Implications of sparsity and high triangle density for graph representation learning

Hannah Sansford
University of Bristol

Alexander Modell
University of Bristol

Nick Whiteley
University of Bristol

Patrick Rubin-Delanchy
University of Bristol

Abstract

Recent work has shown that sparse graphs containing many triangles cannot be reproduced using a finite-dimensional representation of the nodes, in which link probabilities are inner products. Here, we show that such graphs can be reproduced using an infinite-dimensional inner product model, where the node representations lie on a low-dimensional manifold. Recovering a global representation of the manifold is impossible in a sparse regime. However, we can zoom in on local neighbourhoods, where a lower-dimensional representation is possible. As our constructions allow the points to be uniformly distributed on the manifold, we find evidence against the common perception that triangles imply community structure.

1 INTRODUCTION

Network analysis is a common activity in many areas of science and technology, and has diverse goals as a result. Sometimes, it is of interest to estimate just a few, salient, parameters which generally drive connectivity, for example to inform policy (as the 'R' number did, in Britain, during the COVID-19 pandemic (UK-government, 2020)). A considerable amount of research has gone into developing simple models, such as the Erdös-Renyi graph, preferential attachment (Barabási and Albert, 1999), or the Chung-Lu model (Chung and Lu, 2002), to explain global properties of real-world networks. However, in machine-learning, the goal is often to uncover hidden structure or make predictions about the nodes or edges, and these simple models are often felt to say too little about the data. Instead, the paradigm of graph embedding has emerged, in which each node is represented as a vector. There are several theoretical reasons to do this, but the most straightforward is convenience, including compatibility with existing machine-learning techniques, which often take feature vectors as input.

In the design of graph embedding algorithms, providing a realistic generative model for the data has not always been high priority. In fact, in some of the most popular approaches, such as DeepWalk (Perozzi et al., 2014) or Node2vec (Grover and Leskovec, 2016), there is no obvious generative model. But this is a problem for many applications, particularly in unsupervised settings, as there is no straightforward measure of goodness-of-fit. Moreover, data retention and privacy issues have increased the need for synthetic graph generation techniques. It would be useful to be able to harness the power of graph embeddings to reproduce rich structure present in real graphs, *as well as their global properties*.

There are approaches to graph embedding which do purport to model the data. Of particular interest here is the highly studied random dot product graph (RDPG) (Young and Scheinerman, 2007; Athreya et al., 2017), in which the nodes each have a latent position $X_i \in \mathbb{R}^d$, which we aim to estimate via graph embedding, and the inner product $\langle X_i, X_j \rangle$ provides the corresponding connection probability. The model has been particularly useful for understanding graph spectral embedding, where the node representations are obtained from the eigenvectors of some matrix representation of graph, such as the normalised Laplacian (Tang and Priebe, 2018). However, a recent result has put its adequacy as a *generative model* in doubt. The random dot product graph is unable to produce graphs in which the nodes make few connections but still form large numbers of triangles (Seshadhri et al., 2020). These properties, known as sparsity and transitivity, are regarded as ubiquitous in network science (Newman, 2018) and are of course consistent with our own experience of social networks: most of us have few connections, relative to the size of the network and yet, often, 'a friend of a friend is a friend'.

A follow-on paper (Chanpuriya et al., 2020), which proved an inspiration to us, was able to reproduce sparsity and high triangle density (formal definition to come), in theory, by modifying the connection probability to $\max\{0, \min(\langle X_i, Y_j \rangle, 1)\}$ and, in practice, using the ker-

Proceedings of the 26th International Conference on Artificial Intelligence and Statistics (AISTATS) 2023, Valencia, Spain. PMLR: Volume 206. Copyright 2023 by the author(s).

nel logistic$(\langle X_i, Y_j \rangle)$ (the latter already in common use), assigning two latent positions, X_i and Y_i, to each node i. More recently, Stolman et al. (2022) have attained the properties of interest using the softmax kernel over a low-dimensional embedding.

It is important to appreciate that, in our approach, the inner product is not an arbitrary kernel. It provides a default choice, because under mild assumptions it can be used to represent any other positive definite kernel, although typically in infinite dimension, by Mercer's theorem, known in machine-learning as the 'kernel trick' (Steinwart and Christmann, 2008). We find a transparent relationship between the eigenvalues of the kernel and the triangle density (see the supplementary material — Lemma 7), which shows negative eigenvalues can only *reduce* the expected number of triangles in the graph. As a result, if a carefully crafted kernel seems to reproduce high triangle density (Chanpuriya et al., 2020), and negative eigenvalues aren't the reason, the likely problem with the RDPG is with its finite-dimension assumption.

There would be little hope for practical inference in infinite dimension without structure. Instead, this paper shows that *sparse graphs with high triangle density can be modelled using an infinite-dimensional inner product model, with representations on a low dimensional manifold.*

An implication of this manifold structure, that we later explore in depth, is the plausibility of global versus local embedding. Under levels of sparsity which are generally deemed reasonable, the manifold does not get populated everywhere, and perfect reconstruction is impossible. However, since a d-dimensional manifold (informally) locally resembles d-dimensional Euclidean space, we can 'zoom in' to the subgraph containing nodes in a small neighbourhood of the latent space, and form an embedding which represents a flat approximation of the manifold over that region. This encompasses, but generalises, the notion of finding local community structure (Rohe and Qin, 2013; Spielman and Teng, 2013; Chen et al., 2020). A key technical contribution of our work is to construct manifolds for which we can simultaneously control: sparsity, triangle density, and the on-manifold distribution. One point of the latter is that we can place a uniform distribution on the manifold, and still achieve sparse graphs with high triangle density. Because there is no reasonable notion of community structure in this case, we find evidence against the common perception that triangles imply community structure.

Related work Several earlier works have proposed infinite dimensional versions of the RDPG and/or the hypothesis that latent positions live on a low-dimensional manifold (Tang et al., 2013; Athreya et al., 2021; Rubin-Delanchy, 2020; Lei, 2021; Whiteley et al., 2021). The main goal of this paper, to generatively achieve sparsity and high triangle density, was not considered in these papers and indeed they contain several assumptions which make the goal impossible, including a finite rank or trace-class kernel. The paper Whiteley et al. (2021) provides a crucial element of our theory, allowing a manifold construction for which we can control the on-manifold distribution. We principally focus on triangle density, rather than (the formal definition of) transitivity, to align our work with Seshadhri et al. (2020); however, our graphs also achieve high transitivity.

2 ACHIEVING HIGH TRIANGLE DENSITY IN SPARSE GRAPHS

Setup We consider a sequence of simple undirected graphs, $\mathbf{A}^{(n)}$, on nodes indexed $1, \ldots, n$. We make no distinction between the graph and its adjacency matrix, $\mathbf{A}^{(n)}$, which is binary, $n \times n$, and symmetric, with the standard convention that $\mathbf{A}_{ij}^{(n)} = 1$ if nodes i and j have an edge, and zero otherwise.

We will assign to each node, i, an infinite-dimensional latent position, $X_i^{(n)} \in \mathbb{R}^{\mathbb{N}}$, from a probability distribution F_n whose support $\mathcal{M}_n$ (the precise definition of which comes later) only allows inner products that are valid probabilities, that is,

$$\langle x, y \rangle := \sum_{k=1}^{\infty} x_k y_k \in [0, 1] \quad \text{for all } x, y \in \mathcal{M}_n. \qquad (1)$$

For convenience, the superscript (n) will be suppressed from the latent positions, but there is no correspondence between the latent positions of graphs of different sizes, i.e. X_i associated with $\mathbf{A}^{(n)}$ is not the same as X_i associated with $\mathbf{A}^{(m)}$, for $n \neq m$.

Definition 1 (Infinite-dimensional random dot product graph)**.** Let $X_1, \ldots, X_n \overset{i.i.d.}{\sim} F_n$. The matrix $\mathbf{A}^{(n)}$ represents an infinite-dimensional random dot product graph (IRDPG) if

$$\mathbf{A}^{(n)} \overset{ind}{\sim} \text{Bernoulli}(\langle X_i, X_j \rangle),$$

conditional on $X_1, \ldots, X_n$, for $1 \leq i \leq j \leq n$.

This model has appeared, at different levels of generality, in many previous works (Young and Scheinerman, 2007; Tang et al., 2013; Athreya et al., 2017; Lei, 2021), and encodes several assumptions which we review at the end of this section. Of particular interest here is the trade-off between two key properties: the graph sparsity factor, ρ_n, and triangle density, Δ_n, where (He, 2015; Seshadhri et al., 2020)

$$n\rho_n = n \iint \langle x, y \rangle \mathrm{d}F_n(x) \mathrm{d}F_n(y);$$

$$n\Delta_n = \binom{n}{3} \iiint \langle x, y \rangle \langle y, z \rangle \langle z, x \rangle \mathrm{d}F_n(x) \mathrm{d}F_n(y) \mathrm{d}F_n(z),$$

are respectively the graph's expected degree, $\mathbb{E}(\sum_j \mathbf{A}_{ij}^{(n)})$, and expected number of triangles, $\mathbb{E}(\sum_{i\neq j\neq k} \mathbf{A}_{ij}^{(n)}\mathbf{A}_{jk}^{(n)}\mathbf{A}_{ki}^{(n)})$.

Definition 2 (Spectral embedding into $\mathbb{R}^d$)**.** Given $\mathbf{A}^{(n)}$ has eigendecomposition $\mathbf{A}^{(n)} = \sum_i \lambda_i u_i u_i^\top$, its *adjacency spectral embedding* into d dimensions is given by the rows of

$$\hat{\mathbf{X}} := (|\lambda_1|^{1/2}u_1, \dots, |\lambda_d|^{1/2}u_d).$$

The point cloud $\hat{\mathbf{X}}$ can be interpreted as a geometric representation of the best rank-d approximation of $\mathbf{A}^{(n)}$, in the sense that, provided $\lambda_1, \dots, \lambda_d > 0$ (as is generally assumed throughout this paper),

$$\hat{\mathbf{X}} = \underset{\mathbf{X}:\mathrm{rank}(\mathbf{X})\leq d}{\mathrm{argmin}} \|\mathbf{X}\mathbf{X}^\top - \mathbf{A}^{(n)}\|_{\mathrm{F}}.$$

The matrix input may also be rectangular (e.g. the 'core-periphery slice' (Section 3 - local embedding)), in which case, singular values and left singular vectors replace eigenvalues and eigenvectors respectively.

For the reader's convenience, we recall the standard rate-of-growth notation

$$\begin{aligned} o(n) &: \text{'strictly slower than } n\text{'};\\ O(n) &: \text{'slower than } n\text{'};\\ \omega(n) &: \text{'strictly faster than } n\text{'};\\ \Omega(n) &: \text{'faster than } n\text{'};\\ \Theta(n) &: \text{'as fast as } n\text{', i.e., } O(n) \text{ and } \Omega(n). \end{aligned}$$

Question Motivated by recent impossibility results concerning triangle densities under sparse finite-dimensional dot product models (Seshadhri et al., 2020), the question which is the starting point for our contribution is: can a sparse IRDPG, i.e. $\rho_n = o(1)$, have high triangle density, i.e., $\Delta_n = \Omega(1)$?

Main result The answer is yes. Moreover, in contrast to the results of Seshadhri et al. (2020) which show it is impossible to have sparsity and high triangle density when F_n is supported on a finite-dimensional vector space, we show that sparsity and high triangle density *is* possible when F_n is supported on a low-dimensional manifold $\mathcal{M}_n$ embedded in an infinite-dimensional space. We can even obtain the extreme regime $\rho_n = \Theta(1/n), \Delta_n = \Theta(1)$, which includes asymptotically constant expected degree and triangle density.

Approach Our proof is by example: we construct $\mathcal{M}_n$ through a non-distortive homeomorphism ϕ_n applied to a compact topological manifold $\mathcal{Z}_n \subset \mathbb{R}^d$. In our examples, $\mathcal{Z}_n$ is either a growing interval or a growing circle. By non-distortive we mean: the Euclidean length of a curve $\eta(t)$ on $\mathcal{Z}_n$ is equal to the (infinite dimensional) Euclidean length of the image of this curve $\phi_n\{\eta(t)\}$ on $\mathcal{M}_n$. If $\mathcal{Z}_n = [0,1]^2$, for example, we can think of a sheet of paper which is being bent or folded, but not cut or stretched. This property allows us to easily control the on-manifold distribution, an implication of which is given in Section 3 (triangles and community structure). Moreover, the map ϕ_n is chosen in such a way that we can calculate the sparsity and triangle density of the resulting graph. The technical details of the construction are delegated to the supplementary material.

Example 1 Let G_n be a Gaussian distribution with mean zero and variance σ_n^2, truncated to the interval $\mathcal{Z}_n = [-t_n, t_n]$, for $t_n > 0$. We let $t_n = \sigma_n^2$ so that mass of $\mathcal{N}(0, \sigma_n^2)$ on the complement of $\mathcal{Z}_n$ tends to zero as $n \to \infty$.

Lemma 3. There exists a non-distortive homeomorphism $\phi_n^{\mathrm{RBF}} : \mathcal{Z}_n \to \mathbb{R}^{\mathbb{N}}$ such that for $\mathcal{M}_n = \phi_n^{\mathrm{RBF}}(\mathcal{Z}_n)$ and with F_n the pushforward by ϕ_n^{RBF} of the truncated Gaussian distribution G_n,

$$\rho_n = \Theta(\sigma_n^{-1}), \quad \Delta_n = \Theta(n^2\sigma_n^{-2}),$$

so that $\rho_n = \Theta(1/n)$ and $\Delta_n = \Theta(1)$ if $\sigma_n = \Theta(n)$. For this example:

Example 2 Let G_n be the uniform distribution on the circle $\mathcal{Z}_n = \{x \in \mathbb{R}^2 : \|x\| = r_n\}$. For this example:

Lemma 4. There exists a non-distortive homeomorphism $\phi_n^{\mathrm{S}} : \mathcal{Z}_n \to \mathbb{R}^{\mathbb{N}}$ such that for $\mathcal{M}_n = \phi_n^{\mathrm{S}}(\mathcal{Z}_n)$ and F_n the pushforward of G_n by ϕ_n^{S},

$$\rho_n = O(r_n^{-1}), \quad \Delta_n = \Omega(n^2 r_n^{-2}),$$

so that $\rho_n = O(1/n)$ and $\Delta_n = \Omega(1)$ if $r_n = \Theta(n)$.

We note that in the context of Lemma 4, F_n is the uniform probability measure on $\mathcal{M}_n$, since G_n is uniform and ϕ_n^{S} is non-distortive. This is also an example where the expected degrees are identical and therefore bounded for each node, as in Seshadhri et al. (2020).

