Dynamic Persistent Homology for Brain Networks via Wasserstein Graph Clustering

Moo K. Chung, Shih-Gu Huang, Ian C. Carroll,Vince D. Calhoun, H. Hill Goldsmith

Abstract We present the novel Wasserstein graph clustering for dynamically changing graphs. The Wasserstein clustering penalizes the topological discrepancy between graphs. The Wasserstein clustering is shown to outperform the widely used k-means clustering. The method applied in more accurate determination of the state spaces of dynamically changing functional brain networks.

1 Introduction

In standard graph theory based network analysis, network features such as node degrees and clustering coefficients are obtained from the adjacency matrices after thresholding weighted edges [50, 14]. The final statistical analysis results change depending on the choice of threshold or parameter [13, 32]. There is a need to develop a multiscale network analysis framework that provides consistent results and

Moo K. Chung
Department of Biostatistics and Medical Informatics, University of Wisconsin-Madison, USA e-mail: mkchung@wisc.edu

Shih-Gu Huang
Department of Biostatistics and Medical Informatics, University of Wisconsin-Madison, USA e-mail: shihgu@gmail.com

Ian C. Carroll
Department of Psychology, University of Wisconsin-Madison, USA e-mail: icarroll@wisc.edu|

Vince D. Calhoun
Tri-institutional Center for Translational Research in Neuroimaging and Data Science (TReNDS), Georgia State, Georgia Tech, Emory Georgia State University, Georgia, USA e-mail: vcalhoun@gsu.edu

H. Hill Goldsmith
Department of Psychology, University of Wisconsin-Madison, USA e-mail: hill.goldsmith@wisc.edu

interpretation regardless of the choice of parameter. Persistent homology, a branch of algebraic topology, offers a novel solution to this multiscale analysis challenge [21]. Instead of examining networks at one fixed scale, persistent homology identifies persistent topological features that are robust under different scales [39, 47]. Unlike existing graph theory approaches that analyze networks at one different fixed scale at a time and captures the changes of topological features over different scales and then identifies the most persistent topological features that are robust under noise perturbations. This robust performance under different scales is needed for *dynamic networks* that change over time.

Persistent homological network approaches are shown to be more robust and outperforming many existing graph theory measures and methods. In [31, 32], persistent homology was shown to outperform eight existing graph theory features such as clustering coefficient, samll-worldness and modularity. In [15, 17], persistent homology was shown to outperform various matrix norm based network distances. In [56], persistent homology was shown to outperform the power spectral density and local variance methods. In [55], persistent homology was shown to outperform topographic power maps. In [60], center persistency was shown to outperform the network-based statistic and element-wise multiple corrections. However, the method has been mainly used on *static* networks or as a static summary of time varying networks [5]. The dynamic pattern of persistent homology for time varying brain network was rarely investigated expect few [59, 44, 48].

In this paper, we propose to develop the novel *dynamic persistent homology* framework for time varying graph data. We will show that the proposed method based on the Wasserstein distance can capture the topological patterns that are consistently observed across different time points. The Wasserstein distance or Kantorovich–Rubinstein metric is originally defined between probability distributions [54]. Due to the connection to the optimal mass transport, which enjoys various optimal properties, the Wasserstein distance has been applied to various imaging applications. However, there are not many applications of Wasserstein distance in network data. [36] used the Wasserstein distance in resampling brain surface meshes. [46] used the Wasserstein distance in classifying brain cortical surface shapes. [43] used the Wassterstein distance for manifold regression problem in the space of positive definite matrices for the source localization problem in EEG. [57] used the Wasserstein distance in predicting Alzheimer's disease progression in magnetoencephalography (MEG) brain networks. However, the Wassterstein distance in these applications are all geometric in nature.

The main contribution of our paper is as follows. We present a coherent scalable framework for the computation of Wasserstein distance on graphs. We directly build the Wasserstein distance using the edge weights in graphs making the method far more accessible and adaptable. We achieve $O(n \log n)$ run time in most graph manipulation tasks such as matching and averaging. Such scalable computation enables us to perform a computationally demanding graph clustering task with ease. The method is applied in the determination of he state spaces of dynamically changing functional brain networks.

2 Graphs as simplical complexes

A high dimensional object such as brain networks can be modeled as weighted graph $\mathcal{X} = (V, w)$ consisting of node set V indexed as $V = \{1, 2, \cdots, p\}$ and edge weights $w = (w_{ij})$ between nodes i and j. If we order the edge weights in the increasing order, we have the sorted edge weights:

$$\min_{j,k} w_{jk} = w_{(1)} < w_{(2)} < \cdots < w_{(q)} = \max_{j,k} w_{jk},$$

where $q \leq (p^2 - p)/2$. The subscript $_{(\,)}$ denotes the order statistic. In terms of sorted edge weight set $W = \{w_{(1)}, \cdots, w_{(q)}\}$, we may also write the graph as $\mathcal{X} = (V, W)$. If we connect nodes following some criterion on the edge weights, they will form a simplical complex which will follow the topological structure of the underlying weighted graph [21, 61]. Note that the k-simplex is the convex hull of $k+1$ points in V. A simplicial complex is a finite collection of simplices such as points (0-simplex), lines (1-simplex), triangles (2-simplex) and higher dimensional counter parts.

