Graph Pseudometrics from a Topological Point of View

Ana Lucia Garcia-Pulido, Kathryn Hess, Jane Tan, Katharine Turner, Bei Wang, Naya Yerolemou

Abstract We explore pseudometrics for directed graphs in order to better understand their topological properties. The directed flag complex associated to a directed graph provides a useful bridge between network science and topology. Indeed, it has often been observed that phenomena exhibited by real-world networks reflect the topology of their flag complexes, as measured, for example, by Betti numbers or simplex counts. As it is often computationally expensive (or even unfeasible) to determine such topological features exactly, it would be extremely valuable to have pseudometrics on the set of directed graphs that can both detect the topological differences and be computed efficiently. To facilitate work in this direction, we introduce methods to measure how well a graph pseudometric captures the topology of a directed graph. We then use these methods to evaluate some well-established pseudometrics, using test data drawn from several families of random graphs.

A. L. García-Pulido
Department Of Computer Science, University of Liverpool, Liverpool L69 3BX, UK,
e-mail: a.l.garcia-pulido@liverpool.ac.uk

K. Hess
Laboratory for Topology and Neuroscience, Brain Mind Institute, Ecole polytechnique fédérale de Lausanne (EPFL), 1015 Lausanne, Switzerland, e-mail: kathryn.hess@epfl.ch

J. Tan, N. Yerolemou
Mathematical Institute, University of Oxford, Oxford OX2 6GG, UK, e-mail: jane.tan@maths.ox.ac.uk, naya.yerolemou@maths.ox.ac.uk

K. Turner
Mathematical Sciences Institute, Australian National University, Canberra, ACT 2601, Australia, e-mail: katharine.turner@anu.edu.au

B. Wang
School of Computing, University of Utah, Salt Lake City, UT 84112, USA, e-mail: beiwang@sci.utah.edu

1 Introduction

A typical strategy for studying complex networks is to extract features (i.e., parameters, properties, etc.) of the networks that are simpler to analyze and compare than the networks themselves, yet still capture their essential structures. Depending on the context, these features can be local or global and include, for instance, the number of connected components, the number of cycles, the lengths of shortest paths/cycles, graphlet counts, and the graph spectrum, as well as many non-numerical properties.

One method for extracting properties of a directed network begins with the construction of its associated directed flag complex, which is a topological space built from the directed cliques in the graph. Topological features of the directed flag complex provide revealing properties of the original network. For instance, the *homology groups* of the directed flag complex reflect how cliques assemble to form the graph globally, enabling us to distinguish between some graphs: if the homology groups of two graphs are different, then the two graphs are not isomorphic. From the homology groups one can extract *Betti numbers*, which measure the rank of the homology groups. While the 0-th Betti number is simply the number of connected components of the graph, higher Betti numbers measure how intricately higher dimensional cliques intersect. In contrast to classical graph parameters, these invariants capture higher order structures.

Applications of topological methods to the analysis of networks have been motivated, in particular, by the desire to understand the relation between function and structure of biological networks (see e.g. [8, 9, 18, 24, 28, 29]). The work in these articles strongly suggests that biological function reflects topological structure, as measured by Betti numbers for example. A significant barrier to confirming and exploiting this observation, however, is the computational complexity of determining these features for real-world networks. For instance, directly comparing the homology groups or the Betti numbers of two flag complexes is an NP-hard problem [1].

In this paper, we explore the hypothesis that there are pseudometrics on the set of directed graphs that detect differences in their topological features. As a first step to identifying or constructing such a pseudometric, we introduce methods to measure how well a given pseudometric reflects differences in the topological features of two graphs.

We implement our methods of comparison and apply them to existing candidate pseudometrics on the set of directed graphs. There are already many high-performing pseudometrics on the space of graphs that take into account 1-dimensional structural features (see, e.g., [14, 21, 31]), but there is a dearth of literature on whether they also capture any high-dimensional structure. We note that these methods can also be applied to undirected graphs by considering instead the usual flag complex. However, the analysis in the present paper is restricted to the directed case, since many real-world networks (particularly biological networks) are naturally directed. Our test data set is drawn from several families of directed random graphs for which we can control the parameters and behavior.

We present here an experimental study, comparing topological pseudometrics based on Betti numbers and simplex counts with several well-established pseudometrics for directed graphs. Our analysis is four-fold. First, we study the similarities between clusterings based on these pseudometrics by computing both their Fowlkes-Mallows indices and the distance correlations between them. Almost all of the pseudometrics tested are shown to be closely related with high Fowlkes-Mallows indices and high distance correlations. Second, we apply k-nearest neighbors (k-NN) classification as a measure of classification accuracy for three models of random graphs. We find that all pseudometrics achieve near perfect classification accuracy. Third, we test for relationships between each pseudometric and the various random graph parameters by performing permutation

tests with distance correlation and the Fowlkes-Mallows indices. Using the permutation tests, we can reject the null hypothesis of independence between all the different pseudometrics when considered over pooled sets of directed graphs with multiple parameter values. However, we cannot in general reject the null hypothesis of independence once we restrict to a specific model and parameter value. This indicates that the latent variable of the parameter of the model is important. Finally, we apply k-NN regression to our pseudometrics to try to predict the topological feature vectors of given graphs.

Outline We begin by reviewing the requisite topological and combinatorial background in Section 2. This includes the definitions of our topological feature vectors together with the topological pseudometrics that they induce, as well as a brief introduction to each of the existing pseudometrics that we compare to our topology-based ones: TriadEuclid, TriadEMD, and Portrait Divergence. We present two methods for comparing pseudometrics in Section 3, one based on clustering and the other on distance correlation. Here, we also describe the permutation test and k-NN regression, as well as the classification methods that we use. Technical details of our experiments can be found in Section 4. The final results of the comparison are presented in Section 5.

2 Directed Graph Pseudometrics

A *directed graph* G consists of a set of *vertices* V together with a set of *edges* E, which are ordered pairs of vertices. All graphs in this paper are *finite*, meaning V and E are both finite sets. The direction of an edge $(u, v) \in E$ is taken to be from u, the *origin*, to v, the *destination*. We require that G not contain any self-loops, that is, for each $(u, v) \in E$, $u \neq v$. Secondly, for each pair of vertices $(u, v) \in V \times V$, there is at most one directed edge from u to v. Note, however, that we do allow both $(u, v) \in E$ and $(v, u) \in E$. In other words, the directed graphs we consider are simple except for bigons, and we shall simply refer to them as *graphs* or *digraphs*.

We use several standard definitions associated with digraphs. The *out-degree* of a vertex v is the number of edges having v as the origin, while its *in-degree* is the number of edges having v as the destination. A vertex v is a *sink* if its out-degree is zero and its in-degree is at least one and a *source* if its in-degree is zero and its out-degree is at least one. A *path* in G is a list of distinct edges such that the destination of the i-th edge is the same as the origin of the $(i + 1)$-st edge and such that no vertex is traversed more than once, except the path may end at the vertex where it started. If the first and last vertices of a path are the same, we call it a *cycle*.

Let $\mathcal{G}$ denote the set of all finite directed graphs. We will view $\mathcal{G}$ as a space endowed with several natural *pseudometrics*. First, recall that a pseudometric on a set $\mathcal{X}$ is a function $d_{\mathcal{X}} \colon \mathcal{X} \times \mathcal{X} \to [0, \infty)$ such that for all $x, y, z \in X$,

1. $d_{\mathcal{X}}(x, x) = 0$;
2. $d_{\mathcal{X}}(x, y) = d_{\mathcal{X}}(y, x)$ (symmetry); and
3. $d_{\mathcal{X}}(x, z) \leq d_{\mathcal{X}}(x, y) + d_{\mathcal{X}}(y, z)$ (triangle inequality).

Importantly, points need not all be distinguishable by a pseudometric: it is possible that $x \neq y$, even though $d_{\mathcal{X}}(x, y) = 0$. The pair $(\mathcal{X}, d_{\mathcal{X}})$ is a *pseudometric space*.

We will describe three well-established pseudometrics on $\mathcal{G}$: TriadEuclid in Section 2.2, TriadEMD in Section 2.3, and portrait divergence in Section 2.4. Before this though, we define two topological summaries for elements in $\mathcal{G}$ based on Betti numbers and simplex counts in Section 2.1, which provide a crucial point of comparison.

2.1 Betti Numbers and Simplex Counts

The key construction we study in this paper is the directed flag complex of a directed graph. For a more detailed account we refer to the work of Luetgehetmann *et. al.* [16] and Reimann *et. al.* [24]. We assume familiarity with standard notions of abstract simplicial complexes and simplicial homology, introduced for instance in [11, 19].

Definition 1 (Abstract Directed Simplicial Complex) An *abstract directed simplicial complex* on a vertex set V is a collection $\mathcal{K}$ of lists (i.e., totally ordered sets) of elements of V such that for every sequence $\sigma \in \mathcal{K}$, every subsequence τ of σ belongs to $\mathcal{K}$.

An element $\sigma \in \mathcal{K}$ is called a *(directed) simplex*. If σ is of length $p+1$, then we call it an *p-simplex*. The collection of p-simplices of $\mathcal{K}$ is denoted $\mathcal{K}_p$. If $\sigma \in \mathcal{K}$ and $\tau \subset \sigma$, then τ is called a *face* of σ. The i-th face of an p-simplex $\sigma = (v_0, \ldots, v_p)$ is the $(p-1)$-simplex obtained by removing the v_i from the list σ.

The notion of an abstract directed simplicial complex is a variant of the more common notion of abstract simplicial complex. Henceforth, we always mean abstract directed simplicial complexes when we say *simplicial complexes*. The following definition illustrates how directed simplicial complexes arise naturally from directed graphs.