The two examples are illustrated in Figure 1, with $n = 2000$, $\sigma_n = n/2000$ and $r_n = n/2000$. Seshadhri et al. (2020) note that for various real-world graphs, the subgraphs containing nodes of degree $c \in [10, 50]$ or less have triangle density greater than one (and in many cases greater than ten). Our examples, and several others, are able to produce graphs with (constant) expected degrees and triangle densities in the same range, which we demonstrate in the supplementary materials. We also illustrate that the logistic model studied in Chanpuriya et al. (2020), and employed in Spotify's recommendation system (Johnson, 2014), forms a manifold in (indefinite) inner product space.

Model assumptions Allowing self-loops is not crucial, and is simply done to make the analysis more reader-friendly.

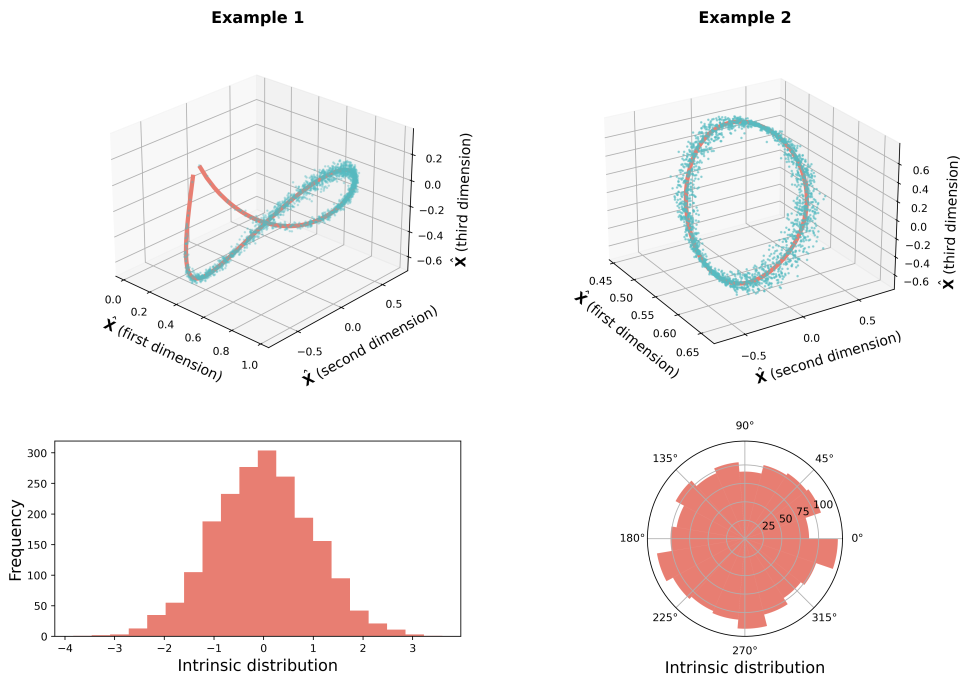

Figure 1: The upper plots show the manifolds of our concrete examples (red curve) with the spectral embeddings (blue points) of a graph simulated from the model, and aligned using Procrustes analysis. The lower plots are histograms showing the intrinsic distribution of X_i, with $n = 2000$, $\sigma_n = n/2000$ and $r_n = n/2000$.

Forbidding self-loops will not change the expected number of triangles, and can only reduce the expected degree by at most one.

That the latent positions are i.i.d., edges conditionally independent, with probability given by an inner product, are testable assumptions. Given an embedding, $\hat{X}_1, \ldots, \hat{X}_n$, the first can be investigated using point process analysis techniques to detect dependence, such as repulsion (Baddeley et al., 2015). The second could be addressed using held out information, such as time, to uncover edge correlations not captured by $\hat{X}_1, \ldots, \hat{X}_n$. A recent paper has observed that edge independence can be problematic for reproducing triangle density while controlling a notion of overlap, for a fixed edge probability matrix (Chanpuriya et al., 2021). However, in our model, the edge probability matrix is random. Regarding the assumption of an inner product, the model cannot explain any large-magnitude negative eigenvalue in $\mathbf{A}^{(n)}$.

These assumptions have been relaxed in many works (Caron and Fox, 2017; Xie and Xu, 2021; Padilla et al., 2019; Rubin-Delanchy et al., 2022; Lei, 2021). However, it is important to appreciate that they are appropriate for our goal: to show that sparsity and high triangle density are possible with node representations on a low-dimensional manifold, under a setup as close as possible to Seshadhri et al. (2020), contrasting their result. Moreover, if these properties are possible within a set of strong assumptions, they are possible more generally.

Relation to graphon models Suppose the node representations are i.i.d. uniform on the unit interval, with connection probabilities given by an *unknown* function $g : [0, 1] \times [0, 1] \to [0, 1]$. If the graphon g is fixed, sparse graphs are impossible since expected degree grows as $\Theta(n)$. If we scale g by a sparsity factor, as is common in the literature (Lunde and Sarkar, 2019; Lei, 2021; Wolfe and Olhede, 2013), then high triangle density becomes impossible. In particular, if $n\rho_n = o(n^{1/3})$, then $\Delta_n = o(1)$. The proof is a straightforward application of the definition of Δ_n and can be found in the supplementary material. Caron et al. (2017) prove that an extension of the graphon model, called *graphex processes*, can simultaneously achieve sparsity and constant clustering coefficients.

3 IMPLICATIONS

We now turn our attention away from the generative properties of the IRDPG, i.e. producing sparsity and high triangle density, to the implications of our model for performing inference on network data.

Global manifold learning Most real world graphs are perceived to be very sparse, in that their average degree is much smaller than n, and it is common in network theory to equate this with the expected degree being $o(\log n)$ (Krivelevich and Sudakov, 2003; Amini et al., 2013; Le et al., 2017). The threshold $\log(n)$ has special significance because it is the rate under which an Erdös-Rényi graph is disconnected with high probability (Bollobás, 1998), and exact recovery of network communities is impossible (Abbe, 2017). The same threshold appears here, but for a different reason.

We can only expect to recover the manifold from $X_1, \ldots, X_n$ if they don't leave large parts of $\mathcal{M}_n$ uncovered. We measure the largest gap using the Hausdorff distance

$$d_H(\{X_i\}, \mathcal{M}_n) := \sup_{x \in \mathcal{M}_n} \left\{ \min_{i \in \{1,\ldots,n\}} d_\circ(X_i, x) \right\},$$

where $d_\circ(x, y)$ denotes geodesic distance on $\mathcal{M}_n$, the length of the shortest curve between x and y on $\mathcal{M}_n$.

Lemma 5. If $\mathcal{M}_n$ is homeomorphic to a closed interval, F_n is uniform on $\mathcal{M}_n$, and there exists $c > 0$ such that $\|x\| \geq c$, for all $x \in \mathcal{M}_n$, then

$$d_H(\{X_i\}, \mathcal{M}_n) \xrightarrow{p} 0 \Rightarrow n\rho_n = \Omega\{\log(n)\}.$$

Given just $X_1, \ldots, X_n$, finding a d-dimensional representation of the nodes in which geodesic distances are faithfully reproduced is therefore impossible in general, unless $n\rho_n = \Omega\{\log(n)\}$. In practice, we only have access to estimates $\hat{X}_1, \ldots, \hat{X}_n$, which adds further complications to manifold learning.

In fact, in this semi-sparse regime, there are problems relating to estimating X_i. For example, existing works on spectral embedding in the infinite-dimensional case make heavy use of the trace-class assumption (Tang et al., 2013; Rubin-Delanchy, 2020; Lei, 2021), $\rho_n^{-1} \sum_{i=1}^{\infty} \lambda_i < \infty$, whereas $n\rho_n = O\{\log(n)\}$ and $\Delta_n = \Omega(1)$ require this sum to be unbounded in n.

Local embedding When $\mathcal{M}_n$ is a manifold, as it is in the examples above, we now explore the idea that we can 'zoom in' to a particular location in latent space and faithfully represent the rich local structure with low-dimensional embeddings.

Given a ball with centre $x \in \mathcal{M}_n \setminus \{0\}$ and radius r, denoted $B_r(x)$, we consider two local views of the graph: the *core*, $\mathbf{A}^{\text{C}}$, is the subgraph on the nodes with latent position in $B_r(x)$; the *core-periphery*, $\mathbf{A}^{\text{C-P}}$, is a graph on the full nodeset with adjacency matrix

$$\mathbf{A}^{\text{C-P}}_{ij} = \begin{cases} \mathbf{A}_{ij} & \text{if } X_i \in B_r(x) \text{ or } X_j \in B_r(x), \\ 0 & \text{otherwise.} \end{cases}$$

In what follows, we will refer to a 'slice' of this matrix, $\mathbf{A}^{\text{C-P slice}} = \mathbf{A}^{\text{C-P}}_{i:X_i \in B_r(x), j \in \{1,\ldots,n\}}$.

We propose to find a low-rank approximate factorisation of these 'local adjacency matrices', $\mathbf{A}^{\text{C}}$ and $\mathbf{A}^{\text{C-P slice}}$, and argue that, under certain conditions, this low-rank assumption is reasonable. We provide some theoretical reasons to prefer the core-periphery; however, our experiments with real data find them comparable.

To describe these graphs as n grows, we have to assume the manifolds are nested, that is, $\mathcal{M}_n \subseteq \mathcal{M}_{n+1}$, for all n. We impose two other regularity conditions. First, $n/v_n \to \infty$, where $v_n = \int_{\mathcal{M}_n} 1 \mathrm{d}x$ is the intrinsic volume of the manifold, which would preclude a constant expected degree in the examples of the previous section, but allows any faster rate. Second, there exists $R \in (0, \|x\|/\sqrt{6})$ and a constant $C > 0$ such that

$$\mathbb{P}\{X_i \in B_r(x)\} \geq \frac{C}{v_n} r^d, \quad \text{for all } r \leq R.$$

Under these assumptions, the core is *dense*, with high triangle density.

Theorem 6. Fix $x \in \mathcal{M}_n \setminus \{0\}$ and let $r_n = \omega\{(v_n/n)^{1/d}\}, r_n \leq R < \|x\|/\sqrt{6}$, which could be fixed or shrinking. The graph core corresponding to the ball $B_{r_n}(x)$ has $\Omega(nr^d/v_n)$ nodes in expectation, expected degree $\Omega(nr^d/v_n)$ and triangle density $\Omega(nr^d/v_n)^2$.

If the manifold is smooth, a local neighbourhood $\mathcal{M}_n \cap B_r(x)$ can be approximated up to arbitrary precision by a d-dimensional plane, the approximation improving as $r \to 0$. However, we can go too far, and collect very few points $X_i \in B_r(x)$. Theorem 6 tells us how r should shrink with n to avoid this.

Under a flat approximation, *the core and core-periphery slice have low rank.* To be precise, the matrices $\mathbf{P}^{\text{C}} := \mathbf{P}_{i,j \in \mathcal{C}}$ and $\mathbf{P}^{\text{C-P slice}} := \mathbf{P}_{i \in \mathcal{C}, j \in \{1,\ldots,n\}}$, where $\mathbf{P} = \langle X_i, X_j \rangle_{i,j \in \{1,\ldots,n\}}$ and $\mathcal{C} = \{i : X_i \in B_r(x)\}$, have rank at most $d+1$. As a result, we can hope to recover local, finite-dimensional estimates of $X_i \in B_r(x)$ using approximate matrix factorisation of $\mathbf{A}^{\text{C}}$ or $\mathbf{A}^{\text{C-P slice}}$. Moreover, the approximate ranks of those matrices provide a local estimate of the manifold dimension.

We find however that, given 'noise-free' factorisations $\mathbf{P}^{\text{C}} = \mathbf{X}^{\text{C}}(\mathbf{X}^{\text{C}})^\top$ and $\mathbf{P}^{\text{C-P slice}} = \mathbf{X}^{\text{C-P}}(\mathbf{Y}^{\text{C-P}})^\top$, where $\mathbf{X}^{\text{C}}, \mathbf{X}^{\text{C-P}} \in \mathbb{R}^{|\mathcal{C}| \times (d+1)}, \mathbf{Y}^{\text{C-P}} \in \mathbb{R}^{n \times (d+1)}$, the original $X_i \in B_r(x)$ are linear images of the rows of $\mathbf{X}^{\text{C-P}}$, but are not necessarily recoverable from $\mathbf{X}^{\text{C}}$.

We have not discussed how a neighbourhood might be found in practice. There are many algorithms suited to this purpose (Haveliwala, 2003; Spielman and Teng, 2004; Andersen et al., 2006), and even a simple similarity measure such as "number of shared neighbours" could be expected to be consistent (Chen and Xu, 2016). The supplementary material contains an experiment demonstrating that this measure forms a reasonable surrogate for the theoretical neighbourhood. There are two other ways a proxy for neighbourhood could be available. First, the graph may have associated covariates, such as time or location, which allow us to zoom

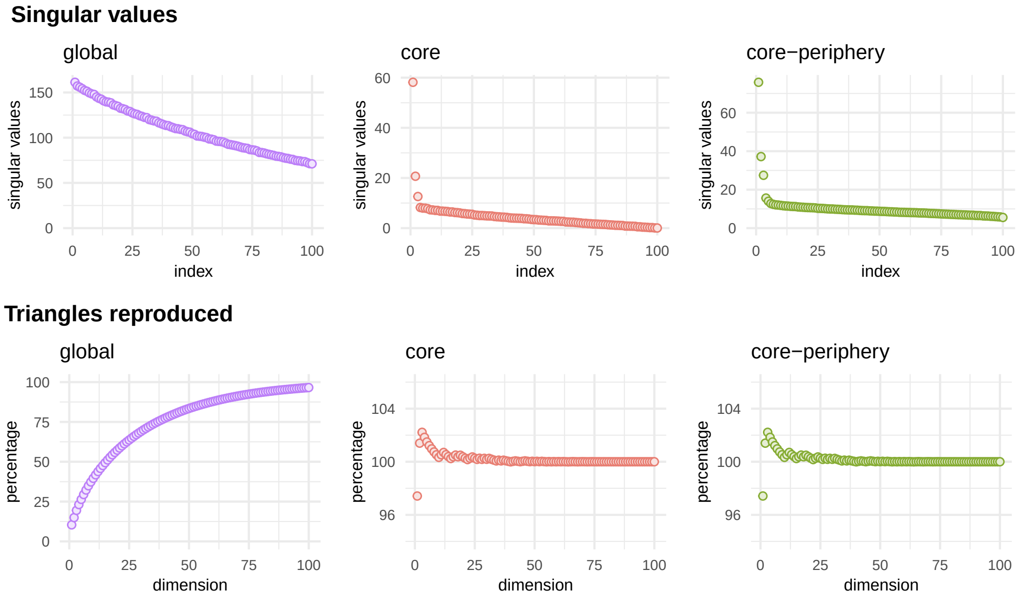

Figure 2: Global versus local embedding. For a graph on 20,000 nodes on a 2-dimensional manifold, we estimate a 200-node neighbourhood around an arbitrary query node (details in main text). Top row: top 100 singular values of the graph, core, and core-periphery slice. The latter, but not the first, roughly distinguish three ($d + 1$) large singular values, indicating the manifold dimension. Bottom row: percentage of triangles reproduced in simulations in the relevant graph using spectral embedding as the dimension increases. The triangle density of the core (middle and right) is well-explained using only a few dimensions (e.g. 1).

into particular subgraphs (see classification experiment). Second, in many applications, a core-periphery graph is precisely all we have. For example, there is a real sense in which the famous Enron email dataset (Cohen, 2009) is a core-periphery view of the network of worldwide emails, focussed on a group of company executives at a period in time.