The *Rips complex* $\mathcal{X}_\epsilon$ is a simplicial complex, whose k-simplices are formed by $(k+1)$ nodes which are pairwise within distance ϵ [23]. While a graph has at most 1-simplices, the Rips complex has at most $(p-1)$-simplices. The Rips complex induces a hierarchical nesting structure called the Rips filtration

$$\mathcal{X}_{\epsilon_0} \subset \mathcal{X}_{\epsilon_1} \subset \mathcal{X}_{\epsilon_2} \subset \cdots$$

for $0 = \epsilon_0 < \epsilon_1 < \epsilon_2 < \cdots$, where the sequence of ϵ-values are called the filtration values. The filtration is quantified through a topological basis called *k-cycles*. 0-cycles are the connected components, 1-cycles are 1D closed path or loop while 2-cycles are a 2-simplices without interior. Any k-cycles can be represented as a linear combination of basis k-cycles. The Betti numbers β_k counts the number of independent k-cycles. During the Rips filtration, the i-th k-cycles are born at filtration value b_i and die at d_i. The collection of all the paired filtration values

$$P(\mathcal{X}) = \{(b_1, d_1), \cdots, (b_q, d_q)\}$$

displayed as 1D intervals is called the *barcode* and displayed as scatter points in 2D plane is called the *persistent diagram*. Since $b_i < d_i$, the scatter points in the persistent diagram are displayed above the line $y = x$ line by taking births in the x-axis and deaths in the y-axis.

For dynamically changing brain network $\mathcal{X}(t) = (V, w(t))$, we assume the node set is fixed while edge weights are changing over time t. If we build persistent homology at each fixed time, the resulting barcode is also time dependent:

$$P(\mathcal{X}(t)) = \{(b_1(t), d_1(t)), \cdots, (b_q(t), d_q(t))\}.$$

2.1 Graph filtrations

As the number of nodes p increases, the resulting Rips complex becomes very dense. As the filtration values increases, there exists an edge between every pair of nodes. At higher filtration values, Rips filtration becomes an ineffective representation of networks. To remedy this issue, graph filtration was introduced [31, 32]. Given weighted graph $\mathcal{X} = (V, w)$ with edge weight $w = (w_{ij})$, the binary network $\mathcal{X}_\epsilon = (V, w_\epsilon)$ is a graph consisting of the node set V and the binary edge weights $w_\epsilon = (w_{\epsilon,ij})$ given by

$$w_{\epsilon,ij} = \begin{cases} 1 & \text{if } w_{ij} > \epsilon; \\ 0 & \text{otherwise.} \end{cases}$$

Note w_ϵ is the adjacency matrix of $\mathcal{X}_\epsilon$, which is a simplicial complex consisting of 0-simplices (nodes) and 1-simplices (edges) [23]. While the binary network $\mathcal{X}_\epsilon$ has at most 1-simplices, the Rips complex can have at most $(p-1)$-simplices. By choosing threshold values at sorted edge weights $w_{(1)}, w_{(2)}, \cdots, w_{(q)}$ [13], we obtain the sequence of nested graphs:

$$\mathcal{X}_{w_{(1)}} \supset \mathcal{X}_{w_{(2)}} \supset \cdots \supset \mathcal{X}_{w_{(q)}}.$$

The sequence of such a nested multiscale graph is called as the *graph filtration* [31, 32]. Figure 1 illustrates a graph filtration in a 4-nodes example. Note that $\mathcal{X}_{w_{(1)}-\epsilon}$ is the complete weighted graph for any $\epsilon > 0$. On the other hand, $\mathcal{X}_{w_{(q)}}$ is the node set V. By increasing the threshold value, we are thresholding at higher connectivity so more edges are removed.

For dynamically changing brain networks, we can similarly build time varying graph filtrations at each time point $\{\mathcal{X}_w(t) : t \in \mathbb{R}^+\}$.

2.2 Birth-death decomposition

Unlike the Rips complex, there are no higher dimensional topological features beyond the 0D and 1D topology in graph filtration. The 0D and 1D persistent diagrams (b_i, d_i) tabulates the life-time of 0-cycles (connected components) and 1-cycles (loops) that are born at the filtration value b_i and die at value d_i. The 0th Betti number $\beta_0(w_{(i)})$ at filtration value $w_{(i)}$ counts the number of 0-cycles and shown to be non-decreasing over filtration (Figure 1) [17]: $\beta_0(w_{(i)}) \leq \beta_0(w_{(i+1)})$. On the other hand the 1st Betti number $\beta_1(w_{(i)})$ counts the number of independent loops and shown to be non-increasing over filtration (Figure 1) [17]: $\beta_1(w_{(i)}) \geq \beta_1(w_{(i+1)})$.