Definition 2 (Directed Flag Complex) Given a directed graph $G = (V, E)$, the *directed flag complex* of G, denoted $\mathcal{K} = \mathcal{K}(G)$, is defined as follows.

- Take $\mathcal{K}_0 = V$.
- For $p \geq 1$, a *directed p-simplex* σ in $\mathcal{K}_n$ is an $(p+1)$-tuple of vertices $(v_0, \ldots, v_p)$ such that there is a directed edge from v_i to v_j for every pair of vertices v_i, v_j with $0 \leq i < j \leq p$.

For $\sigma = (v_0, \ldots, v_p) \in K$, we call v_0 the *source* of the simplex, since there is a directed edge from v_0 to v_i for every $i > 0$. Similarly, we call v_p the *sink* of the simplex, since there is a directed edge from v_i to v_p for every $i < p$. Note that these are consistent with the equivalent digraph notions since a p-simplex is characterised by the *ordered* sequence of vertices and not by the underlying set of vertices.

Throughout this paper, we consider simplicial homology of directed flag complexes with $\mathbb{F}_2$-coefficients, where homology is defined in the usual way. For a simplicial complex $\mathcal{K}$, let $H_p(\mathcal{K})$ denote its *p-th homology group* and $\beta_p(\mathcal{K}) = \dim H_p(\mathcal{K})$ its *p-th Betti number*.

Given $G \in \mathcal{G}$, we work with two simple topology-based feature vectors on G, defined as follows. Let $p \in \mathbb{Z}_{\geq 0}$ denote the maximum homological dimension of interest, which is context-dependent; in our experiments, $p = 6$. For $0 \leq k \leq p$, let $b_k(G) = \max\{0, \log(\beta_k(\mathcal{K}(G)))\}$ and

$$b(G) = (b_0(G), \ldots, b_p(G)).$$

Let $\gamma_k(G)$ be the number of directed k-simplices of $\mathcal{K}(G)$, $c_k(G) = \max\{0, \log(\gamma_k(\mathcal{K}(G)))\}$, and

$$c(G) = (c_0(G), \ldots, c_p(G)).$$

The vectors $b(G)$ and $c(G)$ consist of the logarithms of the Betti numbers and simplex counts of $\mathcal{K}(G)$, respectively. Let $\|\cdot\|_2$ denote the Euclidean norm on $\mathbb{R}^n$.

Definition 3 (Topological Pseudometrics) The pseudometrics d_β and d_Δ on $\mathcal{G}$ are specified by

$$d_\beta(G, G') = \|b(G) - b(G')\|_2; \tag{1}$$
$$d_\Delta(G, G') = \|c(G) - c(G')\|_2. \tag{2}$$

for any pair of directed graphs $G, G' \in \mathcal{G}$.

2.2 TriadEuclid

First introduced by Przulj *et. al.* [23], the term *graphlet* is often used to mean a small connected graph up to a fixed size. A *graphlet-based* pseudometric compares the graphlet counts in pairs of graphs. Graphlet-based pseudometrics are a well established tool in network analysis, as they capture local structural similarity of two graphs.

Xu and Reinert [35] introduced two graphlet-based pseudometrics that outperform the best previously defined directed graphlet methods in graph classification tasks: *TriadEuclid* and *TriadEMD*. Both of these consider only directed graphlets on three vertices. In this section we focus on TriadEuclid, which measures the difference between 3-graphlet counts of two directed graphs in terms of their Euclidean distance. We remark that there are exactly 13 isomorphism classes of connected directed graphlets with 3 vertices (referred to as 3-graphlets). A complete list can be found in [35].

Let $G = (V, E)$ a directed graph. If the induced subgraph determined by three vertices of G is connected, it must be isomorphic to one of the 13 3-graphlets. Let $n_i(G)$ be the number of induced subgraphs of G that are isomorphic to the i-th graphlet, $i \in I = \{1, \ldots, 13\}$. Define $\phi(G) \in \mathbb{N}^{13}$ by

$$\phi(G)_j = \frac{n_j(G)}{\sum_{i \in I} n_i(G)} \tag{3}$$

for $j \in I$.

Definition 4 The *TriadEuclid* pseudometric on the set $\mathcal{G}$ of finite directed graphs is defined by

$$\text{TriadEuclid}(G, G') = \|\phi(G) - \phi(G')\|_2, \tag{4}$$

for all $G, H \in \mathcal{G}$, where $\|\cdot\|_2$ denotes the Euclidean norm.

The complexity of this algorithm is $O(nd^2)$, where d is the maximum degree of vertices in G and G', n is the maximum number of vertices in G and G'.

2.3 TriadEMD

TriadEMD is another graphlet-based pseudometric defined by Xu and Reinert [35], computed in terms of the earth mover distance between generalized degree distributions of two directed graphs.

Definition 5 Given a directed graph $H = (V, E)$, an *automorphism* of H is a graph isomorphism $h\colon H \to H$. The set of automorphisms of H forms a group under composition, denoted $\text{Aut}(H)$, which acts on H.

Let $v \in V$. The *orbit* of v under the action of $\mathrm{Aut}(H)$ is defined by

$$\mathrm{Orb}(v) = \{h(v) | h \in \mathrm{Aut}(H)\}. \tag{5}$$

Projecting to orbits, one considers the genuinely different positions that a vertex takes in a fixed graphlet. Figure 1 shows an example of a three-vertex graph with only two distinct orbits:

$$\mathrm{Orb}(1) = \{1\}, \quad \mathrm{Orb}(2) = \mathrm{Orb}(3) = \{2, 3\}.$$

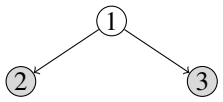

Fig. 1 A graph with two vertex orbits $\{1\}$ and $\{2, 3\}$.

A complete list of the 30 orbits of 3-graphlets can be found in [35]. A list of graphlets with between two and four vertices is given in [26].

Degree distributions can be used to study *hubs* in a network, i.e., vertices of high degree. The *degree distribution* P of an undirected graph is defined so that $P(k)$ is the proportion of vertices in G with degree k:

$$P(k) = \frac{\#\{v \in V | \deg(v) = k\}}{\#V}. \tag{6}$$

When G is directed, there are analogous in-degree and out-degree distributions.

TriadEMD is defined in terms of the degree distributions of orbits in triads. For any orbit i of a fixed graphlet, the *orbit-i-degree* of a vertex v of a digraph G is the number of copies of orbit i in G of which v is a vertex [35]. The *orbit-i-degree distribution* P_i of G is defined so that $P_i(k)$ is the proportion of vertices in G with orbit-i-degree k.

Consider, for example, the graph G on four vertices shown in Figure 2, for which we now calculate the Orb $(\{1\})$-degree distribution of the graphlet in Figure 1. The Orb $(\{1\})$-degrees of vertices 1, 2, 3, and 4 are 1, 3, 0, and 0 respectively; Table 1 gives the Orb $(\{1\})$-degree distribution.

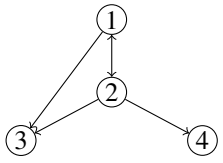

Fig. 2 Graph G with Orb $(\{1\})$-degree distribution given in Table 1.

Table 1 Orb $(\{1\})$-degree distribution of G Figure 2

k =degree	0	1	2	3
$P_{\{1\}}(k)$	0.5	0.25	0	0.25

The original definitions of the orbit degrees and orbit graphlet metrics for undirected and directed graphs, together with additional examples can be found in [22, 26, 34, 35].

Definition 6 Let Θ be the set of 30 orbits of 3-graphlets. The *TriadEMD* pseudometric on the set $\mathcal{G}$ of directed graphs is given by

$$\text{TriadEMD}(G, G') = \frac{1}{30} \sum_{i \in \Theta} \text{EMD}(P_i, P_i'), \tag{7}$$

where P_i (resp. P_i') denotes the orbit-i-degree distribution of G (resp. G'), and EMD denotes the earth mover distance (EMD) between the distributions. Recall that the EMD between two probability distributions P and Q on the real line is defined by

$$\text{EMD}(P, Q) = \int_{-\infty}^{\infty} |F_P(x) - F_Q(x)| dx, \tag{8}$$

where F_P and F_Q are the cumulative distribution functions of P and Q respectively.

The complexity of this algorithm is $O(n \log n)$, where n is the maximum between the numbers of vertices of G and of G'.

2.4 Portrait Divergence

First introduced by Bagrow and Bollt [2], *Network Portrait Divergence* is a pseudometric that compares the distributions of the shortest-path lengths of two graphs. Recall that the *diameter* of a finite directed graph is the maximum over all pairs of vertices of the length of the shortest (directed) path from the first vertex to the second.

Definition 7 Let $G = (V, E)$ be a directed graph of diameter d, with n vertices, and denote by $d_{sp}(v, u)$ the shortest path distance from v to u. For every vertex $v \in V$ and $0 \leq l \leq d$, define

$$s_v^l = \#\{u \in V | d_{sp}(v, u) = l\}. \tag{9}$$

The *network portrait* [3] of G is the matrix $B = (B_{lk})$ with entries

$$B_{lk} = \#\{v \in V | s_v^l = k\}, \tag{10}$$

where $0 \leq l \leq d$ and $1 \leq k \leq n - 1$. That is, B_{lk} is the number of vertices that are at distance l from exactly k other vertices.

The first row of the network portrait B of G is precisely the degree distribution of G. The second row is the degree distribution of next-nearest neighbors, and so on. The network portrait B also captures structural features of G such as the number of edges, the diameter of G, and the distribution of shortest paths. It has been shown to be a graph invariant [2]. The complexity of computing an unweighted portrait is $O(mN + N^2)$, due to the procedure of finding minimum-length paths, where m is the number of edges and N the number of vertices in G.