Triangles and community structure It is known that certain forms of community structure, particularly when heterophilic (e.g. 'opposites attract'), do not give rise to triangles (Newman, 2018). For example, a bipartite graph has no triangles, but may still have well-defined communities. In other words,

$$\text{community structure} \not\Longrightarrow \text{high triangle density}.$$

However, there is a perception that the reverse implication is true (see e.g. Durak et al. (2012); Seshadhri et al. (2020) and references therein). For example, the number of triangles divided by the number of connected triples is commonly known as the 'clustering coefficient' and a high value of this statistic is considered a manifestation of community structure.

To the contrary, our constructions achieve high triangle density, and high clustering coefficients (e.g. ~0.53 in Figure 1, Example 2), when no sensible notion of community structure is present, for example with positions which are uniformly distributed on a manifold. In other words,

$$\text{high triangle density} \not\Longrightarrow \text{community structure}.$$

4 EXPERIMENTS

Simulated graph To compare local versus global embedding, we simulate a graph on a 2-dimensional manifold based on the ϕ_n^{RBF} map from Lemma 3, setting $d = 2$, $n = 20,000$, and $G_n = \text{uniform}[-10, 10]^2$. Picking an arbitrary 'query' node, i, we construct a neighbourhood of the 200 nodes with the most links in common with i. We see a sharp 'elbow' in the spectrum of the core and core-periphery graphs at $3 = d + 1$, whereas there is no obvious cut-off point in the spectrum of the full graph (at least within the first 100 singular values, which are already quite expensive to compute). The triangle density of the core is well-approximated using core or core-periphery embeddings with only a few dimensions (even just 1), but the full graph requires a much higher dimension.

To be clear, we do not claim that a global embedding is inferior to a local embedding, just that large geodesic distances in the former, irrelevant in the latter, must somehow be 'made up' or grossly underestimated by Euclidean distance. On inspection of the central limits for spectral embedding (Rubin-Delanchy et al., 2022, Th. 4 and 7)— which we only treat as indicative because they assume a finite rank — the error distributions of the core positions are the same, up to

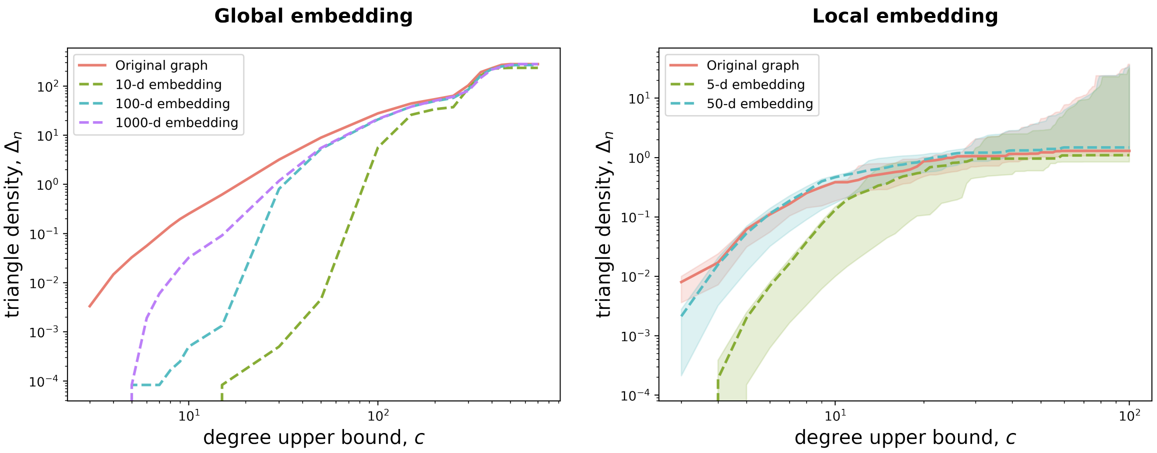

Figure 3: True and recovered triangle counts on subgraphs of a High Energy Physics coauthorship network containing the nodes with degree, at most, c. The number of triangles is divided by size of the graph, n, so that it corresponds to Δ_n. The local neighbourhood used to produce the right-hand plot is constructed by taking the 500 nodes with most links in common with an arbitrary 'query' node. See the main text for details of how the curves and shaded regions are produced. Observe how the local embedding is able to recover triangle counts on much lower degree subgraphs than the global embedding, and with much lower dimensional embeddings.

linear transformation, whether we embed the full graph or the core-periphery. By this (limited) evidence, there is no statistical gain (or loss) in embedding the core-periphery over the whole graph, when we are only interested in the core. The clear, practical advantage of the former is an enormous reduction in dimension.

Recovery of low-degree triangles Seshadhri et al. (2020) empirically validate their claims by demonstrating the inability of various (global) embeddings to recover the low-degree triangles found in real-world graphs. In Figure 3, we illustrate the ability of *local* embeddings to capture this low-degree structure. Specifically, we demonstrate this on the dataset used in Seshadhri et al. (2020): a High Energy Physics coauthorship network with 12,008 nodes (ca-HepPh), obtained from the SNAP graph repository (Leskovec and Krevl, 2014). The plots show the triangle density (y-axis) among nodes with degree, at most, c (x-axis). As in the simulated example, the local neighbourhood is constructed by sampling an arbitrary 'query' node, i, and finding the 500 nodes that have most links in common with i.

The dashed curves on the left-hand (global embedding) plot are the average triangle densities of 100 samples from embeddings of each dimension, while the solid red curve is the true triangle density of the original graph. On the right-hand (local embedding) plot, the solid and dashed curves are the 50th percentiles from 50 sampled neighbourhoods, and the shaded regions encompass the area between the 25th and 75th percentiles. For each sampled neighbourhood, the average triangle density is calculated from 100 graphs sampled from the corresponding embedding.

For an embedding of the core graph into only 5 dimensions, we accurately recover triangle counts on subgraphs with maximum degree greater than 10. Moreover, embedding the core graph into 50 dimensions recovers near-perfect triangle counts for all subgraphs. This is in contrast to the global embedding, which struggles to recover triangles on nodes with degree less than 100, even with a 1000-dimensional embedding. We expect that a graph containing fewer nodes will require a lower dimensional embedding to recover its structure, and hence triangles. However, looking at embedding dimension as a proportion of graph size, a 5-dimensional embedding on our local neighbourhood might be comparable to a 100-dimensional embedding on the entire graph. Likewise, a 50-dimensional embedding on the local graph might be comparable to a 1000-dimensional embedding on the full graph, which is evidently not the case.

Classification To demonstrate the merits of a local core-periphery embedding, we consider a dataset of a diffusion MRI scan of a connectome, first analysed in Priebe et al. (2019). A graph is generated from the scan using the Neuro-Data MR Graph pipeline (Kiar et al., 2018), with vertices representing subregions defined via 'spatial proximity and edges defined by tensor-based fiber streamlines connecting these regions' (Priebe et al., 2019), and we consider only the largest connected component, containing approximately 40,000 nodes. Each node is labelled as belonging to one of 70 regions of the brain and is also labelled with its tissue type: grey or white.

For the six largest regions (each with between 1,351 and 2,316 nodes), we consider the Laplacian spectral embedding (see the supplementary materials for details) of the relevant

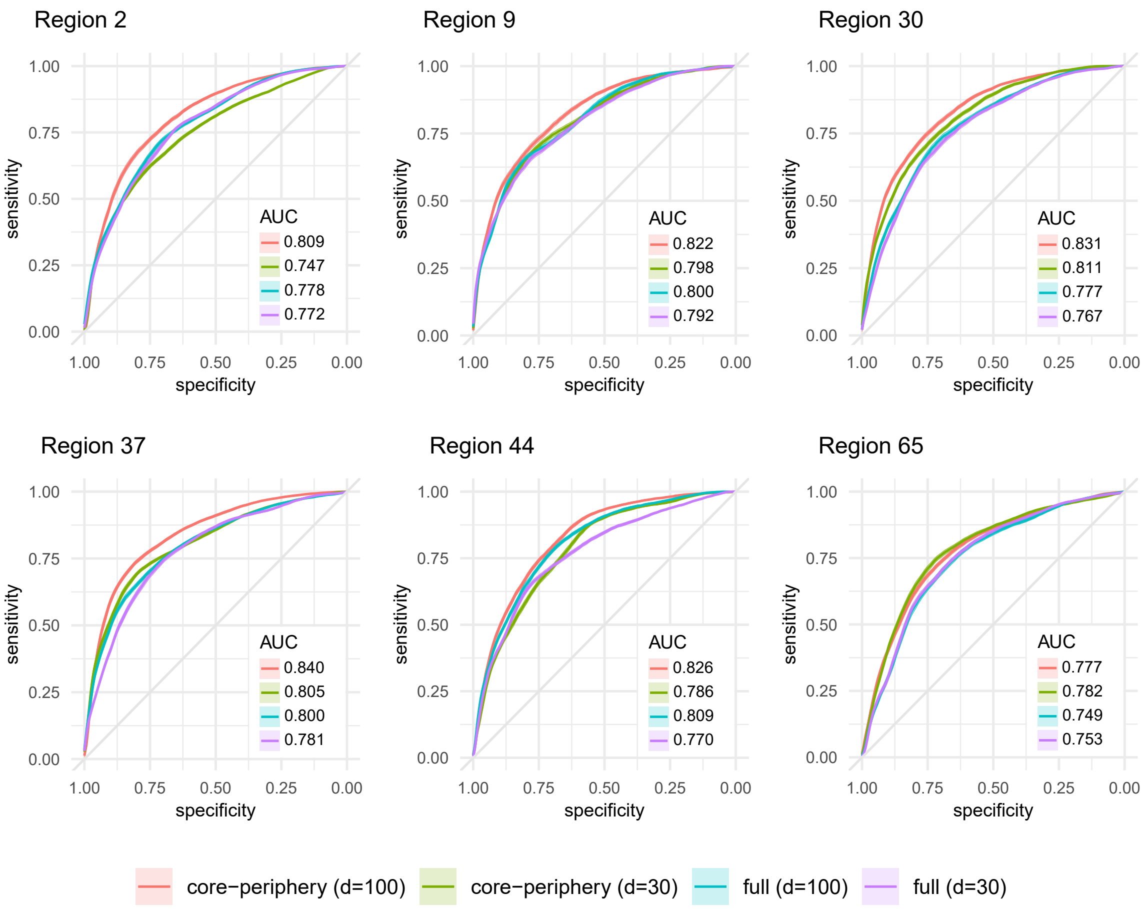

Figure 4: Classification of tissue types from a diffusion MRI scan. For each region, we train a classifier to predict the tissue type of each node using the core-periphery and global embeddings into dimensions $d = 30$ and $d = 100$. For 100 Monte-Carlo iterations, a random sample of 20% of the positions are held out for validation and the corresponding mean receiver-operating curves and mean area-under-the-curve (AUC) statistics are shown. Error bands displaying ± 2 standard errors are shown, but are too small to be visible. The core-periphery embedding performs comparably to the global embedding, even with a lower dimension.

nodes obtained using the full graph, and the corresponding core-periphery 'slice' into both $d = 30$ and $d = 100$ dimensions. 100 and 30 were chosen as roughly logarithmic in the sizes of the full graph and local subgraphs, respectively, the rate required for Wasserstein consistency (Lei, 2021). To give an indication of compute time (which is a loose upper bound on all other experiments in this paper), the 100-dimensional embeddings for the core-periphery and full graphs take approximately 25 seconds and 397 seconds respectively on a 2017 quad-core MacBook Pro.

We use the embeddings as the input to an ℓ_2-regularised logistic regression classifier (the regularisation parameter for which is chosen via 10-fold cross-validation) to predict the tissue type label of each node. For 100 Monte-Carlo iterations, a random subset of 20% of the embedding positions are held out for validation, and the rest are used to train the classifier. The mean receiver-operating curves on the validation sets are shown in Figure 4, with bands showing ± 2 standard errors.

It can be seen that, in general, the core-periphery embedding obtains a classification performance which is comparable to that of the global embedding, even with a lower-dimensional embedding. A comparison of core versus core-periphery classification performance is in the supplementary materials.

5 DISCUSSION

This paper shows that sparse graphs with high triangle density are possible under an inner product model in which the node representations lie on a finite-dimensional manifold in infinite-dimensional space. We believe this is a simpler and more general answer to the corresponding impossibility result for inner product models in finite-dimensional space, compared to existing proposals such as the approach of Chanpuriya et al. (2020), where each node has *two representations*, X_i and Y_i, and nodes i and j connect with probability logistic$(\langle X_i, Y_j \rangle)$' . A simpler model is easier to test, communicate, and reason about, and these are important advantages in network modelling applications affecting society, such as epidemiology.

The presence of a manifold embedded in infinite-dimensional space would indicate several possibilities which we haven't yet pursued. For node classification, it is not actually obvious that graph embedding is necessary at all, just as dimension reduction does not always help regression (Jolliffe, 1982). In standard regression, there is an interesting line of research showing that several techniques, including feedforward neural networks (Nakada and Imaizumi, 2020) and K nearest neighbour methods (Kpotufe, 2011), automatically adapt to manifold structure in feature space. In a similar vein, classification algorithms working directly on the graph, i.e., without embedding, might be developed to exploit the intrinsic dimension of the nodes, on more informed assumptions about the manifold than were available previously. Another interesting direction for future work would be to investigate how one might combine embeddings of local neighbourhoods to build a good global representation. Work in this vein has already been conducted in matrix factorisation in the context of recommender systems (Lee et al., 2013; Mackey et al., 2011).