During the graph filtration, when new components is born, they never dies. Thus, 0D persistent diagrams are completely characterized by birth values b_i only. Loops are viewed as already born at $-\infty$. Thus, 1D persistent diagrams are completely

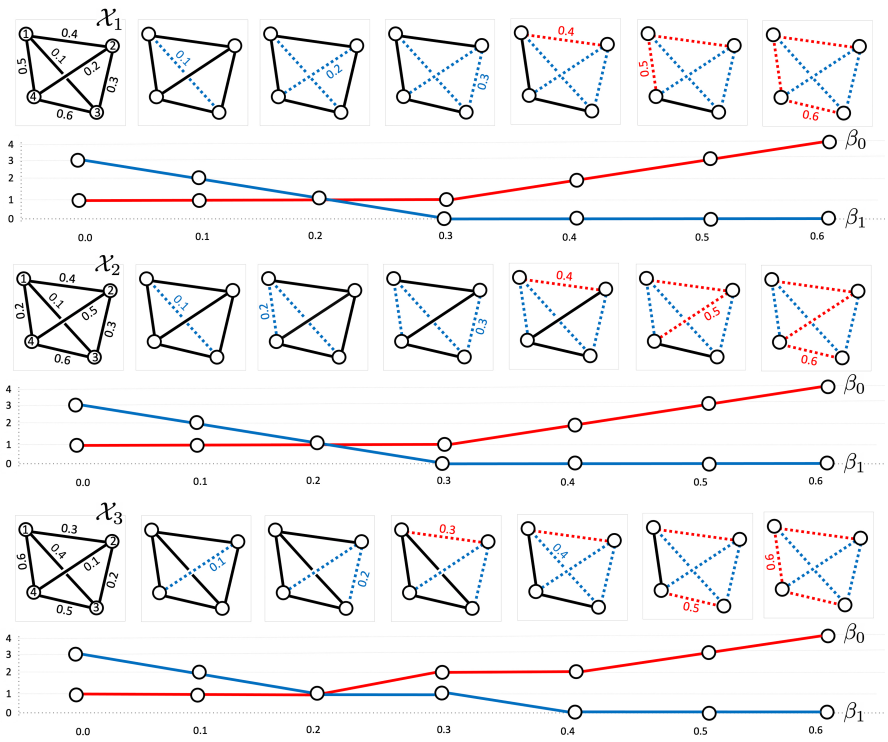

Fig. 1 Graph filtrations are obtained by sequentially thresholding graphs in increasing edge weights. The 0-th Betti number β_0 (number of connected components) and the first Betti number β_1 (number of cycles) are then plotted over the filtration values. The Betti curves are monotone over graph filtrations. However, different graphs (top vs. middle) can yield identical Betti curves.

characterized by death values d_i only. We can show that the edge weight set W can be partitioned into 0D birth values and 1D death values [49]:

Theorem 1 (Birth-death decomposition) *The edge weight set* $W = \{w_{(1)}, \cdots, w_{(q)}\}$ *has the unique decomposition*

$$W = W_b \cup W_d, \quad W_b \cap W_d = \emptyset \tag{1}$$

where birth set $W_b = \{b_{(1)}, b_{(2)}, \cdots, b_{(q_0)}\}$ *is the collection of 0D sorted birth values and death set* $W_d = \{d_{(1)}, d_{(2)}, \cdots, d_{(q_1)}\}$ *is the collection of 1D sorted death values with* $q_0 = p - 1$ *and* $q_1 = (p-1)(p-2)/2$. *Further* W_b *forms the 0D persistent diagram while* W_d *forms the 1D persistent diagram.*

Proof During the graph filtration, when an edge is deleted, either a new component is born or a cycle dies [17]. These events are disjoint and does not happen at the same time. The claim is proved by contradiction. Assume the both events happen at the same time in contrary. Then β_0 increases by 1 while β_1 decreases by 1. When

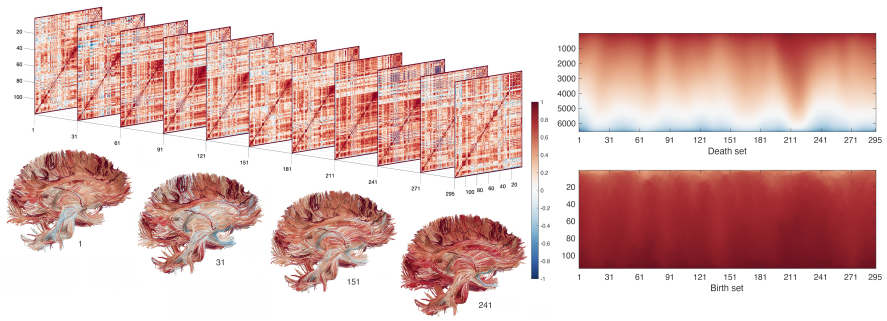

Fig. 2 Left: Dynamically changing correlation matrices computed from rs-fMRI using the sliding window of size 60 for a subject [26]. The constructed correlation matrices are superimposed on top of the white matter fibers [16], which are colored based on correlation values. Right: The corresponding birth and death sets that are changing over time. Columns are the sorted birth and death edge values at that particular time point.

an edge is deleted, the number of nodes p is fixed while the number of edges q is reduced to $q-1$. Thus the Euler characteristic $\chi = p - q$ of the graph increases by 1. The Euler characteristic can be also given by an alternating sum $\chi = \beta_0 - \beta_1$ [1]. Subsequently, the Euler characteristic increases by 2, which contradict the previous computation. Thus, both events cannot occur at the same time. This establishes the decomposition $W = W_b \cup W_d, W_b \cap W_d = \emptyset$.

In a complete graph with p nodes, there are $q = p(p-1)/2$ unique edge weights. There are $q_0 = p - 1$ number of edges that produces 0-cycles. This is equivalent to the number of edges in the maximum spanning tree of the graph. Since W_b and W_d partition the set, there are

$$q_1 = q - q_0 = \frac{(p-1)(p-2)}{2}$$

number of edges that destroys 1-cycles.