Let $G = (V, E)$ be a directed graph. Let $P(k, l)$ be the probability of choosing two vertices $(u, v) \in V \times V$ uniformly at random such that $d_{sp}(u, v) = l$ and $s_u^l = k$. If $s_u^l = k$, then there exist k vertices $u_1, \ldots, u_k$ with $d_{sp}(u, u_i) = l$. Therefore,

$$\begin{aligned}\#\{(u,v) \in V \times V : d_{sp}(u,v) = l, s_u^l = k\} &= k \cdot \#\{v \in V : s_v^l = k\} \\ &= k \cdot B_{lk},\end{aligned}$$

and in consequence,

$$P(k,l) = \frac{kB_{lk}}{n^2}. \tag{11}$$

Given graphs G and G', we define (joint) distributions P and P' for all rows of their portraits B and B'. The KL-divergence between P and P' is then defined by

$$KL(P||P') = \sum_{l=0}^{\max(d,d')} \sum_{k=0}^{N} P(k,l) \log \frac{P(k,l)}{P'(k,l)}. \tag{12}$$

Definition 8 The *Portrait Divergence* (PD) pseudometric on the set $\mathcal{G}$ of finite directed graphs is the Jensen-Shannon divergence,

$$PD(G,G') = \frac{1}{2}KL(P||M) + \frac{1}{2}KL(P'||M), \tag{13}$$

for all $G, G' \in \mathcal{G}$, where $M = (P + P')/2$ is the mixture distribution of P and P'.

3 Statistical Tools

We are interested in comparing known pseudometrics on $\mathcal{G}$ (e.g., TriadEuclid, TriadEMD, and PD) to d_β and d_Δ, which requires some care. A natural first idea would be to use Gromov-Hausdorff distance, but since it measures how close two spaces are to being isometric, it is too rigid for practical purposes. For example, changing even just one Betti number of a single graph can result in a very large Gromov-Hausdorff distance. Multiplying a topological pseudometric by a constant has a similar effect as well.

Instead, we employ two comparison methods that are invariant under rescaling either pseudometric and more robust under perturbations of points. The first method is based on the distance correlation between pseudometrics (Section 3.1), while the second is based on the Fowlkes-Mallows index of the clusterings of the two pseudometrics (Section 3.2). Both rely on taking finite samples $\mathcal{S} \subset \mathcal{G}$ and performing an analysis on $\mathcal{S}$.

We remark that similar comparison methods can also be applied to undirected graphs by replacing the directed flag complex with the flag complex, though we restrict to the directed case in this paper.

3.1 Distance Correlation

The concept of distance correlation was first introduced in by Szekely *et. al.*[30] for two paired random vectors in Euclidean space and generalized to metric spaces by Lyons [17]. Distance correlation measures linear and non linear relationships between two distributions lying in possibly

different metric spaces. We follow the definition of the sample distance correlation of a paired sample as formulated by Turner and Spreemann [33]. We use the sample distance correlation as an estimation of the distance correlation and refer to the sample distance correlation simply as the distance correlation.

Let $(\mathcal{X}, d_{\mathcal{X}})$ and $(\mathcal{Y}, d_{\mathcal{Y}})$ be pseudometric spaces, and let $(X, Y) = \{(x_i, y_i)\}_{1 \leq i \leq l} \subset \mathcal{X} \times \mathcal{Y}$ be paired samples. For $1 \leq i, j \leq l$, let $a_{i,j} = d_{\mathcal{X}}(x_i, x_j)$ and $b_{i,j} = d_{\mathcal{Y}}(y_i, y_j)$, so that $a = (a_{i,j})$ and $b = (b_{i,j})$ denote matrices of pairwise distances in $\mathcal{X}$, and $\mathcal{Y}$, respectively. Let $\overline{a}^i$ and $\overline{b}^i$ denote the row means and $\overline{a}_j$ and $\overline{b}_j$ the column means of the matrices a and b. Let $\overline{a}$ and $\overline{b}$ denote the total matrix means. Define *doubly centred matrices* $(A_{k,l})$ and $(B_{k,l})$ by $A_{k,l} = a_{k,l} - \overline{a}^k - \overline{a}_l + \overline{a}$ and $B_{k,l} = b_{k,l} - \overline{b}^k - \overline{b}_l + \overline{b}$.

Definition 9 The *sample distance covariance* of the paired sample (X, Y) is

$$\operatorname{dcov}(X, Y) = \frac{1}{n^2} \sum_{k,l=1}^{n} A_{k,l} B_{k,l}. \tag{14}$$

When $\operatorname{dcov}(X, Y) \geq 0$, let $\operatorname{dCov}(X, Y) = (\operatorname{dcov}(X, Y))^{1/2}$.

The *sample variance* of the sample X is defined to be

$$\operatorname{dVar}(X) = \left(\frac{1}{n^2} \sum_{k,l=1}^{n} A_{k,l}^2 \right)^{1/2}. \tag{15}$$

If $\operatorname{dVar}(X) \operatorname{dVar}(Y) \neq 0$ and $\operatorname{dcov}(X, Y) \geq 0$, the *sample distance correlation* is given by

$$\operatorname{dCor}(X, Y) = \frac{\operatorname{dCov}(X, Y)}{(\operatorname{dVar}(X) \operatorname{dVar}(Y))^{1/2}}. \tag{16}$$

If $\operatorname{dVar}(X) \operatorname{dVar}(Y) = 0$, then we set $\operatorname{dCor}(X, Y) = 0$.

For strongly negative metrics spaces, which includes Euclidean space, the sample distance covariance between X and Y is always non-negative, and X and Y are independent if and only if $\operatorname{dCov}(X, Y) = 0$, see [17]. This is not true for all metric spaces, but in all of our cases the sample covariance is non-negative, and therefore the sample distance correlation is well-defined.

Notice that sample distance correlation is invariant under scalar multiplication of either of the two metrics. Cauchy-Schwarz implies that $|\operatorname{dCor}(X, Y)| \leq 1$ and that equality is attained when the doubly centred matrices are scalar multiples of each other. Thus, high correlation measures the (possibly non-linear) relationship between X and Y arising from a linear relationship between their corresponding metrics.

In our case we have $\mathcal{Y} = \mathcal{X}$, and we consider the sample distance correlation on paired samples (X, X).

3.2 Fowlkes-Mallows Index

Our second method of comparing pseudometric spaces is based on clustering. Suppose we have pseudometric spaces $(\mathcal{X}, d_0)$ and $(\mathcal{X}, d)$. Given a finite sample $X \subset \mathcal{X}$, we can compute hierarchical

clusterings A_0 and A corresponding to both pseudometrics and hence apply standard techniques of cluster evaluation. In particular, we use the Fowlkes-Mallows index [7], which provides a measure of similarity between two clusterings. Making such comparisons across many choices of sample X provides a means of comparing $(\mathcal{X}, d)$ to $(\mathcal{X}, d_0)$.

We employ a standard agglomerative hierarchical clustering algorithm (see documentation of [12] for details), initialised with every point in a separate cluster. In each step, two clusters of minimum distance to each other are chosen and merged into one to move up the hierarchy. We use complete linkage, so the distance between clusters is the maximum distance between any two points of those clusters. The algorithm terminates when all points are in a single cluster.

In some applications, the number of expected clusters is known from the outset, allowing early termination of the algorithm. As this does not apply to our case, we first run the clustering algorithm and then use *silhouette analysis* to choose the most natural number of clusters with respect to d_0. This method, introduced by Rousseeuw [25], assesses how well a clustering captures the structure of the data using only internal distance information. Informally, we judge how well a single point x has been placed within a cluster based on a *silhouette value*, which quantifies how close x is to the other points in its own cluster in contrast to points in other clusters, and then extend to a numerical score for the clustering.

Definition 10 Suppose that we have points in a pseudometric space $(\mathcal{X}, d)$ partitioned into clusters $C_1, \ldots, C_k$. Given a point $u \in C_i$, let $a(u)$ be the mean distance between u and other points in C_i, that is

$$a(u) = \frac{1}{|C_i| - 1} \sum_{v \neq u, v \in C_i} d(u, v). \tag{17}$$

For every $j \neq i$, let $a_j(u)$ be the mean distance between u and points in cluster C_j,

$$a_j(u) = \frac{1}{|C_j|} \sum_{v \in C_j} d(u, v). \tag{18}$$

Set $b(u) = \min_{j \neq i} a_j(u)$. If $|C_i| > 1$, define the *silhouette value* of u by

$$s(u) = \frac{b(u) - a(u)}{\max\{a(u), b(u)\}}, \tag{19}$$

and set $s(u) = 0$ if $|C_i| = 1$.

It follows immediately from the definition that $s(u) \in [-1, 1]$ and that

- $s(u) \approx 1$ indicates that u is appropriately clustered;
- $s(u) \approx -1$ means that u is closer to points from a different cluster,
- if $s(u) \approx 0$, then u is almost equidistant between C_i and at least one other cluster C_j.

Based on silhouette values of the points in each of the clusters, one can define a score for the entire clustering.

Definition 11 The *silhouette coefficient* of the clustering $C_1, \ldots, C_k$ of l points $u_1, .., u_l$ is their average silhouette value:

$$SC = \frac{1}{l} \sum_{1 \leq i \leq l} s(u_i). \tag{20}$$

The definition above enables us to determine if the clustering $C_1, \ldots, C_k$ reflects a natural structure present in our samples or if instead it seems forced.