References

Abbe, E. (2017). Community detection and stochastic block models: recent developments. *The Journal of Machine Learning Research*, 18(1):6446–6531.

Amini, A. A., Chen, A., Bickel, P. J., and Levina, E. (2013). Pseudo-likelihood methods for community detection in large sparse networks. *The Annals of Statistics*, 41(4):2097–2122.

Andersen, R., Chung, F., and Lang, K. (2006). Local graph partitioning using pagerank vectors. In *2006 47th Annual IEEE Symposium on Foundations of Computer Science (FOCS'06)*, pages 475–486. IEEE.

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(1):8393–8484.

Athreya, A., Tang, M., Park, Y., and Priebe, C. E. (2021). On estimation and inference in latent structure random graphs. *Statistical Science*, 36(1):68–88.

Baddeley, A., Rubak, E., and Turner, R. (2015). *Spatial point patterns: methodology and applications with R*. CRC press.

Barabási, A.-L. and Albert, R. (1999). Emergence of scaling in random networks. *Dcience*, 286(5439):509–512.

Bollobás, B. (1998). *Modern graph theory*, volume 184. Springer Science & Business Media.

Brislawn, C. (1991). Traceable integral kernels on countably generated measure spaces. *Pacific Journal of Mathematics*, 150(2):229–240.

Caron, F. and Fox, E. B. (2017). Sparse graphs using exchangeable random measures. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 79(5):1295–1366.

Caron, F., Panero, F., and Rousseau, J. (2017). On sparsity, power-law and clustering properties of graphex processes. *arXiv preprint arXiv:1708.03120*.

Chanpuriya, S., Musco, C., Sotiropoulos, K., and Tsourakakis, C. (2020). Node embeddings and exact low-rank representations of complex networks. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., and Lin, H., editors, *Advances in Neural Information Processing Systems*, volume 33, pages 13185–13198. Curran Associates, Inc.

Chanpuriya, S., Musco, C., Sotiropoulos, K., and Tsourakakis, C. (2021). On the power of edge independent graph models. *Advances in Neural Information Processing Systems*, 34.

Chen, F., Zhang, Y., and Rohe, K. (2020). Targeted sampling from massive block model graphs with personalized PageRank. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 82(1):99–126.

Chen, Y. and Xu, J. (2016). Statistical-computational tradeoffs in planted problems and submatrix localization with a growing number of clusters and submatrices. *The Journal of Machine Learning Research*, 17(1):882–938.

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

Cohen, W. (2009). Enron email dataset. `http://www.cs.cmu.edu/enron`.

Devroye, L. (1981). Laws of the iterated logarithm for order statistics of uniform spacings. *The Annals of Probability*, 9(5):860–867.

Durak, N., Pinar, A., Kolda, T. G., and Seshadhri, C. (2012). Degree relations of triangles in real-world networks and graph models. In *Proceedings of the 21st ACM International Conference on Information and Knowledge Management*, CIKM '12, page 1712–1716, New York, NY, USA. Association for Computing Machinery.

Grover, A. and Leskovec, J. (2016). node2vec: Scalable feature learning for networks. In *Proceedings of the 22nd ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 855–864.

Haveliwala, T. H. (2003). Topic-sensitive PageRank: A context-sensitive ranking algorithm for web search.

IEEE transactions on knowledge and data engineering, 15(4):784–796.

He, R. (2015). *A Graphon-based Framework for Modeling Large Networks*. Columbia University.

Johnson, C. C. (2014). Logistic matrix factorization for implicit feedback data. *Advances in Neural Information Processing Systems*, 27(78):1–9.

Jolliffe, I. T. (1982). A note on the use of principal components in regression. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 31(3):300–303.

Kiar, G., Bridgeford, E. W., Roncal, W. R. G., Chandrashekhar, V., Mhembere, D., Ryman, S., Zuo, X.-N., Margulies, D. S., Craddock, R. C., Priebe, C. E., et al. (2018). A high-throughput pipeline identifies robust connectomes but troublesome variability. *bioRxiv*, page 188706.

Kpotufe, S. (2011). k-NN regression adapts to local intrinsic dimension. *Advances in neural information processing systems*, 24.

Krivelevich, M. and Sudakov, B. (2003). The largest eigenvalue of sparse random graphs. *Combinatorics, Probability and Computing*, 12(1):61–72.

Le, C. M., Levina, E., and Vershynin, R. (2017). Concentration and regularization of random graphs. *Random Structures & Algorithms*, 51(3):538–561.

Lee, J., Kim, S., Lebanon, G., and Singer, Y. (2013). Local low-rank matrix approximation. In Dasgupta, S. and McAllester, D., editors, *Proceedings of the 30th International Conference on Machine Learning*, volume 28 of *Proceedings of Machine Learning Research*, pages 82–90, Atlanta, Georgia, USA. PMLR.

Lei, J. (2021). Network representation using graph root distributions. *The Annals of Statistics*, 49(2):745–768.

Leskovec, J. and Krevl, A. (2014). SNAP Datasets: Stanford large network dataset collection. `http://snap.stanford.edu/data`.

Luke, Y. L. (1972). Inequalities for generalized hypergeometric functions. *Journal of Approximation Theory*, 5(1):41–65.

Lunde, R. and Sarkar, P. (2019). Subsampling sparse graphons under minimal assumptions.

Mackey, L., Jordan, M., and Talwalkar, A. (2011). Divide-and-conquer matrix factorization. In Shawe-Taylor, J., Zemel, R., Bartlett, P., Pereira, F., and Weinberger, K., editors, *Advances in Neural Information Processing Systems*, volume 24. Curran Associates, Inc.

Nakada, R. and Imaizumi, M. (2020). Adaptive approximation and generalization of deep neural network with intrinsic dimensionality. *J. Mach. Learn. Res.*, 21:174–1.

Newman, M. (2018). *Networks: an introduction*. Oxford university press.

Padilla, O. H. M., Yu, Y., and Priebe, C. E. (2019). Change point localization in dependent dynamic nonparametric random dot product graphs. *arXiv preprint arXiv:1911.07494*.

Perozzi, B., Al-Rfou, R., and Skiena, S. (2014). Deepwalk: Online learning of social representations. In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 701–710.

Priebe, C. E., Park, Y., Vogelstein, J. T., Conroy, J. M., Lyzinski, V., Tang, M., Athreya, A., Cape, J., and Bridgeford, E. (2019). On a two-truths phenomenon in spectral graph clustering. *Proceedings of the National Academy of Sciences*, 116(13):5995–6000.

Rohe, K. and Qin, T. (2013). The blessing of transitivity in sparse and stochastic networks. *arXiv preprint arXiv:1307.2302*.

Rubin-Delanchy, P. (2020). Manifold structure in graph embeddings. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., and Lin, H., editors, *Advances in Neural Information Processing Systems*, volume 33, pages 11687–11699. Curran Associates, Inc.

Rubin-Delanchy, P., Cape, J., Tang, M., and Priebe, C. E. (2022). A statistical interpretation of spectral embedding: the generalised random dot product graph. *Journal of the Royal Statistical Society: Series B*. To appear (arxiv preprint arXiv:1709.05506).

Seshadhri, C., Sharma, A., Stolman, A., and Goel, A. (2020). The impossibility of low-rank representations for triangle-rich complex networks. *Proceedings of the National Academy of Sciences*, 117(11):5631–5637.

Shi, T., Belkin, M., and Yu, B. (2009). Data spectroscopy: Eigenspaces of convolution operators and clustering. *The Annals of Statistics*, pages 3960–3984.

Spielman, D. A. and Teng, S.-H. (2004). Nearly-linear time algorithms for graph partitioning, graph sparsification, and solving linear systems. In *Proceedings of the thirty-sixth annual ACM symposium on Theory of computing*, pages 81–90.

Spielman, D. A. and Teng, S.-H. (2013). A local clustering algorithm for massive graphs and its application to nearly linear time graph partitioning. *SIAM Journal on computing*, 42(1):1–26.

Steinwart, I. and Christmann, A. (2008). *Support vector machines*. Springer Science & Business Media.

Stolman, A., Levy, C., Seshadhri, C., and Sharma, A. (2022). Classic graph structural features outperform factorization-based graph embedding methods on community labeling. In *Proceedings of the 2022 SIAM International Conference on Data Mining (SDM)*, pages 388–396. SIAM.

Sutherland, W. A. (2009). *Introduction to metric and topological spaces*. Oxford University Press.

Tang, M. and Priebe, C. E. (2018). Limit theorems for eigenvectors of the normalized laplacian for random graphs. *The Annals of Statistics*, 46(5):2360–2415.

Tang, M., Sussman, D. L., and Priebe, C. E. (2013). Universally consistent vertex classification for latent positions graphs. *The Annals of Statistics*, 41(3):1406–1430.

UK-government (2020). The R value and growth rate. `https://www.gov.uk/guidance/the-r-value-and-growth-rate`.

Whiteley, N., Gray, A., and Rubin-Delanchy, P. (2021). Matrix factorisation and the interpretation of geodesic distance. *Advances in Neural Information Processing Systems*, 34.

Wolfe, P. J. and Olhede, S. C. (2013). Nonparametric graphon estimation.

Xie, F. and Xu, Y. (2021). Efficient estimation for random dot product graphs via a one-step procedure. *Journal of the American Statistical Association*, pages 1–14.

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*, pages 138–149. Springer.

A APPENDIX

A.1 Technical details of construction of manifolds

Suppose we have a symmetric positive-definite kernel $f : \mathcal{Z}_n \times \mathcal{Z}_n \to [0,1]$, where $\mathcal{Z}_n \subset \mathbb{R}^d$ is a compact topological manifold, such as the unit cube or sphere. Also suppose that the latent positions of the graph are from a probability distribution G_n on $\mathcal{Z}_n$. Then nodes i and j, with latent positions z_i and z_j, connect with probability $f(z_i, z_j)$.

To characterise the sparsity and triangle density of the graph, observe that

$$\begin{aligned}\rho_n &= \iint f(x,y)\mathrm{d}G_n(x)\mathrm{d}G_n(y),\\ \Delta_n &= \Theta(n^2)\iiint f(x,y)f(y,z)f(z,x)\mathrm{d}G_n(x)\mathrm{d}G_n(y)\mathrm{d}G_n(z).\end{aligned} \tag{2}$$

Now, consider the operator:

$$\mathcal{A}g(x) = \int f(x,y)g(y)\mathrm{d}G_n(y),$$

on functions g which are square-integrable with respect to the measure G_n. The operator $\mathcal{A}$ is self-adjoint since f is symmetric, and Hilbert-Schmidt, since $0 \leq f(x,y) \leq 1$. To compute the triangle density we will need to work with the eigenvalues of $\mathcal{A}$, denoted $\lambda_1, \lambda_2, \ldots$, which are nonnegative since f is positive-definite. The dependence of these eigenvalues and $\mathcal{A}$ on n is not shown in the notation.

The following lemma makes a crucial connection between the eigenvalues of $\mathcal{A}$ and the expected triangle density Δ_n – recall (2).

Lemma 7. *If f is Lipschitz continuous in each variable,*

$$\iiint f(x,y)f(y,z)f(z,x)\mathrm{d}G_n(x)\mathrm{d}G_n(y)\mathrm{d}G_n(z) = \sum_{k=1}^{\infty}\lambda_i^3.$$

Thus by understanding the behaviour of $\sum_{k=1}^{\infty}\lambda_i^3$ we can understand the behaviour of the expected triangle density as $n \to \infty$. This is a key part of our proofs for our explicit examples.

Now, let us consider the Mercer representation (see e.g. Steinwart and Christmann (2008)) of our kernel f:

$$f(x,y) = \sum_{k=1}^{\infty}\gamma_k u_k(x)u_k(y),$$

where $u_1, u_2, \ldots$ and $\gamma_1, \gamma_2, \ldots$ are respectively the orthonormal eigenfunctions and corresponding (positive) eigenvalues of f (their dependence on n implicit), with respect to Lebesgue measure on some ball $\mathcal{B}_n \subset \mathbb{R}^d$ containing $\mathcal{Z}_n$, and the convergence of the sum is absolute and uniform. Finally, we set:

$$\phi_n(z) = [\gamma_1^{1/2}u_1(z)\ \gamma_2^{1/2}u_2(z)\ \cdots]^\top,$$

that is, $\phi_n(z)$ is a vector in $\mathbb{R}^{\mathbb{N}}$, and as a consequence of this definition, $f(x,y) = \langle\phi_n(x), \phi_n(y)\rangle$ for all $x, y \in \mathcal{Z}_n$. Hence, the graph defined by this kernel also defines and infinite-dimensional RDPG, with latent positions $\phi_n(z_i)$. It is neither obvious nor generally true that a kernel produces a map ϕ_n which is homeomorphic and non-distortive, but our examples will use kernels that do.

We can then build a manifold $\mathcal{M}_n$ using a homeomorphism $\mathcal{M}_n := \phi_n(\mathcal{Z}_n)$ which is non-distortive: the Euclidean length of a curve $\eta(t)$ on $\mathcal{Z}_n$ is equal to the (infinite dimensional) Euclidean length of the image of this curve $\phi_n\{\eta(t)\}$ on $\mathcal{M}_n$. If $\mathcal{Z}_n = [0,1]^2$, for example, we can think of a sheet of paper which is being bent or folded, but not cut or stretched.

As a result, the set $\mathcal{M}_n$ is a topological manifold of the same dimension as $\mathcal{Z}_n$, and from a probability distribution G_n on $\mathcal{Z}_n$, which we push forward to $\mathcal{M}_n$ using ϕ_n, we obtain an identical distribution F_n on $\mathcal{M}_n$, when viewed as a density with respect to intrinsic Euclidean volume on the manifold. In simpler terms, a uniform (or normal) distribution on $\mathcal{Z}_n$ gives a uniform (respectively, normal) distribution on $\mathcal{M}_n$.

Therefore, we are in the setting of an infinite-dimensional RDPG, with latent positions on a low-dimensional manifold, and, as long as we have access to the eigenvalues of the $\mathcal{A}$, we can calculate the expected triangle density of the resulting graph.

A.2 Proof of Lemma 7

Note that since it is bounded, the kernel f is Hilbert-Schmidt with respect to G_n in the sense that:

$$\int |f(x,y)|^2 dG_n(x) dG_n(y) < \infty,$$

since the above is in fact bounded by 1, and so the eigenvalues $\lambda_1, \lambda_2, \dots$ are square-summable .