The 0D persistent diagram of the graph filtration is given by $\{(b_{(1)}, \infty), \cdots, (b_{(q_0)}, \infty)\}$. Ignoring ∞, W_b is the 0D persistent digram. The 1D persistent diagram of the graph filtration is given by $\{(-\infty, d_{(1)}), \cdots, (-\infty, d_{(q_1)})\}$. Ignoring $-\infty$, W_d is the 1D persistent digram. □

Numerical implementation. The algorithm for decomposing the birth and death set is as follows. As the corollary of Theorem 1, we can show that the birth set is the maximum spanning tree (MST). The identification of W_b is based on the modification to Kruskal's or Prim's algorithm and identify the MST [32]. Then W_d is identified as W/W_d. Figure 1 displays graph filtration on 2 different graphs with 4 nodes, where the birth sets consists of 3 red edges and the death sets consist of 3 blue edges. Figure 2 displays how the birth and death sets change over time in the brain network of a single subject. We made the computer codes available at `http://www.stat.wisc.edu/~mchung/dynamicTDA`. Given edge weight matrix

W as input, the Matlab function `WS_decompose.m` outputs the birth set W_b and the death set W_d.

2.2.1 Algebra on birth-death decompositions

Consider graph $\mathcal{X} = (V, w)$ with the birth-death decompositions $W = W_b \cup W_d$:

$$W_b = \{b_{(1)}, \cdots, b_{(q_0)}\}, \quad W_d = \{d_{(1)}, \cdots, d_{(q_1)}\}.$$

Let $\mathcal{F}(W) = w$ be the function that maps each edge in the ordered edge set W back to the original edge weight matrix w. $\mathcal{F}^{-1}(w) = W$ is the function that maps each edge in the edge weight matrix to the birth death decomposition. Such maps are one-to-one. Since W_b and W_d are disjoint, we can write as

$$\mathcal{F}(W_b \cup W_d) = \mathcal{F}(W_b) \oplus \mathcal{F}(W_d).$$

Define the *scalar multiplication* on the ordered set W as

$$cW = (cW_b) \cup (cW_d) = \{cb_{(1)}, \cdots, cb_{(q_0)}\} \cup \{cd_{(1)}, \cdots, cd_{(q_1)}\}$$

for $c \in \mathbb{R}$. Then we have $\mathcal{F}(cW) = c\mathcal{F}(W)$ for $c > 0$. The relation does not hold for $c < 0$ since it is not order preserving. Define the *scalar addition* on the ordered set W as

$$c + W = (c + W_b) \cup (c + W_d) = \{c + b_{(1)}, \cdots, c + b_{(q_0)}\} \cup \{c + d_{(1)}, \cdots, c + d_{(q_1)}\}$$

for $c \in \mathbb{R}$. Since the addition is order preserving, $\mathcal{F}(c+W) = c+\mathcal{F}(W)$ for all $c \in \mathbb{R}$.

Define scalar multiplication of c to graph $\mathcal{X} = (V, w)$ as $c\mathcal{X} = (V, c\mathcal{F}(W))$. Define the scalar addition of c to graph $\mathcal{X}$ as $c+\mathcal{X} = (V, c+\mathcal{F}(W))$. Let $c = c_b \cup c_d$ be an ordered set with $c_b = (c^b_{(1)}, \cdots, c^b_{(q_0)})$ and $c_d = (c^d_{(1)}, \cdots, c^d_{(q_1)})$. Define the *set addition* of c to the ordered set W as

$$c + W = (c_b + W_b) \cup (c_d + W_d)$$

with $c_b + W_b = \{c^b_{(1)} + b^k_{(1)}, \cdots, c^b_{(q_0)} + b^k_{(q_0)}\}$ and $c_d + W_d = \{c^d_{(1)} + d^k_{(1)}, \cdots, c^d_{(q_1)} + d^k_{(q_1)}\}$. Then we have the following decomposition.

Theorem 2 *For graph $\mathcal{X} = (V, w)$ with the birth-death decompositions $W = W_b \cup W_d$ and positive ordered sets c_b and c_d, we have*

$$\mathcal{F}((c_b + W_b) \cup W_d) = (c_b + \mathcal{F}(W_b)) \oplus \mathcal{F}(W_d) \tag{2}$$

$$\mathcal{F}(W_b \cup (c_d - c_\infty + W_d)) = \mathcal{F}(W_b) \oplus (\mathcal{F}(c_d - c_\infty + W_d), \tag{3}$$

where c_∞ is a large number bigger than any element in c_d.

Proof Note $c_b + W_b$ is order preserving. W_b is the MST of graph $\mathcal{X}$. The total edge weights of MST does not decrease if we change all the edge wights of MST from

W_b to $c_b + W_b$. Thus $c_b + W_b$ will be still MST and $\mathcal{F}(c_b + W_b) = c_b + \mathcal{F}(W_b)$. The death set W_d does not change when the edges in MST increases. This proves (2).