The actual comparative element in our method based on the silhouette coefficient comes from the use of the Fowlkes-Mallows index [7], which measures either the similarity between two clusterings with k clusters or the similarity between one clustering and a benchmark classification. We use the former, but both proceed by considering whether pairs of points are consistently assigned to clusters or classified. Specifically, let X be a set with n elements, and let A and A' be two clusterings of X with k clusters each. Say that a pair of points are *together* in a clustering if and only if they are assigned to the same cluster. Let TP (true positives) be the number of pairs of points that are together in both A and A', FP (false positives) the number of pairs together in A but not in A', FN (false negatives) the number of pairs together in A' but not in A, and TN (true negatives) the number of pairs that are not together in either A or A'.

Definition 12 The *Fowlkes-Mallows* index FM_k of the pair A, A' (for clusterings with k clusters) is given by

$$FM_k := \sqrt{\frac{TP}{TP+FP} \cdot \frac{TP}{TP+FN}}. \tag{21}$$

Notice that $FM_k \in [0,1]$ and that the clusterings A and A' are identical when $FM_k = 1$ and the most dissimilar when $FM_k = 0$.

In our experiments, given a directed graph, we apply silhouette analysis to determine the number of clusters k that gives the highest silhouette score for d_β (respectively, d_Δ). We then compare a known pseudometric d (triadEuclid, triadEMD, and PD) to d_β (respectively, d_Δ) in terms of the Fowlkes-Mallows index of their associated clusterings.

Remark A particular challenge in comparing clusterings is the absence of benchmark datasets. It is also difficult to determine the optimal number of clusters for a given dataset. In order to circumvent this issue, we use silhouette analysis to determine the most appropriate number of clusters. In our experiments, we obtain a unique number of clusters that realizes the maximal silhouette coefficient; however, this is not guaranteed in general. In the case where multiple numbers of clusters attain the same maximum silhouette score, one could easily adapt the present method by computing Fowlkes-Mallows indices for every viable number of clusters and then taking, for instance, the mean score.

3.3 Permutation Tests for Paired Data

The permutation test has become a default method for testing independence. It is a provably valid and consistent test for any consistent dependency measure, such as distance correlation. For more details concerning permutation tests with distance correlation we refer to [27].

A permutation test for paired data involves computing some summary statistic (in our case, either distance correlation or the Fowlkes-Mallows index) on the original pairing and then comparing this summary statistic to that of a shuffled pairing. If the two pseudometrics are truly independent, then the percentile of the summary statistic amongst all those with permuted pairings is drawn uniformly over possible percentiles.

In particular, we can test for a null-hypothesis of independence for two pseudometrics with a permutation test where the sampled *p-value* is the percentile of the original distance correlation (resp. Fowlkes-Mallows index) among those for the permuted pairings.

Since we compute many p-values for comparing various pairs of pseudometrics over many different sample sets of directed graphs, we must correct for multiple hypothesis testing. The *family-wise error rate* is the probability of making one or more false discoveries when performing multiple hypotheses tests. The simplest, most conservative method to bound the family-wise error rate is the Bonferroni correction, which can be applied to any collection of hypothesis tests, regardless of any dependency structure among the variables. In the Bonferroni correction for N tests, we simply divide the target significance level by N; to bound the family-wise error rate by α we use α/N as the threshold for rejecting the null hypothesis. In order to apply the Bonferroni correction, we consider in our analyses each table of p-values as a family.

To consider all the permutation tests in all of the tables of this paper simultaneously, we must instead bound the *false discovery rate*: the number of discoveries (tests where we reject the null hypothesis) that may be false positives as a proportion of all discoveries made, including both true and false discoveries. To affirm that the expected number of discoveries that are false is at most α, we apply the conservative approach by Benjamini and Yekutieli [4] formulated below, since we do not know what dependency structures there may be among the variables.

Theorem 1 *[4] Let* $C(N) = \sum_{j=1}^{N} \frac{1}{j}$. *Let* $p_{(1)} \leq p_{(2)} \leq \ldots \leq p_{(N)}$ *be the ordered observed p-values. Define*

$$k := \max \left\{ i : p_{(i)} \leq \frac{i}{mC(N)} \alpha \right\}. \tag{22}$$

If no such i exists, reject no hypothesis. If an i exists, and we reject the null hypotheses $H_{(1)}, H_{(2)} \ldots H_{(k)}$*, then we have controlled the false discovery rate at a level less than or equal to* α.

3.4 k-NN Classification and Regression

Given a function f on a data set and a set of training data for which the function value is known, one intuitive approach to estimating f on a test point is to extrapolate from the function values on the training points closest to it. This is the method of *k-nearest neighbors* (k-NN). There are two slightly different algorithms for this method, depending on whether the unknown function is categorical or real-/vector-valued.

For a categorical functions, one applies *k-NN classification*. For a given test point, one considers the k nearest training points and assigns to the test point the category that appears the most frequently among them. We measure success in terms of *classification rate*, i.e., the proportion of classifications that are correct.

For real- or vector-valued functions, *k-NN regression* is more appropriate. For a given test point, one considers the k nearest training points and assigns to the test point the average of their function values. The difference between the function estimate and the true function value is called the *residual*. The goal is for the residuals to be small as possible. We therefore measure success via the *mean squared error* (MSE), i.e., the mean of the squares of the errors/residuals.

There are many ways to split data into training and testing sets. In this paper, we use leave-one-out, a form of cross-validation, where at each stage we pick a single data point as the testing set and use everything else as the training set. We then repeat the process so that every data point is considered as the test sample at some point.

4 Experimental Setup

In this section, we provide the technical and implementation details of our experimental setup. First, we describe the random graph models used to generate our test graphs and the associated parameter selection. We then compute topological pseudometrics d_β and d_Δ for the collections of test graphs in order to perform a comparative study.

4.1 Random Graph Models and Parameters

In our experiments, we analyze directed graphs from families based on three random graph models: directed Erdös–Rényi random graphs (ER), directed geometric random graphs (GR), and random k-out graphs with preferential attachment (PA). There are numerous standard references for these models (see, for instance, [20]). We focused on these three random models to initialise the study of pseudometrics from a topological perspective. In particular, GR and PA are used to model real world networks, such as mobile ad hoc networks, the World Wide Web, and social networks. We include a brief description with the purpose of describing the specifications used to generate our test data.

Random graph models A directed ER random graph is generated by starting with a fixed set of n vertices and adding a directed edge between each ordered pair of vertices independently with probability ρ. Note that the edges $u \to v$ and $v \to u$ are also chosen independently of each other, and in particular it is possible for both to be present.

A classic geometric random graph is generated by placing vertices uniformly at random in the unit square, and then adding an edge between two vertices whenever the (Euclidean) distance between the vertices is at most equal to a fixed parameter r. We consider (biased) oriented GR random graphs obtained by taking an undirected graph generated in the classical sense with vertex set $\{1, \ldots, n\}$, and then for each edge uv (with $u < v$) choosing the direction $u \to v$ with probability 1/3 and $v \to u$ with probability 2/3. The directions are chosen independently for each edge in the undirected graph, but in this collection it is not possible to have both directed edges between a single pair of vertices.

A PA random graph with parameter k is generated as follows: give each vertex an initial weight of 1, and select a vertex u with out-degree less than k, uniformly at random. Choose another vertex v with probability proportional to its weight, add a directed edge from u to v and increase the weight of v by 1. This process terminates when every vertex has out-degree k. The initial output may have parallel (repeated) directed edges, which we subsequently reduce to a single directed edge leaving at most one edge in each direction between any pair of vertices.

Test graphs Our test graphs consist of two collections of directed graphs, all on 500 vertices, generated according to the preceding descriptions. The first collection consists of 120 graphs, with 10 for each of the following parameters: ER with $\rho \in \{0.03, 0.06, 0.1, 0.15, 0.2, 0.25\}$, GR with

$r \in \{0.1, 0.175, 0.3\}$, and PA with $k \in \{20, 40, 70\}$. We refer to this collection as the *point-drawn collection*, since the parameters are chosen from a discrete set of values.

For the second collection, we generated 300 directed graphs with 100 for each of the three random graph models. The parameters for a fixed model were obtained by generating 25 values independently uniformly at random from a set of four predefined intervals: ρ values are chosen from intervals in $\{(0, 0.01), (0.02, 0.03), (0.05, 0.07), (0.09, 0.1)\}$ for the ER graphs, r values from $\{(0, 0.02), (0.04, 0.05), (0.08, 0.12), (0.15, 0.175)\}$ for GR graphs, and k values from $\{(4, 7), (12, 18), (22, 25), (27, 30)\}$ for PA graphs. We refer to this collection as the *interval-drawn collection*.

With the point-drawn collection, we aim to understand the relationships of the topological pseudometrics to the model parameter and to be able to determine if strong relationships between a given pseudometric and a topological metric may have originated from the latent parameters of the models. The interval point-drawn collection allows us to study how well a pseudometric predicts topological features. We chose model parameters that ensure that our datasets include graphs with genuinely different topologies and graphs with non-trivial 6th Betti numbers.

4.2 Computing Topological Pseudometrics d_β and d_Δ

For each test graph, we run our experiments by comparing well-established pseudometrics (TriadEuclid, TriadEMD, and PD) against the topological pseudometrics d_β and d_Δ defined in Section 2.

Computing d_β and d_Δ To compute Betti numbers of directed graphs, we use the Flagser [16] software available via a Python implementation pyflagser [32]. For homology computations, Flagser comes with an approximation option, which speeds up computation time while maintaining remarkable accuracy [16]. In particular, it can be used to approximate the homology of a directed flag complex (associated with a test graph), by skipping columns that require more reduction steps than a user-chosen approximation parameter (denoted ϵ).