From Brislawn (1991), we have the following:

Lemma 8. Let μ be a σ-finite Borel measure on a second-countable space, X, and let K be a trace class operator on $L^2(X, \mu)$. If the kernel $K(x,x)$ is continuous at (x,x) for almost every x then

$$\text{tr } K = \int K(x,x) d\mu(x).$$

Let us denote by $\mathcal{A}^3$ the operator

$$\mathcal{A}^3 g(u) = \int \left(\int f(u,y) f(y,z) f(z,x) dG_n(y) dG_n(z) \right) g(x) dG_n(x),$$

with eigenvalues λ_i^3. Since the λ_i are square-summable, it follows that $\mathcal{A}^3$ is trace-class.

Now, assuming that f is Lipschitz continuous, we will show that the Lipschitz continuity of the kernel of $\mathcal{A}^3$,

$$f^3(u,x) = \int f(u,y) f(y,z) f(z,x) dG_n(y) dG_n(z),$$

on the diagonal follows. Let $K > 0$ be such that $|f(x_1,y) - f(x_2,y)| < K|x_1 - x_2|$ for all $x_1, x_2, y \in \mathcal{Z}_n$. Then, with $f^2(x,y) = \int f(x,z) f(z,y) \mathrm{d}G_n(z)$, we have

$$\begin{aligned}
|f^2(x,x) - f^2(y,y)| &= \left| \int f(x,z) f(z,x) - f(y,z) f(z,y) \mathrm{d}G_n(z) \right| \\
&\leq \int |f(x,z) f(z,x) - f(x,z) f(z,y) + f(x,z) f(z,y) - f(y,z) f(z,y)| \, \mathrm{d}G_n(z) \\
&\leq \int f(x,z) |f(z,x) - f(z,y)| + f(z,y) |f(x,z) - f(y,z)| \mathrm{d}G_n(z) \\
&\leq 2K|x-y|.
\end{aligned}$$

Hence f^2 is Lipschitz continuous on the diagonal. By a similar argument, we can see that f^2 is Lipschitz continuous in each argument:

$$\begin{aligned}
|f^2(x,y_1) - f^2(x,y_2)| &= \left| \int f(x,z) f(z,y_1) - f(x,z) f(z,y_2) \mathrm{d}G_n(z) \right| \\
&\leq \int f(x,z) |f(z,y_1) - f(z,y_2)| \mathrm{d}G_n(z) \\
&\leq K|y_1 - y_2|,
\end{aligned}$$

and therefore it follows by induction that f^3 is Lipschitz continuous on the diagonal.

Furthermore, since G_n is a probability measure it is σ-finite, hence we can apply Lemma 8 to $\mathcal{A}^3$ to get:

$$\iiint f(x,y) f(y,z) f(z,x) dG_n(x) dG_n(y) dG_n(z) = \sum_{i=1}^{\infty} \lambda_i^3.$$

□

A.3 Proof of Lemma 3

Let $f(x,y)=\exp\{-\frac{1}{2}|x-y|^2\}$, be the Gaussian radial basis function (RBF) and let $\bar{G}_n$ be a Gaussian distribution with mean zero and variance σ_n^2.

Proposition 9. The map ϕ_n^{RBF} associated with f is a non-distortive homeomorphism.

The eigenvalues of the integral operator associated with f and $\bar{G}_n$ are given explicitly by Shi et al. (2009):

$$\gamma_k=\sqrt{\frac{2}{1+2\sigma_n^2+\sqrt{1+4\sigma_n^2}}}\left(\frac{2\sigma_n^2}{1+2\sigma_n^2+\sqrt{1+4\sigma_n^2}}\right)^{k-1}$$

for $k=1,2,\dots$. Let $r_n=2\sigma_n^2/(1+2\sigma_n^2+\sqrt{1+4\sigma_n^2})$ be the ratio that is raised to $(k-1)$ in the expression above. Then using the geometric sum to infinity formula,

$$\sum_{k=1}^{\infty}(r_n^3)^{k-1}=\frac{1}{1-r_n^3}=\frac{(1+2\sigma_n^2+\sqrt{1+4\sigma_n^2})^3}{(1+2\sigma_n^2+\sqrt{1+4\sigma_n^2})^3-8\sigma_n^6},$$

we see that $\sum_{k=1}^{\infty}(r_n^3)^{k-1}=\Theta(\sigma_n)$ and thus $\sum_{k=1}^{\infty}\gamma_k^3=\Theta(\sigma_n^{-2})$.

Now, let G_n be a Gaussian distribution with mean zero and variance σ_n^2, truncated to the interval $[-t_n,t_n]$. We let $t_n=\sigma_n^2$ so that the mass being removed shrinks to zero. Denote the eigenvalues of the integral operator associated with f and G_n by λ_k. Then,

$$\begin{aligned}\sum_{k=1}^{\infty}\lambda_k^3&=\int_{-t_n}^{t_n}f(x,y)f(y,z)f(z,x)\mathrm{d}G_n(x)\mathrm{d}G_n(y)\mathrm{d}G_n(z)\\&=\frac{1}{(1-2\Phi(-t_n/\sigma_n))^3}\int_{-t_n}^{t_n}f(x,y)f(y,z)f(z,x)\mathrm{d}\bar{G}_n(x)\mathrm{d}\bar{G}_n(y)\mathrm{d}\bar{G}_n(z),\end{aligned}$$

where Φ is the standard normal cumulative distribution function.

Using that

$$\sum_{k=1}^{\infty}\gamma_k^3=\int_{-\infty}^{\infty}f(x,y)f(y,z)f(z,x)\mathrm{d}\bar{G}_n(x)\mathrm{d}\bar{G}_n(y)\mathrm{d}\bar{G}_n(z)=\Theta(\sigma_n^{-2}),$$

we will show that $\sum_{k=1}^{\infty}\lambda_k^3=\Theta(\sigma_n^{-2})$. Let

$$\begin{aligned}I_{t_n}&=\int_{-t_n}^{t_n}f(x,y)f(y,z)f(z,x)\mathrm{d}\bar{G}_n(x)\mathrm{d}\bar{G}_n(y)\mathrm{d}\bar{G}_n(z),\\I_{\text{res}}&=\int_{-\infty}^{-t_n}f(x,y)f(y,z)f(z,x)\mathrm{d}\bar{G}_n(x)\mathrm{d}\bar{G}_n(y)\mathrm{d}\bar{G}_n(z)+\int_{t_n}^{\infty}f(x,y)f(y,z)f(z,x)\mathrm{d}\bar{G}_n(x)\mathrm{d}\bar{G}_n(y)\mathrm{d}\bar{G}_n(z),\end{aligned}$$

so that $\sum_{k=1}^{\infty}\gamma_k{}^3=I_{t_n}+I_{\text{res}}$, and

$$\sum_{k=1}^{\infty}\lambda_k^3=\frac{1}{(1-2\Phi(-t_n/\sigma_n))^3}I_{t_n}.$$

First, we will show that $I_{\text{res}}=o(\sigma_n^{-2})$ and therefore $I_{t_n}=\Theta(\sigma_n^{-2})$. Since f is bounded above by 1, we have that

$$\begin{aligned}I_{\text{res}}&\leq\int_{-\infty}^{-t_n}\mathrm{d}\bar{G}_n(x)\mathrm{d}\bar{G}_n(y)\mathrm{d}\bar{G}_n(z)+\int_{t_n}^{\infty}\mathrm{d}\bar{G}_n(x)\mathrm{d}\bar{G}_n(y)\mathrm{d}\bar{G}_n(z)\\&\leq 3\int_{-\infty}^{-t_n}\mathrm{d}\bar{G}_n(x)+3\int_{t_n}^{\infty}\mathrm{d}\bar{G}_n(x)\\&=6\Phi\left(-\frac{t_n}{\sigma_n}\right)\end{aligned}$$

Plugging in $t_n=\sigma_n^2$, we get that $I_{\text{res}}\leq 6\Phi(-\sigma_n)\leq C\exp(-v\sigma_n^2)$, for positive constants $C,v,$. Hence, $I_{\text{res}}=o(\sigma_n^{-2})$. Now, with $t_n=\sigma_n^2$, we have that $(1-2\Phi(-t_n/\sigma_n))^{-3}=(1-2\Phi(-\sigma_n))^{-3}\to 1$ as $n\to\infty$. Therefore, there exists an

$N \in \mathbb{N}$ such that for all $n \geq N$, $(1 - 2\Phi(-\sigma_n))^{-3} \in [1, 1.1]$. It follows that $\sum_{k=1}^{\infty} \lambda_k^3 = \Theta(\sigma_n^{-2})$. Finally, from Lemma 2, we know that

$$\Delta_n = \Theta(n^2) \sum_{k=1}^{\infty} \lambda_k^3,$$

and therefore, $\Delta_n = \Theta(n^2 \sigma_n^{-2})$.

Now, we will show that the sparsity factor, ρ_n, for the graph generated by f and G_n is $\Theta(\sigma_n^{-1})$. The sparsity factor for the graph generated by f and $\bar{G}_n$ is given by

$$\begin{aligned}
\bar{\rho}_n &= \int\int f(x,y) d\bar{G}_n(x) d\bar{G}_n(y) \\
&= \frac{1}{2\pi\sigma_n^2} \int\int \exp\left(\frac{-(x-y)^2}{2}\right) \exp\left(\frac{-(x^2+y^2)}{2\sigma_n^2}\right) dxdy \\
&= \frac{1}{2\pi\sigma_n^2} \int \exp\left(\frac{-y^2(\sigma_n^2+1)}{2\sigma_n^2}\right) \int \exp\left(\frac{-x^2(\sigma_n^2+1) + 2xy\sigma_n^2}{2\sigma_n^2}\right) dxdy \\
&= \frac{1}{2\pi\sigma_n^2} \int \exp\left(\frac{-y^2(\sigma_n^2+1)}{2\sigma_n^2}\right) \int \exp\left(\frac{-(\sigma_n^2+1)\left(x + \frac{y\sigma_n^2}{\sigma_n^2+1}\right)^2 + \frac{y^2\sigma_n^4}{\sigma_n^2+1}}{2\sigma_n^2}\right) dxdy \\
&= \frac{1}{2\pi\sigma_n^2} \int \exp\left(-y^2\left(\frac{\sigma_n^2+1}{2\sigma_n^2} - \frac{\sigma_n^2}{2(\sigma_n^2+1)}\right)\right) \int \exp\left(\frac{-(\sigma_n^2+1)\left(x + \frac{y\sigma_n^2}{\sigma_n^2+1}\right)^2}{2\sigma_n^2}\right) dxdy \\
&= \frac{1}{2\pi\sigma_n^2} \frac{\sqrt{2\pi}\sigma_n}{\sqrt{\sigma_n^2+1}} \int \exp\left(-y^2\left(\frac{(\sigma_n^2+1)^2 - \sigma_n^4}{2\sigma_n^2(\sigma_n^2+1)}\right)\right) dy \\
&= \frac{1}{\sqrt{2\pi}\sigma_n\sqrt{\sigma_n^2+1}} \cdot \frac{\sqrt{2\pi}\sigma_n\sqrt{\sigma_n^2+1}}{\sqrt{2\sigma_n^2+1}} \\
&= \frac{1}{\sqrt{2\sigma_n^2+1}}.
\end{aligned}$$

Therefore, $\bar{\rho}_n = \Theta(\sigma_n^{-1})$. Let

$$\begin{aligned}
I'_{t_n} &= \int_{-t_n}^{t_n} f(x,y) \mathrm{d}\bar{G}_n(x) \mathrm{d}\bar{G}_n(y), \\
I'_{\text{res}} &= \int_{-\infty}^{-t_n} f(x,y) \mathrm{d}\bar{G}_n(x) \mathrm{d}\bar{G}_n(y) + \int_{t_n}^{\infty} f(x,y) \mathrm{d}\bar{G}_n(x) \mathrm{d}\bar{G}_n(y),
\end{aligned}$$

so that $\bar{\rho}_n = I'_{t_n} + I'_{\text{res}}$ and

$$\rho_n = \int_{-t_n}^{t_n} f(x,y) \mathrm{d}G_n(x) \mathrm{d}G_n(y) = \frac{1}{(1 - 2\Phi(-t_n/\sigma_n))^2} I'_{t_n}.$$

Since f is bounded above by 1, we have that

$$\begin{aligned}
I'_{\text{res}} &\leq \int_{-\infty}^{-t_n} \mathrm{d}\bar{G}_n(x) \mathrm{d}\bar{G}_n(y) + \int_{t_n}^{\infty} \mathrm{d}\bar{G}_n(x) \mathrm{d}\bar{G}_n(y) \\
&\leq 2\int_{-\infty}^{-t_n} \mathrm{d}\bar{G}_n(x) + 2\int_{t_n}^{\infty} \mathrm{d}\bar{G}_n(x) \\
&= 4\Phi\left(-\frac{t_n}{\sigma_n}\right)
\end{aligned}$$

Plugging in $t_n = \sigma_n^2$, we get that $I'_{\text{res}} \leq 4\Phi(-\sigma_n) \leq C\exp(-v\sigma_n^2)$, for positive constants $C, v,$. Hence, $I'_{\text{res}} = o(\sigma_n^{-2})$ and thus $I'_{t_n} = \Theta(\sigma_n^{-1})$. Now, with $t_n = \sigma_n^2$, we have that $(1 - 2\Phi(-t_n/\sigma_n))^{-2} = (1 - 2\Phi(-\sigma_n))^{-2} \to 1$ as $n \to \infty$. Therefore, there exists an $N \in \mathbb{N}$ such that for all $n \geq N$, $(1 - 2\Phi(-\sigma_n))^{-2} \in [1, 1.1]$. It follows that $\rho_n = \Theta(\sigma_n^{-1})$. □

A.4 Proof of Lemma 4

Let $f(x,y) = \exp\{x^\top y\}/\exp(r_n)$ on the circle $\mathcal{Z}_n = \{x \in \mathbb{R}^2 : \|x\| = r_n\}$, with $\mathcal{B}_n = \{x \in \mathbb{R}^2 : \|x\| \leq r_n + \epsilon\}$, for some $\epsilon > 0$.

Proposition 10. The map ϕ_n^{S} associated with f is a non-distortive homeomorphism.

The above is proved by a straightforward application of Whiteley et al. (2021, Proposition 2).