The sequence $(a_1, \cdots, a_{q1}) = c_d - c_\infty$ with $a_i = c^d_{(i)} - c_\infty < 0$ is increasing. Adding $(a_1, \cdots a_{q1})$ to W_d is order preserving. Decreasing edge weights in W_d will not change the total edge weights of MST. Thus the birth set is still identical to W_b. Then the death set is $c_d - c_\infty + W_d$. This proves (3). □

The decomposition (3) does not work if we simply add an arbitrary ordered set to W_d since it will change the MST. Numerically the above algebraic operations are all linear in run time and will not increase the computational load. So far, we demonstrated what the valid algebraic operations are on the birth-death decompositions. Now we address a more important question of *if the birth-death decomposition is addictive.* Given graphs $\mathcal{X}_1 = (V, w^1)$ and $\mathcal{X}_2 = (V, w^2)$ with corresponding birth-death decompositions $W_1 = W_{1b} \cup W_{1d}$ and $W_2 = W_{2b} \cup W_{2d}$, define the sum of graphs $\mathcal{X}_1 + \mathcal{X}_2$ as a graph $\mathcal{X} = (V, w)$ with birth-death decomposition

$$W_b \cup W_d = (W_{1b} + W_{2b}) \cup (W_{1d} + W_{2d}). \tag{4}$$

However, it is unclear if there even exists a unique graph with decomposition (4).

Define *projection* $\mathcal{F}(W_1|W_2)$ as the projection of edge values in the ordered set W_1 onto the edge weight matrix $\mathcal{F}(W_1)$ such that the birth values W_{1b} are sequentially mapped to the $\mathcal{F}(W_{2b})$ and the death values W_{1d} are sequentially mapped to the $\mathcal{F}(W_{2d})$. Trivially, $\mathcal{F}(W_1|W_1) = \mathcal{F}(W_1)$. In general, $\mathcal{F}(W_1|W_2) \neq \mathcal{F}(W_2|W_1)$. The projection can be written as

$$\mathcal{F}(W_1|W_2) = \mathcal{F}(W_{1b}|W_{2b}) \oplus \mathcal{F}(W_{1d}|W_{2d}).$$

Theorem 3 *Given graphs $\mathcal{X}_1 = (V, w^1)$ and $\mathcal{X}_2 = (V, w^2)$ with corresponding birth-death decompositions $W_1 = W_{1b} \cup W_{1d}$ and $W_2 = W_{2b} \cup W_{2d}$, there exists graph $\mathcal{X} = (V, w)$ with birth-death decomposition $W_b \cup W_d$ satisfying*

$$W_b \cup W_d = (W_{1b} + W_{2b}) \cup (W_{1d} + W_{2d}).$$

with

$$w = \mathcal{F}(W_b \cup W_d) = \mathcal{F}(W_{1b} + W_{2b}|W_{1b}) \oplus \mathcal{F}(W_{1d} + W_{2d}|W_{1d}).$$

Proof We prove by the explicit construction in a sequential manner by applying only the valid operations.

1) Let c_∞ be some fixed number larger than any edge weights in w^1 and w^2. Add c_∞ to the decomposition $W_{1b} \cup W_{1d}$ to make all the edges positive:

$$c_\infty + W_{1b} \cup W_{1d} = (c_\infty + W_{1b}) \cup (c_\infty + W_{1d}). \tag{5}$$

The edge weight matrix is given by

$$\mathcal{F}((c_\infty + W_{1b}) \cup (c_\infty + W_{1d})) = c_\infty + \mathcal{F}(W_1).$$

2) We add the ordered set W_{2b} to decomposition (5) and obtain

$$c_\infty + (W_{1b} + W_{2b}) \cup W_{1d} = (c_\infty + W_{1b} + W_{2b}) \cup (c_\infty + W_{1d}). \tag{6}$$

We next determine how the corresponding edge weight matrix changes when the birth-death decomposition changes from (5) to (6). Increasing birth values from $c_\infty + W_{1b}$ to $c + W_{1b} + W_{2b}$ increases the total edge weights in the MST of $c_\infty + \mathcal{X}_1$. Thus, $c + W_{1b} + W_{2b}$ is still MST. The death set does not change from $c_\infty + W_{1d}$. The edge weight matrix is then given by

$$\begin{aligned} &\mathcal{F}((c_\infty + W_{1b} + W_{2b}) \cup (c_\infty + W_{1d})) \\ &= \mathcal{F}(c_\infty + W_{1b} + W_{2b}|W_{1b}) \oplus \mathcal{F}(c_\infty + W_{1d}). \end{aligned} \tag{7}$$

(7) can be also derived from (2) in Theorem 2 as well.

3) Add ordered set $W_{2d} - c_\infty$ to the death set in the decomposition (6) and obtain

$$(c_\infty + W_{1b} + W_{2b}) \cup (W_{1d} + W_{2d}). \tag{8}$$

Decreasing death values from $c_\infty + W_{1d}$ to $W_{1d} + W_{2d}$ does not affect the the total edge weights in the MST of (7). There is no change in MST. The birth set does not change from $c + W_{1b} + W_{2b}$. Thus,

$$\begin{aligned} &\mathcal{F}((c_\infty + W_{1b} + W_{2b}) \cup (W_{1d} + W_{2d})) \\ &= \mathcal{F}(c_\infty + W_{1b} + W_{2b}|W_{1b})\mathcal{F}(W_{1d} + W_{2d}|W_{1d}) \\ &= (c_\infty + \mathcal{F}(W_{1b} + W_{2b}|W_{1b})) \oplus \mathcal{F}(W_{1d} + W_{2d}|W_{1d}) \end{aligned} \tag{9}$$

Since edge weights in $W_{2d} - c_\infty$ are all negative, we can also obtain the above result from Theorem 2.