We use Flagser to compute the Betti numbers of the directed flag complex of each test graph up to dimension 6. We also compute approximate Betti numbers with $\epsilon = 10^0, 10^1, 10^2, 10^3$, and 10^4 to construct an approximate d_β pseudometric. The corresponding errors of the logarithm of the Betti numbers (i.e., vectors of $max\{0, log(\beta_k)\}$) across our full data set is 36.1%, 7.72%, 1.53%, 0.22% and 0.04%, respectively. Since there is no theoretical error estimate for those approximate parameters, we calculated exact Betti numbers and then compared the approximated Betti numbers in order to calculate the errors above. We remark that similar calculations with $\epsilon = 10^5$ produced the exact Betti numbers.

As explained in Section 2, we then compute the distance $d_\beta(G, G')$ between two directed graphs G and G' as the Euclidean distance between the vector of (possibly approximated) Betti numbers associated to the graphs G and G'. The pseudometric d_Δ is defined similarly, with Betti numbers replaced by simplex counts.

A parameter distance d_p In addition to d_β and d_Δ, we compute an additional pseudometric d_p for comparison, defined in terms of the parameters associated with our random graph models. For test graphs G and G' generated by a fixed random graph model (ER, GR, or PA), we define $d_p(G, G')$ to be the absolute value of the difference of the parameters used to generate them. For instance, if G and G' are generated with a ER random graph model with $\rho = 0.03$ and 0.06 respectively, then $d_p(G, G') = |0.03 - 0.06| = 0.03$.

Details of implementation All calculations are performed using Python 3.9. The source code for our experiments can be found in our github repository `https://github.com/winsy17/graph_pseudo_top_view`. Each ER/GR/PA graph is generated using the Python package NetworkX.

In the case of the point-drawn collection, we calculate the exact Betti numbers for all the parameters, except for $\rho \in \{0.15, 0.2, 0.25\}$, $r = 0.3$ and $k = 70$ for which we use the approximation option of pyflagser with $\epsilon = 10^5$ because they are not computationally feasible to calculate exactly. For the interval-drawn collection, we restrict the parameters to values that are computationally feasible and explore in pyflagser. We perform hierarchical clustering with complete linkage as our clustering technique. As a control, we also generate a random positive definite matrix with zeros on the diagonal that we include as a "random" pseudometric, denoted as "random".

Distributions of Betti numbers As illustrated in Figure 3, the Betti numbers seem to be strongly related to the parameters used to generate the test graphs. An analogous figure appears in [13] for the undirected case. This is well aligned with the observations in [15], where Lasalle studied the behavior of Betti numbers for directed ER graphs.

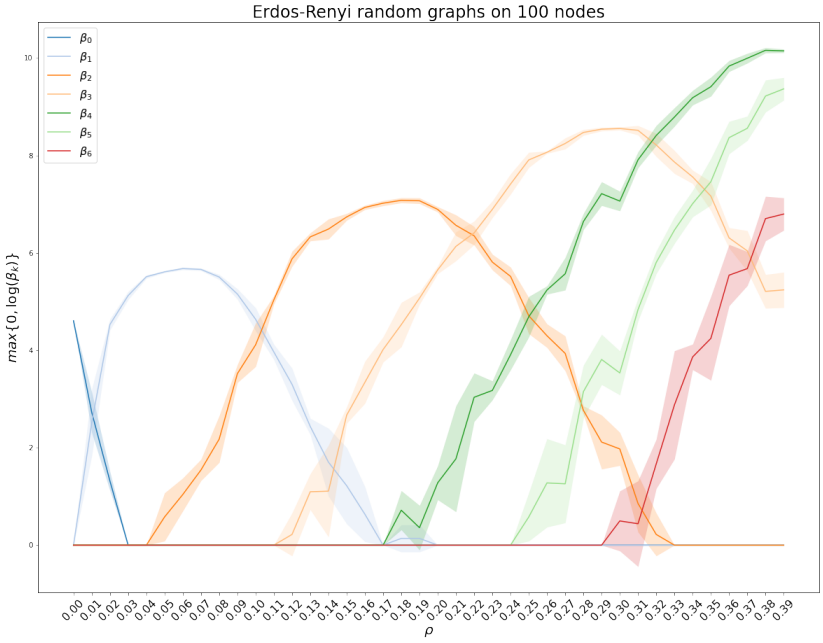

Fig. 3 The logarithm of Betti numbers across a range of parameters for Erdős–Rényi random graphs, each with 100 vertices. The x-axis corresponds to the parameter ρ ranging from 0 to 0.39, while the y-axis corresponds to the maximum of the logarithm of Betti numbers and 0. Similar results are observed for geometric random and preferential attachment random graphs.

5 Experimental Results

In this section, we first study the similarities between clusterings based on d_β and d_Δ and those based on the three well-established pseudometrics by computing the Fowlkes-Mallows indices and distance correlations between them (Section 5.1), first for each of the three random graph models individually, then for all three combined. This analysis shows that when we consider them model by model, the pseudometrics we study are closely related to one another, with a few exceptions, which vary between models. When all three models are considered simultaneously, the differences between the pseudometrics are substantially more stark.

We next apply k-NN classification and report the classification accuracy for each of the three models of random graphs by model parameter and for the full collection of all of our random graphs by model type (Section 5.2). We achieve 100% accuracy in almost all cases for the point-drawn collection. In the interval-drawn case, MSE is very low for the ER and GR models and rather high for the PA model.

We then explore the relationship of each of our pseudometrics to the various random graph parameters by performing permutation tests with distance correlation between a fixed pseudometric and d_β or d_Δ, for datasets from the interval-drawn collections (Section 5.3). When we consider all graph models together, the very small p-value enables us to reject the null hypothesis of independence of the pseudometrics. The results are more nuanced when we instead study the models individually, one parameter at a time, where we can reject the null hypothesis of independence only rarely and then almost only for the relationship between d_β and d_Δ.

Finally, we use k-NN regression on our pseudometrics in the interval-drawn case to predict the topological feature vectors $b(G)$ or $c(G)$ and report the MSE (Section 5.4). Both d_β and d_Δ perform well in predicting $b(G)$ and $c(G)$, better than any other pseudometric for the GR and PA models. The case of the ER model is more complex, as (unsurprisingly) d_Δ and d_β best predict $c(G)$ and $b(G)$, respectively, but the other pseudometrics perform quite well, too.

5.1 Fowlkes-Mallows Indices and Distance Correlation Between Pseudometrics

We compute the Fowlkes-Mallows (FM) indices and distance correlations of various pairs of pseudometrics for the interval-drawn collection of random graphs; recall that this contains 100 graphs for each of the three models, ER, GR, and PA. We first treat each set of 100 graphs separately and report the resulting FM indices and distance correlations for ER, GR, and PA in Figure 4, Figure 5, and Figure 6, respectively. We then combine all 300 graphs and report the results in Figure 7.

Recall that a higher value for the FM index indicates greater similarity between the clusterings induced by two pseudometrics. On the other hand, a higher value of distance correlation measures a higher level of dependence between the two pseudometrics. Both the FM index and distance correlation take values between 0 and 1.

Starting from Figure 4, we study various pseudometrics for a collection of 100 ER random graphs. The first column lists the pseudometrics we consider: from top to bottom, the topological pseudometrics based on Betti numbers (d_β) and simplex counts (d_Δ); the pseudometric d_p based on model parameters; well-established (previously known) pseudometrics including PD (Portrait Divergence), TriadEuclid, and TriadEMD; d_β using approximated Betti numbers with approximating parameters ($\epsilon = 10^0, 10^1, 10^2, 10^3, 10^4$); and a "random" pseudometric generated

with random positive definite matrix with zeros in the diagonal. The column labeled "**FM-**d_β" computes the FM index between d_β and the pseudometric of each row (treated as the benchmark classification). For instance, the FM index between d_β and PD is 0.8145. The number of clusters used for the FM index computation is chosen by a silhouette analysis of the column pseudometric. Notice that this way of choosing the number of clusters may result in different FM indices of the same two pseudometrics. For example, in Figure 4 the FM index of d_Δ with respect to d_β differs to the FM index of d_β with respect to d_Δ. Similarly, the column "**dCor-**d_Δ" computes the sample distance correlation between d_Δ and a pseudometric of each row. For instance, the sample distance correlation between d_Δ and PD is 0.86.

	FM – d_β	**FM** – d_Δ	**FM** – d_p	**dCor** – d_β	**dCor** – d_Δ	**dCor** – d_p
d_β	1	0.8464	0.8145	1	0.968	0.9687
d_Δ	0.982	1	0.8221	0.968	1	0.991
d_p	0.8145	0.8506	1	0.9687	0.991	1
PD	0.8145	0.9885	1	0.9057	0.86	0.8513
TriadEuclid	0.8145	0.8539	1	0.9418	0.9854	0.9819
TriadEMD	0.6907	0.9885	0.7377	0.9518	0.9948	0.9841
$d_\beta,\ \varepsilon = 10^0$	1	0.9885	0.8145	0.9904	0.9816	0.9739
$d_\beta,\ \varepsilon = 10^1$	1	0.8464	0.8145	0.9999	0.9696	0.9701
$d_\beta,\ \varepsilon = 10^2$	1	0.8464	0.8145	1	0.9681	0.9688
$d_\beta,\ \varepsilon = 10^3$	1	0.8464	0.8145	1	0.9681	0.9687
$d_\beta,\ \varepsilon = 10^4$	1	0.8464	0.8145	1	0.968	0.9687
Random	0.2619	0.1947	0.2501	0.1707	0.1819	0.166

Fig. 4 Interval-drawn collection: Erdös–Rényi (ER) random graphs

Considering Figure 4 closely, we notice that almost all the pseudometrics (except the random metric in the last row) are closely related, since the table contains high FM indices and high distance correlations. For simplicity, from now on, we exclude the last row (random) from our discussion.

