j} \frac{1}{4} + \sum_{i\neq j\neq k, i<k} 0$$

$$= \frac{1}{4}\binom{n}{2}.$$

Now, the edge density of $\mathbb{T}_n$ is $\frac{X_n}{\binom{n}{2}}$; its variance, therefore, is

$$\frac{Var[X_n]}{\binom{n}{2}^2} = \frac{1}{4\binom{n}{2}}.$$

References

[1] B. Bollobás, *Random Graphs*, Cambridge University Press, 2001.

[2] D. Johannsen and J. L. Solka, *Metric MDS to Surfaces*, in submission, 2007.

[3] P. D. Hoff, A. E. Raftery, and M. S. Handcock, *Latent Space Approaches to Social Network Analysis*, JASA, 2002, pp. 1090–1098.

[4] G. Marsaglia, W. W. Tsang, and J. Wang, *Evaluating Kolmogorov's distribution*, J. Stat. Soft., 8, no. 18 (2003).

[5] F. J. Massey, *The Kolmogorov-Smirnov Test for Goodness of Fit*, JASA 46, no. 253 (1951), pp. 68-78.

[6] C. E. Priebe and J. M. Conroy and D. J. Marchette and Y. Park, *Scan Statistics on Enron Graphs*, Computational & Mathematical Organization Theory 11 (2005), pp. 229-247.