4) Finally we subtract c_∞ from the brith set in (8) and obtain the projection of sum onto W_1.

$$\mathcal{F}(W_{1b} + W_{2b}|W_{1b}) \oplus \mathcal{F}(W_{1d} + W_{2d}|W_{1d}). \tag{10}$$

□

Remark. Theorem 3 does not guarantee the uniqueness of edge weight matrices. Instated of projecting birth and death values onto the first graph, we can also project onto the second graph

$$\mathcal{F}(W_{1b} + W_{2b}|W_{2b}) \oplus \mathcal{F}(W_{1d} + W_{2d}|W_{2b}).$$

or any other graph. Different graphs can have the same birth-death sets. Figure 3 shows two different graphs with the identical birth and death sets.

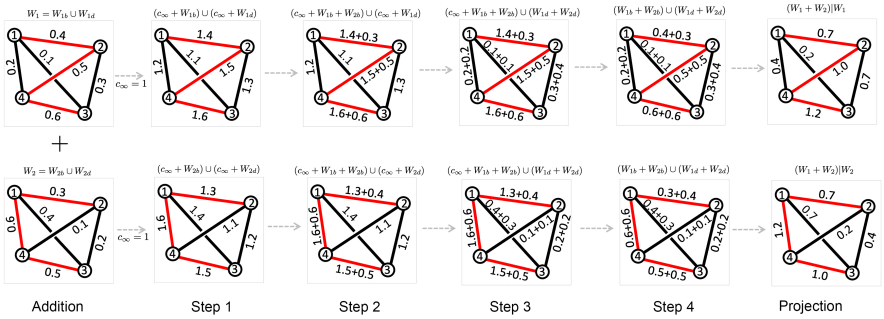

Fig. 3 Schematic of proof of Theorem 3 with 4-nodes examples. Each step of operations yield graphs with valid birth-death decompositions. The first row is the construction of sum operation by projecting to W_1. The second row is the construction of sum operation by projecting to W_2. Red colored edges are the maximum spanning trees (MST). Each addition operation will not change MST. Eventually, we can have two different graphs with the identical birth-death decomposition.

3 Wasserstein graph clustering

Consider persistent diagrams P_1 and P_2 given by

$$P_1 : x_1 = (b_i^1, d_i^1), \cdots, x_q = (b_q^1, d_q^1), \quad P_2 : y_1 = (b_i^2, d_i^2), \cdots, y_q = (b_q^2, d_q^2).$$

Their empirical distributions are given in terms of Dirac-Delta functions

$$f_1(x) = \frac{1}{q}\sum_{i=1}^{q} \delta(x - x_i), \quad f_2(y) = \frac{1}{q}\sum_{i=1}^{q} \delta(y - y_i).$$

Then we can show that the 2-*Wasserstein distance* on persistent diagrams is given by

$$D_W(P_1, P_2) = \inf_{\psi: P_1 \to P_2} \Big(\sum_{x \in P_1} \|x - \psi(x)\|^2 \Big)^{1/2} \tag{11}$$

over every possible bijection ψ between P_1 and P_2 [54]. Optimization (11) is the standard assignment problem, which is usually solved by Hungarian algorithm in $O(q^3)$ [22]. However, for graph filtration, the distance can be computed in $O(q \log q)$ by simply matching the order statistics on birth or death sets [41, 49]:

Theorem 4 *The 2-Wasserstein distance between the 0D persistent diagrams for graph filtration is given by*

$$D_{W0}(P_1, P_2) = \Big[\sum_{i=1}^{q_0} (b_{(i)}^1 - b_{(i)}^2)^2 \Big]^{1/2},$$

where $b^j_{(i)}$ is the i-th smallest birth values in persistent diagram P_j. The 2-Wasserstein distance between the 1D persistent diagrams for graph filtration is given by

$$D_{W1}(P_1, P_2) = \Big[\sum_{i=1}^{q_1} (d^1_{(i)} - d^2_{(i)})^2 \Big]^{1/2},$$

where $d^j_{(i)}$ is the i-th smallest death values in persistent diagram P_j.

Proof 0D persistent diagram is given by $\{(b_{(1)}, \infty), \cdots, (b_{(q_0)}, \infty)\}$. Ignoring ∞, the 0D Wasserstein distance is simplified as

$$D^2_{W0}(P_1, P_2) = \min_{\psi} \sum_{i=1}^{q_0} |b^1_i - \psi(b^1_i)|^2,$$

where the minimum is taken over every possible bijection ψ from $\{b^1_1, \cdots, b^1_{q_0}\}$ to $\{b^2_1, \cdots, b^2_{q_0}\}$. Note $\sum_{i=1}^{q_0} |b^1_i - \psi(b^1_i)|^2$ is minimum only if $\sum_{i=1}^{q_0} b^1_i \psi(b^1_i)$ is maximum. Rewrite $\sum_{i=1}^{q_0} b^1_i \psi(b^1_i)$ in terms of the order statistics as $\sum_{i=1}^{q_0} b^1_{(i)} \psi(b^1_{(i)})$. Now, we prove by *induction*. When $q = 2$, there are only two possible bijections:

$$b^1_{(1)} b^2_{(1)} + b^1_{(2)} b^2_{(2)} \quad \text{and} \quad b^1_{(1)} b^2_{(2)} + b^1_{(2)} b^2_{(1)}.$$