In particular, the **FM-**d_β column of Figure 4 indicates that the clusters based on d_β are very similar to those based on d_Δ (treated as the benchmark classification), with an FM index of **0.982**. Moreover, d_β is highly correlated with d_Δ, with a distance correlation of **0.968** in the **dCor-**d_β column. In addition, d_β is also highly correlated with d_p where $dCor(d_\beta, d_p) =$ **0.9687**. However, as the TriadEMD entries in the **FM-**d_β and **dCor-**d_β columns show, high distance correlation between metrics (d_β vs. TriadEMD, **0.9518**) does not imply similar clusterings: the clusters given by d_β are less similar to those given by TriadEMD than for any other pseudometric, with an FM index of **0.6907**.

We remark that it is not too surprising that some of the numbers of Figure 4 associated to the Fawlkes-Mallows index coincide. For example, this occurs for **FM-**d_β of the pseudometrics d_p, PD and TriadEuclid. This phenomenon is due to identical clustering of the pseudometrics d_p, PD and TriadEuclid at the number of clusters chosen for d_β. This phenomenon can also be observed in subsequent figures/tables.

Careful inspection of Figure 5 and Figure 6 leads to a similar observation that the pseudometrics we study are highly related to one another (with a few exceptions – with FM index less than 0.7).

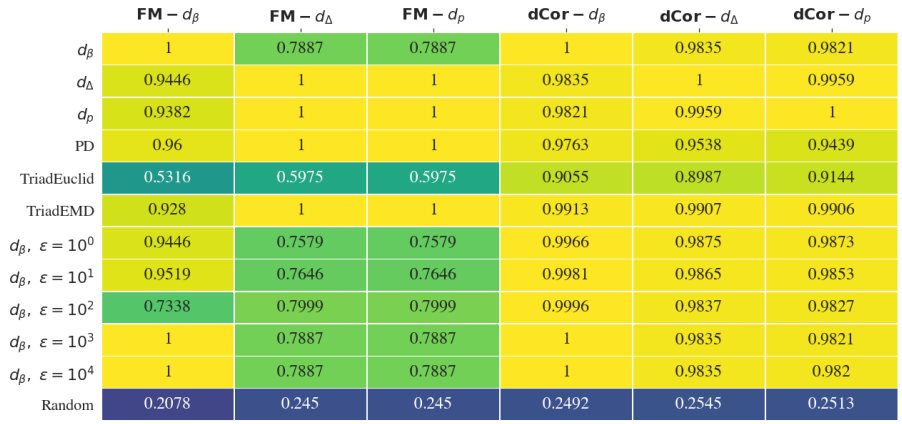

Fig. 5 Interval-drawn collection: geometric random (GR) graphs

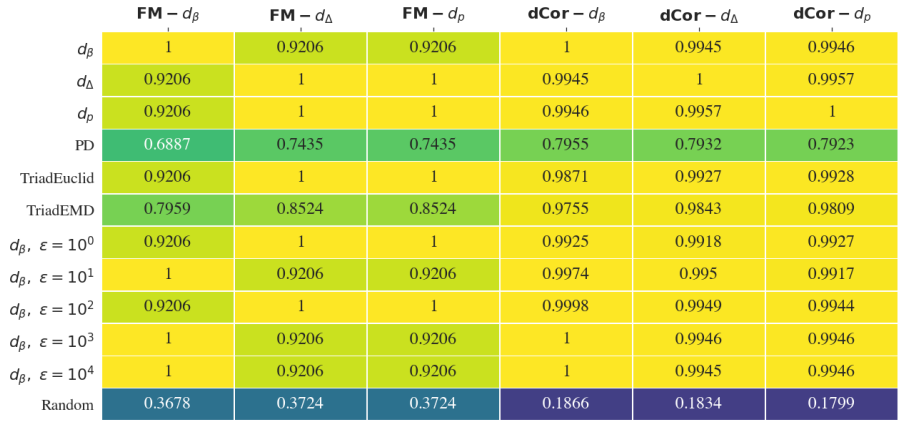

Fig. 6 Interval-drawn collection: preferential attachment (PA) random graphs

In particular, d_β and d_Δ consistently have high FM indices and high distance correlations across all three models. There are also some marked differences among the three models. For instance, the FM index between d_β and PD for PA random graphs is quite low – **0.6887** in the **FM-d_β** column of Figure 6 – in comparison with the other two models.

We combine all 300 random graphs to study FM indices and distance correlations in Figure 7. Both d_β and d_Δ remain reasonably similar to and dependent on one another, with an FM index of 0.7327 and a distance correlation of 0.9154. We also see a sharp fall in the FM indices and distance correlations between the the topological (d_β, d_Δ) and well-established pseudometrics (PD, TriadEuclid, TriadEMD). One potential explanation for these low FM indices and distance correlations is that the relationship observed model by model is due to a latent variable of the chosen model of random graph in addition to the parameter of that model.

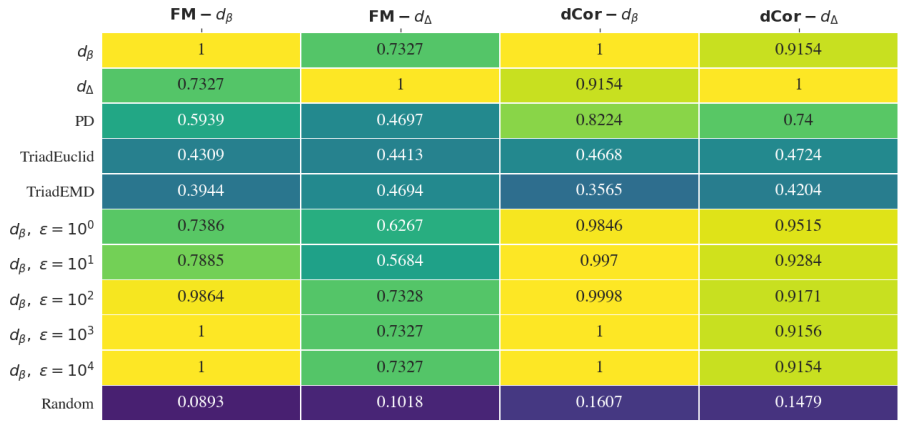

Fig. 7 Interval-drawn collection with all three random graph models

5.2 Classification Accuracy

Recall that our point-drawn collection of random graphs consists of 10 ER random graphs generated for each of six parameters (six labels), 10 GR graphs generated for each of three parameters (three labels), and 10 PA graphs generated for each of three parameters (three labels). We apply k-NN classification and report the classification accuracy for each of the three random graph models, as presented in Table 2, Table 3, and Table 4.

As shown in Table 2, using the PD, d_β, or d_Δ pseudometrics, we achieve 100% accuracy for all three random graph models; the accuracy for TriadEuclid and TriadEMD is slightly lower.

Table 2 Point-drawn collection: parameter classification rate within each set of random graphs

Pseudometrics	**ER**	**GR**	**PA**
PD	1	1	1
TriadEuclid	0.9167	1	1
TriadEMD	0.9167	1	1
d_β	1	1	1
d_Δ	1	1	1
Random	0.1833	0.2667	0.2

If we treat all 120 random graphs in the point-drawn collection together and try to classify them into 3 classes (ER, GR and PA), we achieve 100% classification accuracy for all pseudometrics (excluding the random metric), as shown in Table 3.

Table 3 Point-drawn collection: model classification rate for all three sets of random graphs combined

Pseudometrics	**Classification rate**
PD	1
TriadEuclid	1
TriadEMD	1
d_β	1
d_Δ	1
Random	0.3917

For the interval-drawn collection, we apply k-NN regression and report the MSE in predicting the model parameters. As shown in Table 4, for both ER and GR random graphs, the regression using both well-established and topological pseudometrics achieves very low MSE (excluding the random metric); while for the PA random graphs, none of these pseudometrics performs well.

Table 4 Interval-drawn collection: MSE with k-NN regression in predicting model parameters

Pseudometric	**ER** ($\times e-06$)	**GR** ($\times e-06$)	**PA**
PD	0.9787	7.269	0.1488
TriadEuclid	0.8367	447.6	0.1656
TriadEMD	2.207	11.83	0.6636
d_β	2.956	6.665	0.4048
d_Δ	1.653	9.03	0.4084
Random	1373	4047	98.3

5.3 Permutation Tests

We are interested in the statistical significance of relationships between well-established pseudometrics and d_β (respectively, d_Δ). Permutation tests provide a method to test statistical significance of high distance correlation or Fowlkes-Mallows (FM) index as demonstrating dependence between two pseudometrics. Our null hypothesis is that the two pseudometrics being compared are independent.

For small p-values we can reject the null hypothesis and deduce dependency. For insufficiently small p-value we cannot reject the null hypothesis of independence. This does not imply the two pseudometrics are independent, but rather that we cannot rule out the independence.

We perform permutation tests across all four datasets from the interval-drawn collection, based on either FM index or distance correlation as the dependency measures. One of our null-hypotheses is thus the independence of pseudometrics from our topological metrics (d_β, d_Δ or d_p) as measured by the FM index. We also consider the analogous null-hypothesis of independence with respect to distance correlation.

With exception of the random pseudometric, all of the calculated p-values equal 0.0006667. Hence we can reject both of our null-hypotheses with p-value equal to 0.0006667.