Denote the eigenvalues of the integral operator associated with f and G_n by λ_k. Then, by the symmetry of the uniform distribution on $S^1(r_n)$,

$$\begin{aligned}\sum_{k=1}^{\infty} \lambda_k^3 &= \int\int\int f(x,y)f(y,z)f(z,x)\mathrm{d}G_n(x)\mathrm{d}G_n(y)\mathrm{d}G_n(z)\\ &= \int\int f(x',y)f(y,z)f(z,x)\mathrm{d}G_n(y)\mathrm{d}G_n(z)\end{aligned}$$

for any $x' \in S^1(r_n)$. Let $x' = (r_n, 0)$, $\mathcal{R}_\varepsilon = \{x \in \mathbb{R}^2 : d_\circ(x,x') \leq \varepsilon\}$ for some $\varepsilon > 0$ and $f'(x,y) = \exp\{-\frac{1}{2}d_\circ(x,y)^2\}$, where $d_\circ(x,y)$ represents the geodesic distance between x and y on $S^1(r_n)$. Note that $f(x,y) \geq f'(x,y)$ for all $x, y \in S^1(r_n)$, and thus

$$\begin{aligned}\sum_{k=1}^{\infty} \lambda_k^3 &\geq \int\int f'(x',y)f'(y,z)f'(z,x)\mathrm{d}G_n(y)\mathrm{d}G_n(z)\\ &\geq \int_{\mathcal{R}_\varepsilon} f'(x',y)f'(y,z)f'(z,x)\mathrm{d}G_n(y)\mathrm{d}G_n(z).\end{aligned}$$

Notice that this integral is equivalent to integrating over the flat approximation of $S^1(r_n) \cap \mathcal{R}_\varepsilon$ at $x' = (r_n, 0)$, which is simply the vertical line between (r_n, ε) and $(r_n, -\varepsilon)$. In this case, the first coordinate is r_n for all points we are integrating over, and the second coordinate is uniformly distributed on $[-\varepsilon, \varepsilon]$, therefore

$$\sum_{k=1}^{\infty} \lambda_k^3 \geq \frac{1}{(2\pi r_n)^2}\int_{-\varepsilon}^{\varepsilon}\int_{-\varepsilon}^{\varepsilon} f(x_2', y_2)f(y_2,z_2)f(z_2,x_2)\mathrm{d}y_2\mathrm{d}z_2,$$

where x_2', y_2 and z_2 are the second coordinates of x', y and z, respectively. It follows that,

$$\begin{aligned}\sum_{k=1}^{\infty} \lambda_k^3 &= \frac{1}{(2\pi r_n)^2}\int_{-\varepsilon}^{\varepsilon}\int_{-\varepsilon}^{\varepsilon} \exp\left(-\frac{1}{2}\left(y_2^2 + (y_2 - z_2)^2 + z_2^2\right)\right)\mathrm{d}y_2\mathrm{d}z_2\\ &= \frac{1}{(2\pi r_n)^2}\int_{-\varepsilon}^{\varepsilon}\exp(-z_2^2)\int_{-\varepsilon}^{\varepsilon}\exp\left(-(y_2^2 - y_2 z_2)\right)\mathrm{d}y_2\mathrm{d}z_2\\ &= \frac{1}{(2\pi r_n)^2}\int_{-\varepsilon}^{\varepsilon}\exp\left(-z_2^2 + \frac{z_2^2}{4}\right)\int_{-\varepsilon}^{\varepsilon}\exp\left(-\left(y_2 - \frac{z_2}{2}\right)^2\right)\mathrm{d}y_2\mathrm{d}z_2\\ &= \frac{\sqrt{\pi}}{(2\pi r_n)^2}\int_{-\varepsilon}^{\varepsilon}\exp\left(-\frac{3}{4}z_2^2\right)\left[\Phi\left(\sqrt{2}\left(\varepsilon - \frac{z_2}{2}\right)\right) - \Phi\left(\sqrt{2}\left(-\varepsilon - \frac{z_2}{2}\right)\right)\right]\mathrm{d}z_2,\end{aligned}$$

where Φ is the standard normal cumulative distribution function. Since $z_2 \in [-\varepsilon, \varepsilon]$, we can lower bound this by

$$\begin{aligned}\sum_{k=1}^{\infty} \lambda_k^3 &\geq \frac{\sqrt{\pi}}{(2\pi r_n)^2}\left[\Phi\left(\frac{\varepsilon}{\sqrt{2}}\right) - \Phi\left(\frac{-\varepsilon}{\sqrt{2}}\right)\right]\int_{-\varepsilon}^{\varepsilon}\exp\left(-\frac{3}{4}z_2^2\right)\mathrm{d}z_2\\ &= \frac{\sqrt{\pi}}{(2\pi r_n)^2}\mathbb{P}\left(|Z| \leq \frac{\varepsilon}{\sqrt{2}}\right)\int_{-\varepsilon}^{\varepsilon}\exp\left(-\frac{3}{4}z_2^2\right)\mathrm{d}z_2,\end{aligned}$$

where $Z \sim \mathcal{N}(0,1)$. Then, by Chebyshev's inequality,

$$\begin{aligned}\sum_{k=1}^{\infty} \lambda_k^3 &\geq \frac{\sqrt{\pi}}{(2\pi r_n)^2}\left(1-\frac{2}{\varepsilon^2}\right)\int_{-\varepsilon}^{\varepsilon} \exp\left(-\frac{3}{4}z_2^2\right)\mathrm{d}z_2 \\ &= \frac{\sqrt{\pi}}{(2\pi r_n)^2}\left(1-\frac{2}{\varepsilon^2}\right)\frac{2\sqrt{\pi}}{\sqrt{3}}\left[\Phi\left(\frac{\varepsilon}{\sqrt{2/3}}\right)-\Phi\left(\frac{-\varepsilon}{\sqrt{2/3}}\right)\right] \\ &= \frac{1}{2\sqrt{3}\pi r_n^2}\left(1-\frac{2}{\varepsilon^2}\right)\mathbb{P}\left(|Z| \leq \frac{\varepsilon}{\sqrt{2/3}}\right) \\ &\geq \frac{1}{2\sqrt{3}\pi r_n^2}\left(1-\frac{2}{\varepsilon^2}\right)\left(1-\frac{2}{3\varepsilon^2}\right) \\ &\geq \frac{1}{2\sqrt{3}\pi r_n^2}\left(1-\frac{2}{\varepsilon^2}\right)^2.\end{aligned}$$

Letting $\varepsilon = 2$, we get that

$$\sum_{k=1}^{\infty} \lambda_k^3 \geq \frac{1}{8\sqrt{3}\pi r_n^2},$$

and hence $\sum_{k=1}^{\infty} \lambda_k^3 = \Omega(r_n^{-2})$. Finally, using Lemma 7, we have $\Delta_n = \Omega(n^2 r_n^{-2})$.

Now we will show that the sparsity factor, ρ_n, associated with f and G_n is $O(r_n^{-1})$. We have that, by the symmetry of the uniform distribution on $S^1(r_n)$,

$$\rho_n = \int\int f(x,y)\mathrm{d}G_n(x)\mathrm{d}G_n(y) = \int f(x',y)\mathrm{d}G_n(y)$$

for any $x' \in S^1(r_n)$. Therefore,

$$\begin{aligned}\rho_n &= \int \exp\left(-\frac{||x'-y||^2}{2}\right)\mathrm{d}G_n(y) \\ &= \int \exp\left(\frac{-(x'^{\top}x' + y^{\top}y - 2x'^{\top}y)}{2}\right)\mathrm{d}G_n(y) \\ &= \exp(-r_n^2)\int \exp(-x'^{\top}y)\mathrm{d}G_n(y).\end{aligned}$$

Now, letting $x' = r_n e_1$, where $e_1 = (1,0)$, and making the change of variable $u = y/r_n$, so that u is uniformly distributed on the unit circle $S^1(1)$, we get

$$\rho_n = \frac{\exp(-r_n^2)}{r_n}\int \exp(-r_n^2 e_1^{\top}u)\mathrm{d}G(u)$$

where G is the uniform distribution on $S^1(1)$. Multiplying and dividing by $I_0(r_n^2)$, where I_0 denotes the modified Bessel function of the first kind at order 0, we get that the expression inside the integral is the probability density of the von Mises-Fisher distribution with mean direction $-e_1$ and concentration parameter r_n^2:

$$\begin{aligned}\rho_n &= \frac{\exp(-r_n^2)I_0(r_n^2)}{r_n}\int \exp\frac{1}{I_0(r_n^2)}(-r_n^2 e_1^{\top}u)\mathrm{d}F(u) \\ &= \frac{\exp(-r_n^2)I_0(r_n^2)}{r_n}.\end{aligned}$$

It has been shown that $I_0(x) < \cosh x$ (Luke, 1972), and using that $\cosh x < \exp(x)$, we get the upper bound $\rho_n < r_n^{-1}$. It follows that $\rho_n = O(r_n^{-1})$. □

A.5 Proof of Proposition 9

In order to establish that ϕ_n^{RBF} is a homeomorphism, we need to show it is continuous, injective, and has a continuous inverse on its image.

The continuity of ϕ_n^{RBF} follows from the continuity of $(x, y) \mapsto f(x, y)$ combined with the identities:

$$\begin{aligned}\|\phi_n^{\text{RBF}}(x) - \phi_n^{\text{RBF}}(y)\|^2 &= \|\phi_n^{\text{RBF}}(x)\|^2 + \|\phi_n^{\text{RBF}}(y)\|^2 - 2\langle \phi_n^{\text{RBF}}(x), \phi_n^{\text{RBF}}(y)\rangle \\ &= f(x, x) + f(y, y) - 2f(x, y).\end{aligned}$$

The proof of injectivity is by contradiction. Assume there exists $x, y \in \mathcal{Z}_n$ such that $x \neq y$ and $\phi_n^{\text{RBF}}(x) = \phi_n^{\text{RBF}}(y)$. Then $1 = f(x, x) = \langle \phi_n^{\text{RBF}}(x), \phi_n^{\text{RBF}}(x)\rangle = \langle \phi_n^{\text{RBF}}(x), \phi_n^{\text{RBF}}(y)\rangle = f(x, y) = \exp\{-\frac{1}{2}|x - y|^2\} \neq 1$, giving a contradiction as required.

Since $\mathcal{Z}_n$ is compact, a general result concerning injective maps on compact domains, Sutherland (2009)[Prop 13.26], implies that the inverse of ϕ_n^{RBF} on its image must be continuous. Thus we have established ϕ_n^{RBF} is a homeomorphism.

Now let $\eta : [0, 1] \to \mathcal{Z}_n$ be a curve, i.e., a continuously differentiable function. The length of the image of this curve by ϕ_n^{RBF}, i.e. the length of $t \in [0, 1] \mapsto \phi_n^{\text{RBF}}\{\eta(t)\} \in \mathcal{M}_n$ is:

$$\int_0^1 \left\| \frac{\mathrm{d}}{\mathrm{d}t} \phi_n^{\text{RBF}}\{\eta(t)\} \right\| \mathrm{d}t. \tag{3}$$

By Steinwart and Christmann (2008)[Lem 4.34], ϕ_n^{RBF} is continuously differentiable, hence by the chain rule:

$$\frac{\mathrm{d}}{\mathrm{d}t} \phi_n^{\text{RBF}}\{\eta(t)\} = \frac{\mathrm{d}\phi_n^{\text{RBF}}}{\mathrm{d}x}\{\eta(t)\}\frac{\mathrm{d}\eta}{\mathrm{d}t}$$

and

$$\begin{aligned}\left\| \frac{\mathrm{d}}{\mathrm{d}t} \phi_n^{\text{RBF}}\{\eta(t)\} \right\|^2 &= \left\langle \frac{\mathrm{d}}{\mathrm{d}t} \phi_n^{\text{RBF}}\{\eta(t)\}, \frac{\mathrm{d}}{\mathrm{d}t} \phi_n^{\text{RBF}}\{\eta(t)\} \right\rangle \\ &= \left\langle \frac{\mathrm{d}\phi_n^{\text{RBF}}}{\mathrm{d}x}\{\eta(t)\}, \frac{\mathrm{d}\phi_n^{\text{RBF}}}{\mathrm{d}x}\{\eta(t)\} \right\rangle \left| \frac{\mathrm{d}\eta}{\mathrm{d}t} \right|^2 \\ &= \left.\frac{\partial^2 f}{\partial x \partial y}\right|_{\eta(t),\eta(t)} \left| \frac{\mathrm{d}\eta}{\mathrm{d}t} \right|^2.\end{aligned}$$

For the RBF kernel in question, $\left.\frac{\partial^2 f}{\partial x \partial y}\right|_{\eta(t),\eta(t)} = 1$, and substituting into (3), we find:

$$\int_0^1 \left\| \frac{\mathrm{d}}{\mathrm{d}t} \phi_n^{\text{RBF}}\{\eta(t)\} \right\| \mathrm{d}t = \int_0^1 \left| \frac{\mathrm{d}\eta}{\mathrm{d}t} \right| \mathrm{d}t,$$

the right hand side of which is the Euclidean length of η. Hence ϕ_n^{RBF} is non-distortive, as required.

A.6 Proof of Lemma 5

Let ϕ_n be a homeomorphism satisfying $\mathcal{M}_n = \phi_n([0, 1])$. Define $X_{(0)} := \phi_n(0), X_{(n+1)} := \phi_n(1)$ and

$$X_{(k)} = X_\ell : \phi_n^{-1}(X_\ell) \text{ is the } \ell\text{-th smallest of } \phi_n^{-1}(X_1), \dots, \phi_n^{-1}(X_n), \quad k \in \{1, \dots, n\}.$$

Then

$$\begin{aligned}H_n &:= d_H(\{X_i\}, \mathcal{M}_n) = \max\left\{(d_\circ(X_{(1)}, X_{(0)}), d_\circ(X_{(2)}, X_{(1)})/2, \dots, d_\circ(X_{(n)}, X_{(n-1)})/2, d_\circ(X_{(n+1)}, X_{(n)})\right\} \\ 2H_n &\geq \max_{0 \leq k \leq n} \left\{d_\circ(X_{(k+1)}, X_{(k)})\right\} =: M_n\end{aligned}$$

Pick $\epsilon > 0$. Then

$$\mathbb{P}(H_n \geq \epsilon) \geq \mathbb{P}(M_n \geq 2\epsilon) \geq \mathbb{P}\left\{\frac{n}{v_n}M_n - \log(n) \geq \frac{2n\epsilon}{v_n} - \log(n)\right\} \to S\left\{\frac{2n\epsilon}{v_n} - \log(n)\right\},$$

by a classical result on uniform spacings (Devroye, 1981), where $v_n = d_\circ\{\phi_n(1), \phi_n(0)\}$ and $S(x) = 1 - \exp(-e^{-x})$ is the survivor function of the standard Gumbel distribution. Hence, $\mathbb{P}(H_n \geq \epsilon) \to 0$ implies $n/v_n = \Omega\{\log(n)\}$.