Since $b^1_{(1)} b^2_{(1)} + b^1_{(2)} b^2_{(2)}$ is larger, $\psi(b^1_{(i)}) = b^2_{(i)}$ is the optimal bijection. When $q_0 = k$, assume $\phi(b^1_{(i)}) = b^2_{(i)}$ is the optimal bijection. When $q_0 = k + 1$,

$$\max_{\psi} \sum_{i=1}^{k+1} b^1_{(i)} \psi(b^2_{(i)}) \le \max_{\psi} \sum_{i=1}^{k} b^1_{(i)} \psi(b^1_{(i)}) + \max_{\psi} b^1_{(k+1)} \tau(b^1_{(k+1)}).$$

The first term is maximized if $\psi(b^1_{(i)}) = b^2_{(i)}$. The second term is maximized if $\psi(b^1_{(k+1)}) = b^2_{(k+1)}$. Thus, we proved the statement.

1D persistent diagram of graph filtration is given by $\{(-\infty, d_{(1)}), \cdots, (-\infty, d_{(q)})\}$. Ignoring $-\infty$, the Wasserstein distance is given by

$$D^2_{W1}(P_1, P_2) = \min_{\psi} \sum_{i=1}^{q_1} |d^1_i - \psi(d^1_i)|^2.$$

Then we follow the similar inductive argument as the 0D case. □

3.1 Graph matching via the Wasserstein distance

Using the Wasserstein distance between two graphs, we match graphs at the edge level. In the usual graph matching problem, the node labels do not have to be matched and thus, the problem is different from simply regressing brain connectivity matrices

over other brain connectivity matrices at the edge level [6]. The graph matching has been previously used in matching and averaging heterogenous tree structures such as brain artery trees and neuronal trees [25].

Suppose we have weighted graphs $\mathcal{X}_1 = (V_1, w^1)$ and $\mathcal{X}_1 = (V_2, w^2)$, and corresponding 0D persistent diagrams P_1^0 and P_2^2 and 1D persistent diagrams P_1^1 and P_2^2. We define the Wasserstein distance between graphs $\mathcal{X}_1$ and $\mathcal{X}_2$ as the Wasserstein distance between corresponding persistent diagrams P_1 and P_2:

$$D_{Wj}(\mathcal{X}_1, \mathcal{X}_2) = D_{Wj}(P_1^j, P_2^j).$$

The 0D Wasserstein distance matches birth edges while the 1D Wasserstein distance matches death edges. We need to use both distances together to match graphs. Thus, we use the squared sum of 0D and 1D Wasserstein distances

$$\mathcal{D}(\mathcal{X}_1, \mathcal{X}_2) = D_{W0}^2(\mathcal{X}_1, \mathcal{X}_2) + D_{W1}^2(\mathcal{X}_1, \mathcal{X}_2)$$

as the Wasserstein distance between graphs in the study. Then we can show the distance is translation and scale invariant in the following sense:

$$\begin{aligned} \mathcal{D}(c + \mathcal{X}_1, c + \mathcal{X}_2) &= \mathcal{D}(\mathcal{X}_1, \mathcal{X}_2), \\ \frac{1}{c}\mathcal{D}(c\mathcal{X}_1, c\mathcal{X}_2) &= \mathcal{D}(\mathcal{X}_1, \mathcal{X}_2). \end{aligned}$$

Unlike existing computationally demanding graph matching algorithms, the method is scalable at $O(q \log q)$ run time. The majority of runtime is on sorting edge weights and obtaining the corresponding maximum spanning trees (MST).

3.2 Wasserstein graph mean

Given a collection of graphs $\mathcal{X}_1 = (V, w^1), \cdots, \mathcal{X}_n = (V, w^n)$ with edge weights $w^k = (w_{ij}^k)$, the usual approach for obtaining the average network $\bar{\mathcal{X}}$ is simply averaging the edge weigth matrices in an element-wise fashion

$$\bar{\mathcal{X}} = \Big(V, \frac{1}{n}\sum_{k=1}^{n} w_{ij}^k\Big).$$

However, such average is the average of the connectivity strength. It is not necessarily the average of underlying topology. Such an approach is usually sensitive to topological outliers [17]. We address the problem through the Wasserstein distance. A similar concept was proposed in persistent homology literature through the Wasserstein barycenter [2, 19], which is motivated by Fréchet mean [30, 52]. However, the method has not seen many applications in modeling graphs and networks.