Except for the random control, the FM indices (respectively, distance correlations) calculated in Section 5.1 are so high that for every single random permutation the result is lower than for the original (unpermuted) FM indices (respectively, distance correlations). With 1500 permutations used in these tests, we compute a p-value of 0.000667. In each case we have performed at most 66

tests. Since $0.000667 < 0.05/66$ we reject the null-hypothesis (except the random control) with a family-wise error of 0.05.

If we were to consider all of the permutation tests combined, then the number of tests (264) would require a prohibitive number of permutations to make the Bonferroni correction useful. We would require p-values of at most 0.000189 to be able to establish any test as significant, which would require a minimum of 5280 permutations for each test.

We use instead the Benjamini-Hochberg procedure to combine all the tests. Since

$$0.000667 < \frac{242}{262 * C(262)} 0.005,$$

we can reject all the null hypotheses for all the pseudometrics except the random controls and still have a false discovery rate bounded above by 0.005. Here we have used that $C(264) < \ln(264) + 1 < 6.57$.

The results in Section 5.1 indicated that all of the pseudometrics are related in terms of what they detect in our random graph models. However, it is unclear to what extent this relationship might also rely on or be amplified by the differences between the models and model parameters chosen. It is therefore necessary to test conditional independence while taking these possible effects into account.

One way to control for these latent variables is to use the samples with fixed choices of generating model and model parameter, that is, analysing slices of the point-drawn collection taken based on parameter for each model. Table 5 through Table 10 contain the p-values of the permutation tests based on distance correlations between a pseudometric (in the first 1st column) and d_β or d_Δ.

We reject the null hypothesis test for independence when the p value is 0.0004998, which corresponds to the permutation tests where none of the random permutations had a smaller distance correlation than that of the original labels. These p-values are in bold.

To control for multiple hypothesis testing, we again consider the family-wise error over each table separately and then the false discovery rate over the tables combined. Tables 5 to 10 each have at most 30 entries, so we can say that all the results with values $p < 0.0005$ are significant with a family-wise error bound of 0.015.

We can again use the the Benjamini-Hochberg procedure to bound the false discovery rate on the p-values written in bold in Tables 5 to 10 combined. Since

$$0.0004998 < \frac{7}{120 * C(120)} 0.03,$$

(using $C(120) < \ln(120)+1 < 3.5$) we can report all of the p-values less than 0.0005 as significant, with an expected false discovery rate bounded above by 0.03. This means the expected number of false positives is less that 0.21.

Most of the rejected null hypotheses concern the independence of distances based on Betti numbers and on simplex counts. Furthermore these are occurring with the higher parameter values within the models. This is not surprising, as it ties in well with the literature on limit theorems for Betti numbers (see the survey paper [5] and its references).

What is more surprising is that for the geometric random graph with parameter value 0.3, we can reject independence between the simplex count metric and the portrait divergence metric. This suggests that for geometric graphs, the differences in simplex counts is in fact more closely related to the portrait divergence than it is to differences in the Betti numbers.

Although the various pseudometrics are not independent when we consider sets of directed graphs with multiple model parameters, we generally cannot reject independence once we restrict to specific models and parameter. This indicates that the relationships between these pseudometrics are probably driven by the latent variable of the models and their parameters. They may be capturing different features within these graphs (that are all affected in different ways by the model and parameter) and in doing so provide different perspectives on the graph structure.

Table 5 ER random graphs in point-drawn collection: p-values for distance correlation w.r.t. d_β

Pseudometrics	0.03	0.06	0.1	0.15	0.2	0.25
PD	0.5442	0.7251	0.1529	0.3673	0.8041	0.01349
TriadEuclid	0.1339	0.4803	0.8281	0.6887	0.2304	0.3318
TriadEMD	0.6632	0.1964	0.7206	0.9245	0.2909	0.2539
d_Δ	0.6387	0.7916	0.5922	0.5587	0.1399	**0.0004998**
Random	0.3603	0.5882	0.3693	0.2614	0.8431	0.3003

Table 6 ER random graphs in point-drawn collection: p-values for distance correlations w.r.t. d_Δ

Pseudometrics	0.03	0.06	0.1	0.15	0.2	0.25
PD	0.96	0.5187	0.3018	0.3638	0.5927	0.01299
TriadEuclid	0.7031	0.9535	0.6592	0.3973	0.8661	0.3188
TriadEMD	0.3138	0.4123	0.9355	0.4478	0.988	0.1939
d_β	0.6622	0.8061	0.6047	0.5592	0.1489	**0.0004998**
Random	0.7056	0.6022	0.1084	0.5872	0.9455	0.2454

Table 7 GR random graphs in point-drawn collection: p-values for distance correlation w.r.t. d_β

Pseudometrics	0.1	0.175	0.3
PD	0.03348	0.2034	0.02299
TriadEuclid	0.5642	0.2699	0.2519
TriadEMD	0.1034	0.02049	0.906
d_Δ	0.2769	0.2199	0.05597
Random	0.01449	0.1099	0.6187

Table 8 GR random graphs in point-drawn collection: p-values for distance correlation w.r.t. d_Δ

Pseudometrics	0.1	0.175	0.3
PD	0.3903	0.02199	**0.0004998**
TriadEuclid	0.5312	0.7641	0.8221
TriadEMD	0.7536	0.92	0.7866
d_β	0.2909	0.2214	0.06197
Random	0.5557	0.3488	0.3158

Table 9 PA random graphs in point-drawn collection: p-values for distance correlation w.r.t. d_β

Pseudometrics	20	40	70
PD	0.4858	0.07796	0.4513
TriadEuclid	0.8981	0.7466	0.4943
TriadEMD	0.1904	0.1074	0.5762
d_Δ	0.3058	**0.0004998**	**0.0004998**
Random	0.1979	0.7221	0.08796

Table 10 PA random graphs in point-drawn collection: p-values for distance correlation w.r.t. d_Δ

Pseudometrics	20	40	70
PD	0.947	0.04998	0.4253
TriadEuclid	0.2029	0.6912	0.6617
TriadEMD	0.9185	0.06047	0.6342
d_β	0.3013	**0.0004998**	**0.0004998**
Random	0.6522	0.3288	0.09545

5.4 Comparison of Clusterings and Classification Power

We now turn to testing how well each pseudometric performs in classification/regression tasks with respect to our graph models and parameters. For this, we focus on the interval-drawn collection, applying k-NN regression with respect to our pseudometrics to predict the topological feature vectors $b(G)$ and $c(G)$. Note that we have chosen the model parameters deliberately so as to potentially confuse any Betti number classifier, so that we are being very conservative about the potential power of Betti numbers for classification in a mixed model scenario. In the tables below, we use one of the pseudometrics of the first column to predict our topological feature vector and report the MSE results.

For individual random graph models, it is not surprising that d_β and d_Δ predict both simplex counts and Betti numbers fairly well. In the case of mixed model data in (Table 13 and Table 14), it is clear that PD outperforms the other well-established pseudometrics, as measured by MSE. In this same setting, the very large MSEs from the predictions using TriadEuclid and TriadEMD make them unsuitable methods for these classification tasks.

Table 11 Interval-drawn collection: MSE in predicting Betti numbers

Pseudometric	ER	GR	PA
PD	0.4179	0.6574	0.4238
TriadEuclid	0.4969	3.569	0.4338
TriadEMD	0.4504	0.8846	0.5576
d_β	0.1263	0.2526	**0.08243**
d_Δ	0.5233	0.588	**0.3343**
d_p	0.4736	0.6567	0.421
Random	46.15	39.86	25.92

Table 12 Interval-drawn collection: MSE in predicting simplex counts

Pseudometric	ER	GR	PA
PD	0.3368	0.8528	0.4618
TriadEuclid	0.3044	14.6	0.4803
TriadEMD	0.4215	0.9439	0.4793
d_β	0.5266	0.7196	**0.343**
d_Δ	0.124	0.2695	**0.08164**
d_p	0.2638	0.8092	0.3726
Random	51.31	130.6	33.12

Table 13 Interval-drawn collection: MSE in predicting Betti numbers, combining all models

Pseudometric	MSE
PD	**0.7495**
TriadEuclid	77.17
TriadEMD	77.45
d_β	0.1378
d_Δ	0.4866
Random	45.98

Table 14 Interval-drawn collection: MSE in predicting simplex count, combining all models

Pseudometric	MSE
PD	**2.53**
TriadEuclid	91.72
TriadEMD	92.47
d_β	1.183
d_Δ	0.1407
Random	79.16

6 Conclusion

In recent years several useful pseudometrics have been defined on the set $\mathcal{G}$ of finite directed graphs (digraphs), as tools for quantifying similarities and differences between digraphs. In this preliminary study, inspired by recent successful applications of topological methods to digraphs, we introduced two "topological" pseudometrics on $\mathcal{G}$, d_β and d_Δ, derived from Betti numbers and simplex counts of the directed flag complex of a digraph, respectively. We compared these new pseudometrics with three well-established pseudometrics (PD, TriadEuclid, TriadEMD), to determine to what extent the latter might be sensitive to topological structure.

The relationship between the topological pseudometrics d_β and d_Δ and the previously defined pseudometrics proved to be highly model-dependent. For example, we generally observed high values of Fowlkes-Mallows indices and distance correlations between all pairs of pseudometrics for a fixed random model (Figure 4, Figure 5, Figure 6), but significantly lower values when mixing models (Figure 7).

The relationship between d_β and d_Δ themselves is worth exploring, as calculating simplex counts is considerably computationally cheaper than calculating Betti numbers. When restricting to a specific random model, we obtained Fowlkes-Mallows indices between d_Δ and d_β greater than **0.9206** (column 1 of Figure 4, Figure 5, Figure 6). The scores were noticeably lower, however, when we mixed models (Figure 7), indicating that the relationship between d_β and d_Δ is also model-dependent.