Now, pick any r satisfying $0 < r < \sqrt{c}$ and consider any $x \in \mathcal{M}_n$. Then,

$$\begin{aligned} n\rho_n(x) &:= n \int \langle x, y\rangle \mathrm{d}F_n(y), \\ &\geq n\mathbb{E}\left\{\langle x, X\rangle \mid X \in B_r(x)\right\} \mathbb{P}\{X \in B_r(x)\}, \\ &\geq \frac{nr}{v_n} \inf\{\langle x, y\rangle : y \in B_r(x)\}, \\ &\geq \frac{nr}{v_n}(c^2 - cr), \end{aligned}$$

where $X \sim F_n$, $B_r(x)$ denotes a Euclidean ball of centre x and radius r, we have used that geodesic distance is larger than Euclidean distance in the third line, and $\langle x, y\rangle = (\langle x, x\rangle + \langle y, y\rangle - \|x - y\|^2)/2 \geq (c^2 + (c-r)^2 - r^2)/2 = c^2 - cr$ in the fourth. As a result,

$$n\rho_n = n \int \rho_n(x)\mathrm{d}F_n(x) \geq \frac{nr}{v_n}(c^2 - cr) = \Omega\{\log(n)\}.$$

A.7 The graphon model

Let node representations $Z_i \overset{\text{i.i.d.}}{\sim} \mathcal{U}[0,1]$ and $\rho_n > 0$ be monotone non-increasing. Suppose the probability of an edge between nodes i and j is modelled by

$$p_{ij} = \rho_n g(Z_i, Z_j),$$

where $g : [0,1] \times [0,1] \to [0,1]$ is a symmetric function such that $\iint g(x,y)\mathrm{d}x\mathrm{d}y = 1$.

The triangle density Δ_n of the resulting graph is defined by

$$\begin{aligned} \Delta_n &= \frac{(n-1)(n-2)}{6} \iiint_{[0,1]^3} \rho_n g(x,y)\rho_n g(y,z)\rho_n g(z,x)\mathrm{d}x\mathrm{d}y\mathrm{d}z \\ &= \Theta(n^2)\rho_n^3 \iiint_{[0,1]^3} g(x,y)g(y,z)g(z,x)\mathrm{d}x\mathrm{d}y\mathrm{d}z \\ &\leq \Theta(n^2)\rho_n^3 \end{aligned}$$

Therefore $\Delta_n \to 0$ if $\rho_n = o(n^{-2/3})$. Hence, in a sparse regime where average degree is constant with the respect to n, i.e. $\rho_n = o(1/n)$, the triangle density $\Delta_n \to 0$.

A.8 Laplacian spectral embedding

We have referred to spectral embedding in various places in the main document. Given an undirected graph whose adjacency matrix $\mathbf{A}$ has eigendecomposition $\mathbf{A} = \sum_i \lambda_i u_i u_i^\top$, its *adjacency spectral embedding* into d dimensions is given by the rows of

$$\hat{\mathbf{X}} := (|\lambda_1|^{1/2}u_1, \ldots, |\lambda_d|^{1/2}u_d).$$

The point cloud $\hat{\mathbf{X}}$ can be interpreted as a geometric representation of the best rank-d approximation of $\mathbf{A}$, in the sense that, providing $\lambda_1, \ldots, \lambda_d > 0$ (as is generally assumed throughout this paper),

$$\hat{\mathbf{X}} = \underset{\mathbf{X}:\mathrm{rank}(\mathbf{X})\leq d}{\operatorname{argmin}} \|\mathbf{X}\mathbf{X}^\top - \mathbf{A}\|_{\mathrm{F}}.$$

The matrix input may also be rectangular (e.g. the ‘core-periphery slice’ (Section 3 - local embedding)), in which case, singular values and left singular vectors replace eigenvalues and eigenvectors respectively.

The graph Laplacian matrix is defined as $\mathbf{L} = \mathbf{D}^{-1/2}\mathbf{A}\mathbf{D}^{-1/2}$ where $\mathbf{D}$ is the diagonal matrix containing the node degrees (the row-sums of $\mathbf{A}$). An alternative spectral embedding is the *Laplacian spectral embedding* defined analogously to above, with $\mathbf{A}$ replaced with $\mathbf{L}$. Figures 1 and 2 employ adjacency spectral embedding while Figure 4 employs Laplacian spectral embedding.

A.9 Proof of Theorem 6

Let M_n denote the number of latent positions in $B_{r_n}(x)$. Then M_n follows a Binomial distribution with success probability exceeding $\frac{C}{v_n} r^d$, and so $\mathbb{E}(M) = \Omega(nr^d/v_n)$.

The expected degree of the core is

$$\sum_{m=0}^{\infty} m \iint_{B_{r_n}(x)} \langle y, z \rangle \mathrm{d}\bar{F}_n(x) \mathrm{d}\bar{F}_n(y) \mathbb{P}(M_n = m),$$

where $\bar{F}_n$ is the distribution F_n conditioned on landing in $B_{r_n}(x)$.

First, for $y, z \in B_{r_n}(x)$,

$$\begin{aligned}
\langle y, z \rangle &\geq (\langle y, y \rangle + \langle z, z \rangle - \|y - z\|^2)/2 \\
&\geq \{\langle y, y \rangle + \langle z, z \rangle - (2r_n)^2\}/2, \\
&\geq \langle x, x \rangle - 6r_n^2,
\end{aligned}$$

using the triangle inequality $\|x\| \leq \|y\| + \|x - y\|$. As a result, $\langle y, z \rangle$ has a lower bound $\langle x, x \rangle - 6R^2$, which is positive by the theorem assumptions, and the expected degree is larger than

$$(\langle x, x \rangle - 6R^2) \mathbb{E}(M_n) = \Omega(nr^d/v_n).$$

The triangle density of the core is

$$\begin{aligned}
\sum_{m=0}^{\infty} \binom{m}{3} / m \iiint_{B_{r_n}(x)} &\langle y, z \rangle \langle z, u \rangle \langle u, y \rangle \mathrm{d}\bar{F}_n(y) \mathrm{d}\bar{F}_n(z) \mathrm{d}\bar{F}_n(u) \mathbb{P}(M_n = m), \\
&\geq \left(\langle x, x \rangle - 6R^2\right)^3 \mathbb{E}\left(\binom{M_n}{3} / M_n\right)
\end{aligned}$$

The assumption $r_n = \omega\{(v_n/n)^{1/d}\}$ ensures that $M_n \to \infty$ in probability, so that

$$\left(\langle x, x \rangle - 6R^2\right)^3 \mathbb{E}\left(\binom{M_n}{3} / M_n\right) = \Omega(\mathbb{E}\{M_n^2\}) = \Omega(nr^d/v_n)^2.$$

A.10 Recovery of the latent positions from the core and core-periphery embeddings

Let $\mathbf{P} = \mathbf{X}\mathbf{X}^\top$ where $\mathbf{X} \in \mathbb{R}^{n \times D}$ is of rank D and let C denote the indices of m ‘core’ nodes. Let $\mathbf{X}^{\text{C}} \in \mathbb{R}^{m \times D}$ denote the rows of $\mathbf{X}$ in C and assume the rows of $\mathbf{X}^{\text{C}}$ are contained within a d-dimensional subspace of $\mathbb{R}^D$, i.e. $\operatorname{rank}(\mathbf{X}^{\text{C}}) = d$, where $m > d$. Let $\mathbf{P}^{\text{C-P slice}} = \mathbf{X}^{\text{C}}\mathbf{X}^\top$ denote the $m \times n$ matrix containing the rows of $\mathbf{P}$ in C and let $\mathbf{P}^{\text{C-P slice}} = \mathbf{Y}^{\text{C-P}}(\mathbf{Z}^{\text{C-P}})^\top$ where $\mathbf{Y}^{\text{C-P}} \in \mathbb{R}^{m \times d}$, $\mathbf{Z}^{\text{C-P}} \in \mathbb{R}^{n \times d}$, be a rank-$d$ factorisation of $\mathbf{P}^{\text{C-P slice}}$. Then

$$\mathbf{X}^{\text{C}} = \mathbf{Y}^{\text{C-P}} \left[(\mathbf{Z}^{\text{C-P}})^\top \mathbf{X} (\mathbf{X}^\top \mathbf{X})^{-1}\right],$$

so $\mathbf{X}^{\text{C}}$ is a linear transformation of $\mathbf{Y}^{\text{C-P}}$.

Let $\mathbf{P}^{\text{C}} = \mathbf{X}^{\text{C}}(\mathbf{X}^{\text{C}})^\top$ denote the $m \times m$ submatrix containing the rows and columns of $\mathbf{P}$ in C and let $\mathbf{P}^{\text{C}} = \mathbf{Y}^{\text{C}}(\mathbf{Y}^{\text{C}})^\top$, where $\mathbf{Y}^{\text{C}} \in \mathbb{R}^{m \times d}$ be a rank-d factorisation of $\mathbf{P}^{\text{C}}$. There is no such relationship between $\mathbf{X}^{\text{C}}$ and $\mathbf{Y}^{\text{C}}$ since by assumption, $\operatorname{rank}(\mathbf{X}^{\text{C}}) = d$ and so $(\mathbf{X}^{\text{C}})^\top \mathbf{X}^{\text{C}}$ is not invertible. To provide a concrete counter example, consider the $\mathbf{P}$-matrix generated from a stochastic block model with $\mathbf{B}$-matrix

$$\mathbf{B} = \begin{pmatrix} a & a & b \\ a & a & c \\ b & c & a \end{pmatrix}.$$

Let C be the nodes in the first and second community. Then $\mathbf{X}^{\text{C}}$ has rank 2 and $\mathbf{Y}^{\text{C}}$ has rank 1, so they cannot be related by a linear transformation.

A.11 Comparison of the core versus core-periphery embeddings

Figure 5 shows the classification performance of the core-periphery and core embeddings in the real data experiment detailed in Section 3.

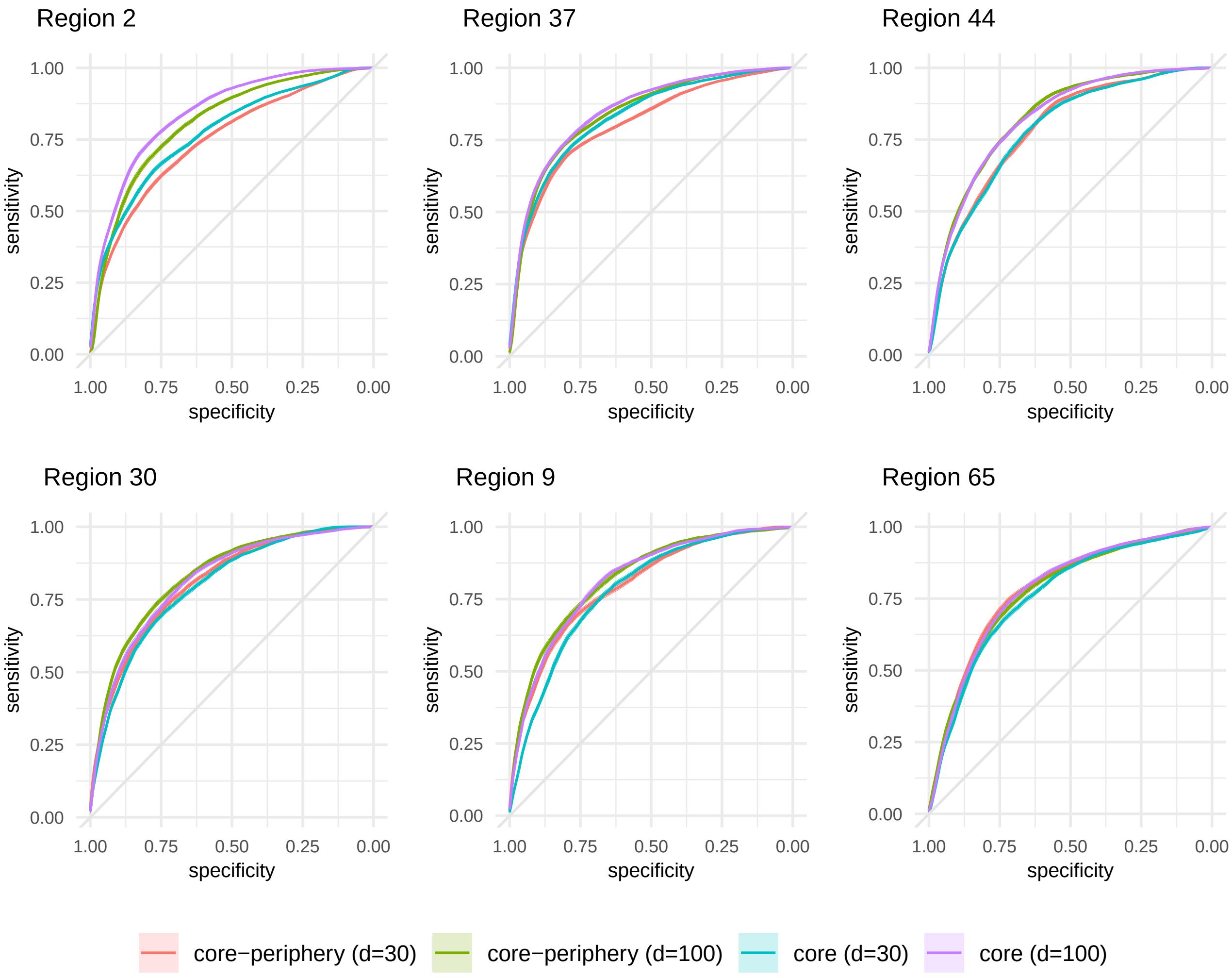

Figure 5: Classification of tissue types from a diffusion MRI scan. For each region, we train a classifier to predict the tissue type of each node using the core-periphery and core embeddings into dimensions $d = 30$ and $d = 100$. For 100 Monte-Carlo iterations, a random sample of 20% of the positions are held out for validation and the corresponding mean receiver-operating curves, and ± 2 standard errors are shown. The core-periphery embedding performs comparably to the core embedding.

A.12 Simulations of concrete examples

Seshadhri et al. (2020) note that for various real-world graphs, when looking only at the subgraphs containing nodes of maximum degree $c \in [10, 50]$, they get a value of $\Delta > 1$ (in many cases $\Delta > 10$). Hence, we choose values of σ_n and r_n (that are $\Theta(n)$) in our examples, which produce graphs of average degree in the same range. Figure 6 shows that setting $\sigma_n = n/20$ and $r_n = n/50$ gives average degrees in this range and average $\Delta_n > 10$ for varying values of n. The figure also illustrates that $n\rho_n$ and Δ_n stay (close to) constant as n varies.

To show that the values of Δ_n and $n\rho_n$ do not vary too greatly over different simulations of graphs from the same model, Figures 7 and 8 show 100 samples from each model, again with $\sigma_n = n/20$ and $r_n = n/50$. We see that the average values (red line) correspond well to the values seen for varying n in Figure 6.

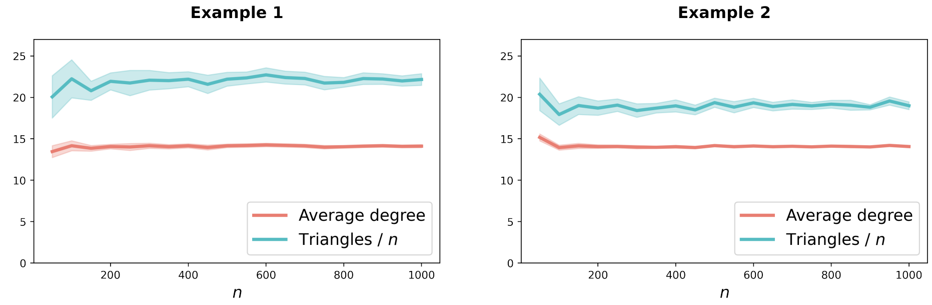

Figure 6: For each n, 20 graphs have been simulated from examples 1 and 2, with $\sigma_n = n/20$ and $r_n = n/50$ respectively, and the average degree ($n\rho_n$) and average triangle density (Δ_n) calculated. The confidence bands shown are ± three standard errors from the mean.

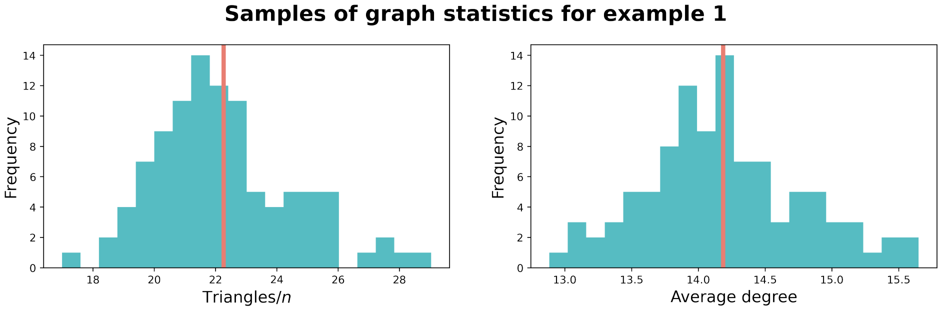

Figure 7: 100 graphs of size $n = 500$ have been generated from the model described in Example 1, with $\sigma_n = n/20$, and the average degree ($n\rho_n$) and triangle density (Δ_n) calculated. The vertical red line represents the mean of these samples.

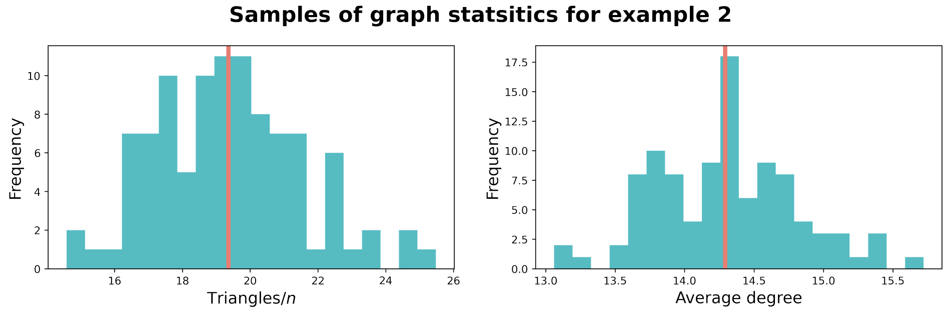

Figure 8: 100 graphs of size $n = 500$ have been generated from the model described in example 2, with $r_n = n/50$, and the average degree ($n\rho_n$) and triangle density (Δ_n) calculated. The vertical red line represents the mean of these samples.

A.13 Simulations of further examples

To illustrate that our results apply to models other than the two that we prove theoretically, in this section we present similar results from simulations of further models.

Example (3). Let $f(x, y) = \exp\{-\frac{1}{2}|x - y|^2\}$, be the radial basis function (RBF) and G_n be the uniform distribution on $[0, a_n]$.

We show through simulations that if we let $a_n = \Theta(n)$, then we get constant average degree and constant triangle density. This is illustrated in Figure 9 by simulating graphs from the model for increasing n with $a_n = n/10$.

Example (4). Let $f(x, y) = \exp\{-\frac{1}{2}||x - y||_2^2\}$, be the radial basis function (RBF) and G_n be the uniform distribution on $S^2(r_n)$.

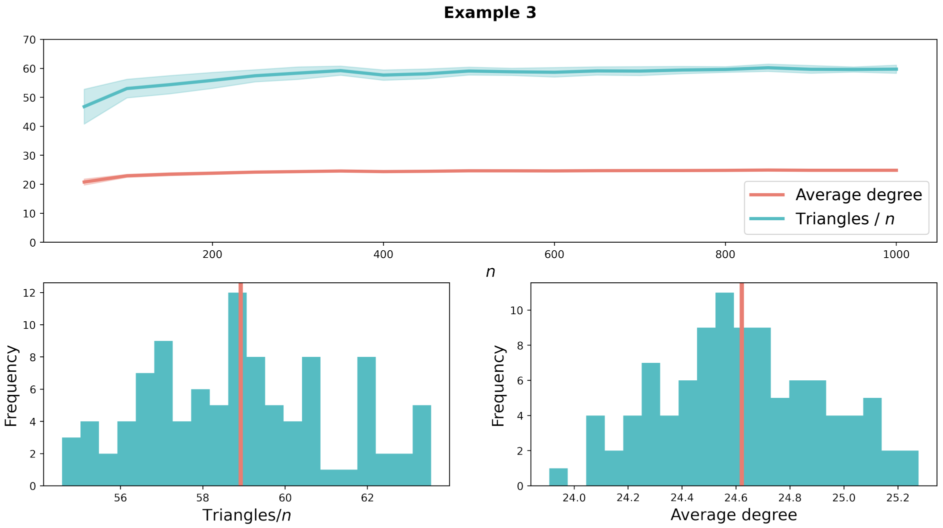

Figure 9: In the upper plot, for each n, 20 graphs have been simulated from the model described in example 3, with $a_n = n/10$, and the average degree $(n\rho_n)$ and triangle density (Δ_n) calculated. The confidence bounds shown are ± two standard deviations from the mean. In the lower plots, the same calculations have been done on 100 simulated graphs of size $n = 500$. The vertical red line represents the mean of these samples.

We show through simulations that if we let $r_n = \Theta(\sqrt{n}))$, then we get constant average degree and constant triangle density. This is illustrated in Figure 10 by simulating graphs from the model for increasing n with $r_n = \sqrt{n}/10$.

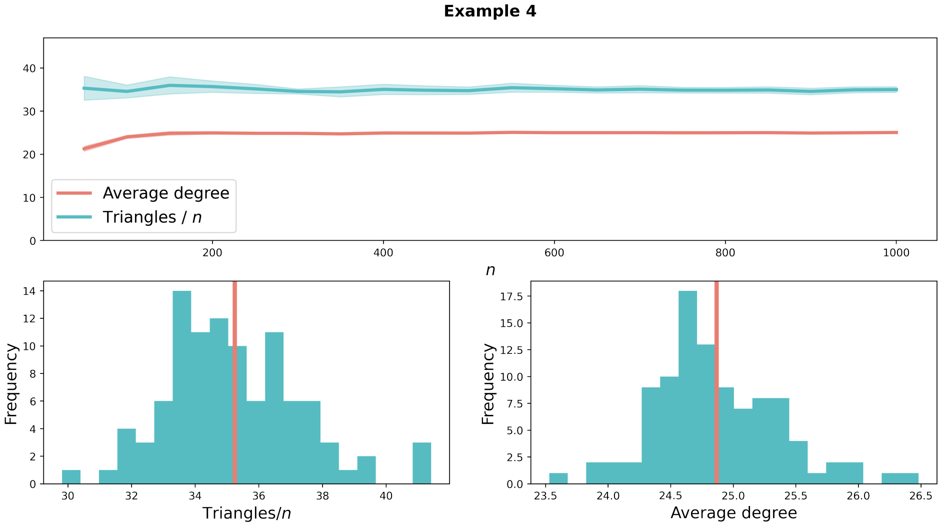

Figure 10: In the upper plot, for each n, 20 graphs have been simulated from the model described in example 3, with $r_n = \sqrt{n}/10$, and the average degree $(n\rho_n)$ and triangle density (Δ_n) calculated. The confidence bounds shown are ± two standard deviations from the mean. In the lower plots, the same calculations have been done on 100 simulated graphs of size $n = 500$. The vertical red line represents the mean of these samples.

A.14 Logistic model

Chanpuriya et al. (2020) propose a latent position model that they find is able to reconstruct sparse, triangle-dense graphs, where link probabilities are given by the logistic function,

$$f(x, y) = \frac{\exp(x^\top y)}{1 + \exp(x^\top y)}.$$

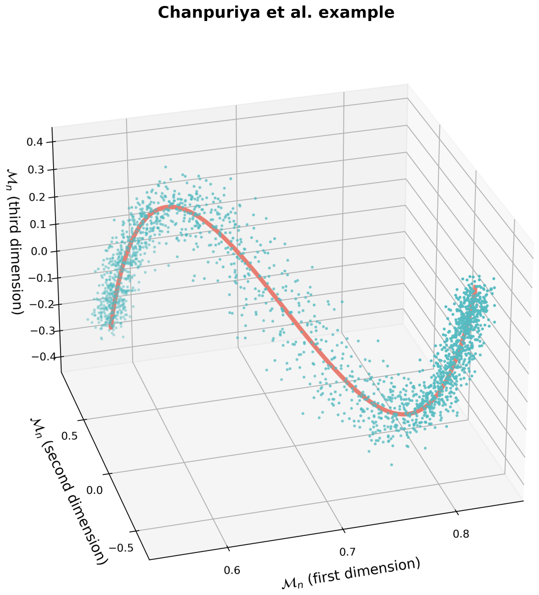

Figure 11: The first 3 dimensions of the manifold (red curve) implied by the logistic kernel with a uniform latent position distribution G_n on the interval $\mathcal{Z}_n = [-3, 3]$ are shown, along with the spectral embeddings (blue points) of a graph simulated from the model and aligned using Procrustes analysis.

As mentioned in the introduction, the effect of allowing two latent positions per node is complex, and not crucial for reproducing high triangle density. We therefore opt to present the manifold associated with the kernel above, rather than the asymmetric version in Chanpuriya et al. (2020). Figure 11 shows the first three dimensions of the manifold implicit in this kernel with a uniform latent position distribution (red curve) and the (Procrustes aligned) spectral embedding of a graph generated from it (blue points). The kernel is distortive (as one might expect, given it was never chosen not to be) and so the intrinsic (on-manifold) distribution F_n is not uniform.

A.15 Neighbourhood approximations

Here we will demonstrate that the 'number of shared neighbours' metric forms a reasonable surrogate for the theoretical neighbourhood described in Section 3 (local embedding). The approximate neighbourhood around an arbitrary query node, i, is found by taking the nodes with most neighbours in common with i.

To investigate this hypothesis, we simulated a number of graphs from the model in Example (3) (Appendix A.13), with $n = 500$ and $a_n = n/10$ so that the levels of sparsity and triangle density are in the range of real-world networks. As well as the shared neighbours metric, we also find an approximate neighbourhood based on the graph distance between nodes, i.e. the number of edges in the shortest path between nodes. Figure 12 shows the Jaccard similarity between the approximate neighbourhoods and the theoretical neighbourhood (nodes with latent positions closest to the latent position of the query node), where the Jaccard similarity between two sets A and B is defined as:

$$J(A, B) = \frac{|A \cap B|}{|A \cup B|}.$$

We see that the shared neighbours metric gives reasonable approximations to the theoretical neighbourhood for neighbourhoods of size between 5-15% of the size of the full graph.

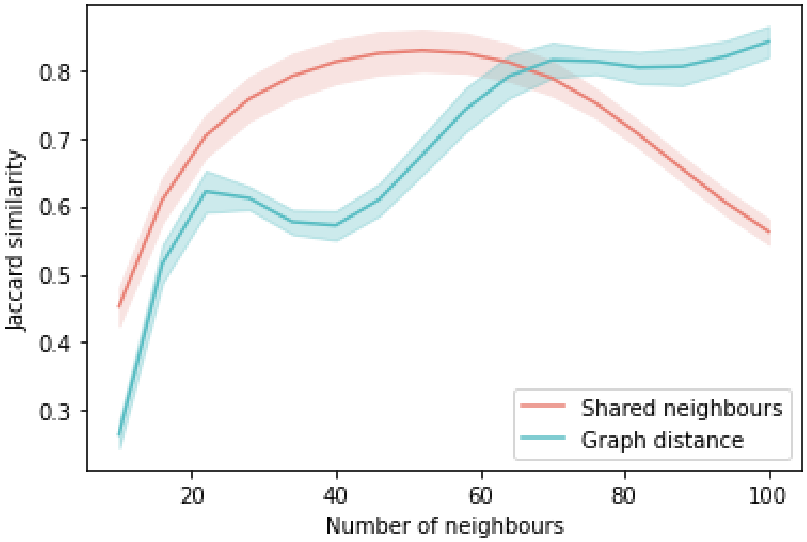

Figure 12: y-axis: Jaccard similarity between theoretical and approximate neighbourhoods found using two metrics: number of shared neighbours and graph distance. x-axis: Size of the neighbourhoods. The mean Jaccard similarity of 20 graphs simulated from the model in example (3), with $n = 500$ and $a_n = n/10$, is shown for each metric, with ± 2 standard deviation error bars. For each of the simulated graphs, we take the mean similarity over the neighbourhoods found for each node in the graph. The shared neighbours metric gives a mean Jaccard similarity greater than 0.7 for neighbourhoods sized between 4.4% and 16.4% of the size of the graph, and outperforms the graph distance metric for neighbourhoods of size less than 12.8% the size of the graph.