We also compared the performance of d_β and d_Δ at two classification tasks for random graphs: estimating parameters within a random model and classifying graphs by random model (Section 5.2). We observed that the performance of d_β and d_Δ is comparable to that of PD, TriadEuclid, and TriadEMD in both of these tasks, achieving almost 100% classification accuracy (see Table 2, Table 3) or very low MSE (see Table 4), except in the case of the PA model. It would be interesting to determine the source of the relatively large MSE in this last case.

We explored, moreover, whether the model parameter explains the observed relationships between our topological metrics and PD, TriadEMD and TriadEuclid, using permutation tests with distance correlation and a null-hypothesis of independence (Section 5.3). In only very few cases, almost always concerning the pair (d_β, d_Δ), were we able to reject the null-hypothesis, so that the independence of most pairs of pseudometrics considered remains an open question for the graph models we studied.

Finally, we tested how well each pseudometric predicts Betti numbers and simplex counts (Section 5.4). With exception of TriadEuclid on GR graphs (Table 12), all of the pseudometrics performed very well with a very low MSE when predicting either Betti or simplex counts for a fixed random model (Table 11, Table 12). When combining random models, applying PD resulted in a very low MSE when predicting either Betti or simplex counts, while using either TriadEuclid or TriadEMD led to significant prediction error for both Betti and simplex counts. On the other hand, both d_β and d_Δ performed very well in the case of combined models. It follows in particular that when models underlying a collection of random graphs are either known and mixed or unknown, one should use either PD or one of d_β and d_Δ to predict Betti numbers or simplex counts. Applying d_Δ is slightly less accurate in its prediction of Betti numbers than d_β, which should be weighed against its ease of computation.

Although we considered only the topology of the directed flag complex in this study, our methods can be applied in a straightforward manner using other topological features, such as those obtained from the flag complex of an undirected graph or from the path homology of a directed graph [6, 10].

One limitation of this study is that all of the graphs analyzed have 500 vertices. It would be interesting to determine whether the relationships observed here hold for graphs of other sizes. In particular, understanding the asymptotic behavior of these relationships, as the number of vertices goes to infinity, should provide important insights. The Betti numbers of various types of random simplicial complexes [13] and random ER graphs [15] have been studied previously. A similar limit theorem for Betti numbers of directed flag complexes seems possible, though nontrivial, to formulate and prove, given the computations displayed in Figure 3.

Acknowledgments This work began at the Women in Computational Topology Workshop in 2019. The authors wish to thank the Mathematical Sciences Institute at ANU, the US National Science Foundation through the award CCF-1841455, the Australian Mathematical Sciences Institute, and the Association of Women in Mathematics for their financial support. ALGP is supported by the EPSRC grant "New Approaches to Data Science: Application Driven Topological Data Analysis" EP/R018472/1. NY is supported by the Alan Turing Institute under the EPSRC grant EP/N510129/1. KT is supported by an ARC DECRA fellowship. BW is supported in part by DOE DE-SC0021015 and NSF DBI-1661375. The authors would also like to thank Erika Roldán for insightful contributions, Gesine Reinert for sharing the source code for TriadEMD and TriadEuclid, the Oxford Mathematical Institute for providing access to computational resources, and the anonymous referees whose comments have helped to clarify this paper.

References

1. Adamaszek, M., Stacho, J.: Complexity of simplicial homology and independence complexes of chordal graphs. Computational Geometry: Theory and Applications **57**, 8–18 (2016)
2. Bagrow, J.P., Bollt, E.M.: An information-theoretic, all-scales approach to comparing networks. Applied Network Science **4**(1), 45 (2019)
3. Bagrow, J.P., Bollt, E.M., Skufca, J.D., ben Avraham, D.: Portraits of complex networks. EPL (Europhysics Letters) **81**(6), 68004 (2008)
4. Benjamini, Y., Yekutieli, D.: The control of the false discovery rate in multiple testing under dependency. The Annals of Statistics **29**(4), 1165–1188 (2001)
5. Bobrowski, O., Kahle, M.: Topology of random geometric complexes: a survey. Journal of Applied and Computational Topology **1**, 331–363 (2018)
6. Chowdhury, S., Mémoli, F.: Persistent path homology of directed networks. Proceedings of the 2018 Annual ACM-SIAM Symposium on Discrete Algorithms (2018)
7. Fowlkes, E.B., Mallows, C.L.: A method for comparing two hierarchical clusterings. Journal of the American Statistical Association **78**(383), 553–569 (1983)
8. Giusti, C., Ghrist, R., Bassett, D.S.: Two's company, three (or more) is a simplex. Journal of Computational Neuroscience **41**(1), 1–14 (2016)
9. Giusti, C., Pastalkova, E., Curto, C., Itskov, V.: Clique topology reveals intrinsic geometric structure in neural correlations. Proceedings of the National Academy of Sciences of the United States of America **112**(44), 13455–13460 (2015)
10. Grigor'yan, A., Lin, Y., Muranov, Y., Yau, S.T.: Path complexes and their homologies. Journal of Mathematical Sciences **248**, 564–599 (2020)
11. Hatcher, A.: Algebraic Topology. Cambridge University Press, Cambridge, United Kingdom (2002)
12. Jones, E., Oliphant, T., Peterson, P., et al.: SciPy: Open source scientific tools for Python (2001–). URL `http://www.scipy.org/`
13. Kahle, M., Meckes, E.: Limit the theorems for Betti numbers of random simplicial complexes. Homology, Homotopy and Applications **15**(1), 343–374 (2013)
14. Kriege, N.M., Johansson, F.D., Morris, C.: A survey on graph kernels. Applied Network Science **5**(1) (2020)
15. Lasalle, E.: Topological analysis of random graphs in the context of neuroscience. Unpublished report (2019)
16. Luetgehetmann, D., Govc, D., Smith, J., Levi, R.: Computing persistent homology of directed flag complexes. arXiv preprint arXiv:1906.10458 (2019)
17. Lyons, R.: Distance covariance in metric spaces. Annals of Probability **41**(5), 3284–3305 (2013)
18. Masulli, P., Villa, A.E.P.: The topology of the directed clique complex as a network invariant. SpringerPlus **5**(1) (2016)
19. Munkres, J.R.: Elements of algebraic topology. Addison-Wesley, Redwood City, CA, USA (1984)
20. Newman, M.: Networks. Oxford University Press, Oxford (2018)
21. Nikolentzos, G., Siglidis, G., Vazirgiannis, M.: Graph kernels: A survey. arXiv preprint arXiv:1904.12218 (2019)

22. Przulj, N.: Biological network comparison using graphlet degree distribution. Bioinformatics **26**(6), 853–854 (2010)
23. Pržulj, N., Corneil, D.G., Jurisica, I.: Modeling interactome: scale-free or geometric? Bioinformatics **20**(18), 3508–3515 (2004)
24. Reimann, M.W., Nolte, M., Scolamiero, M., Turner, K., Perin, R., Chindemi, G., Dłotko, P., Levi, R., Hess, K., Markram, H.: Cliques of neurons bound into cavities provide a missing link between structure and function. Frontiers in computational neuroscience **11**, 48 (2017)
25. Rousseeuw, P.J.: Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. Journal of Computational and Applied Mathematics **20**, 53–65 (1987)
26. Sarajlić, A., Malod-Dognin, N., Yaveroğlu, Ö.N., Pržulj, N.: Graphlet-based characterization of directed networks. Scientific Reports **6**(1) (2016)
27. Shen, C., Priebe, C.E., Vogelstein, J.T.: From distance correlation to multiscale graph correlation. Journal of the American Statistical Association **115**(529), 280–291 (2020)
28. Sizemore, A., Giusti, C., Bassett, D.S.: Classification of weighted networks through mesoscale homological features. Journal of Complex Networks (2016)
29. Sizemore, A.E., Giusti, C., Kahn, A., Vettel, J.M., Betzel, R.F., Bassett, D.S.: Cliques and cavities in the human connectome. Journal of Computational Neuroscience **44**(1), 115–145 (2018)
30. Székely, G.J., Rizzo, M.L., Bakirov, N.K.: Measuring and testing dependence by correlation of distances. Annals of Statistics **35**(6), 2769–2794 (2007)
31. Tantardini, M., Ieva, F., Tajoli, L., Piccardi, C.: Comparing methods for comparing networks. Scientific Reports **9**(1) (2019)
32. Tauzin, G., Lupo, U., Tunstall, L., Perez, J.B., Caorsi, M., Reise, W., Medina-Mardones, A.M., Dassatti, A., Hess, K.: giotto-tda: A topological data analysis toolkit for machine learning and data exploration. In: NeurIPS 2020 Workshop on Topological Data Analysis and Beyond (2020)
33. Turner, K., Spreemann, G.: Same but different: Distance correlations between topological summaries. arXiv preprint arXiv:1903.01051 (2019)
34. Wegner, A.E., Ospina-Forero, L., Gaunt, R.E., Deane, C.M., Reinert, G.: Identifying networks with common organizational principles. Journal of Complex Networks **6**(6), 887–913 (2018)
35. Xu, X., Reinert, G.: Triad-based comparison and signatures of directed networks. In: L.M. Aiello, C. Cherifi, H. Cherifi, R. Lambiotte, P. Lió, L.M. Rocha (eds.) Complex Networks and Their Applications VII. Complex Networks 2018. Studies in Computational Intelligence, vol. 812, pp. 590–602. Springer International Publishing (2018)