With Theorem 3, we define the *Wasserstein graph sum* of graphs $\mathcal{X}_1 = (V, w^1)$ and $\mathcal{X}_2 = (V, w^2)$ as $\mathcal{X}_1 + \mathcal{X}_2 = (V, w)$ with the birth-death decomposition $W_b \cup W_d$

satisfying

$$W_b \cup W_d = (W_{1b} + W_{2b}) \cup (W_{1d} + W_{2d}).$$

with

$$w = \mathcal{F}(W_b \cup W_d).$$

However, the sum is not uniquely defined. Thus, the average of two graphs is also not uniquely defined. The situation is analogous to Fréchet mean, which often does not yield the unique mean [30, 52]. However, this is not an issue since their topology is uniquely defined and produces identical persistent diagrams. Now, we define the *Wasserstein graph mean* $\mathbb{E}\mathcal{X}$ of $\mathcal{X}_1, \cdots, \mathcal{X}_n$ as

$$\mathbb{E}\mathcal{X} = \frac{1}{n}\sum\nolimits_{k=1}^{n} \mathcal{X}_k. \tag{12}$$

The Wasserstein graph mean is the minimizer with respect to the Wasserstein distance, which is analogous to the sample mean as the minimizer of Euclidean distance. However, the Wasserstein graph mean is not unique in geometric sense. It is only unique in topological sense.

Theorem 5 *The Wasserstein graph mean is the graph given by*

$$\mathbb{E}\mathcal{X} = \arg\min_{X} \sum_{i=1}^{n} \mathcal{D}(X, \mathcal{X}_i).$$

Proof Since the cost function is a linear combination of quadratic functions, the global minimum exists and unique. Let $X = (V, W_b \cup W_d)$ be the birth-death decomposition with $W_b = \{b_{(1)}, \cdots, b_{(q_0)}\}$ and $W_d = \{d_{(1)}, \cdots, d_{(q_1)}\}$. From Theorem 4,

$$\sum_{i=1}^{n} \mathcal{D}(X, \mathcal{X}_i) = \sum_{i=1}^{n} \Big[\sum_{i=1}^{q_0} (b_{(i)} - b_{(i)}^k)^2 + \sum_{i=1}^{q_1} (d_{(i)} - d_{(i)}^k)^2 \Big].$$

This is quadratic so the minimum is obtained by setting its partial derivatives with respect to $b_{(i)}$ and $d_{(i)}$ equal to zero:

$$b_{(i)} = \frac{1}{n}\sum_{k=1}^{n} b_{(i)}^k, \quad d_{(i)} = \frac{1}{n}\sum_{k=1}^{n} d_{(i)}^k.$$

Thus, we obtain

$$W_b = \frac{1}{n}\sum_{k=1}^{n} W_{kb}, \quad W_d = \frac{1}{n}\sum_{k=1}^{n} W_{kd}.$$

This is identical to the birth-death decomposition of $\frac{1}{n}\sum_{k=1}^{n} \mathcal{X}_k$ and hence proves the statement. □

The *Wasserstein graph variance* $\mathbb{V}\mathcal{X}$ is defined in a similar fashion:

$$\mathbb{V}\mathcal{X} = \frac{1}{n}\sum_{i=1}^{n} \mathcal{D}(\mathbb{E}\mathcal{X}, \mathcal{X}_i),$$

which is interpreted as the variability of graphs from the Wassterstein graph mean $\mathbb{E}\mathcal{X}$. We can rewrite the Wasserstein graph variance as

$$\begin{aligned}\mathbb{V}\mathcal{X} &= \frac{1}{n}\sum_{i=1}^{n} \mathcal{D}\Big(\frac{1}{n}\sum_{j=1}^{n}\mathcal{X}_j, \mathcal{X}_i\Big) \\ &= \frac{1}{n^2}\sum_{i,j=1}^{n} \mathcal{D}(\mathcal{X}_i, \mathcal{X}_j). \end{aligned} \tag{13}$$

The formulation (13) compute the variance using the pairwise distances without the need for computing the Wasserstein graph mean.

3.3 Wasserstein graph clustering

There are few studies that used the Wasserstein distance for clustering [36, 58]. The existing methods are mainly applied to geometric data without topological consideration. It is not obvious how to apply the method to cluster graph data. We propose to use the Wasserstein graph matching method to cluster collection of graphs $\mathcal{X}_1, \cdots, \mathcal{X}_n$ into k clusters $C_1, \cdots, C_k$ such that

$$\cup_{i=1}^{k} C_i = \{\mathcal{X}_1, \cdots, \mathcal{X}_n\}, \quad C_i \cap C_j = \emptyset.$$

The total number of ways of partitioning n data points into k nonempty clusters is the *Stirling number of the second kind* $S_{n,k}$ [35]. There are $S_{n,1} = 1$ 1-clusters, $S_{n,2} = 2^{n-1} - 1$ 2-clusters and $S_{n,n-1} = \frac{n(n-1)}{2}$ possible $(n-1)$-clusters out of n data points. Asymptotically $S_{n,k}$ increases exponentially as [35]

$$S_{n,k} \sim \frac{k^n}{k!}.$$

Brute-force approaches for searching for every possible clusters is not feasible for large n. We propose to a more scalable approach for clustering.

Let $C = (C_1, \cdots, C_k)$ be the collection of clusters. Let μ_j be the *Wasserstein cluster mean* given by

$$\mu_j = \frac{1}{|C_j|}\sum_{X \in C_j} X$$

with $|C_j|$ number of elements in the cluster C_j. The cluster mean is computed through the birth-death decomposition using (12). Let $\mu = (\mu_1, \cdots, \mu_k)$ be the cluster mean vector. The within-cluster distance is given by

$$l_W(C;\mu) = \sum_{j=1}^{k} \sum_{X \in C_j} \mathcal{D}(X, \mu_j)., \tag{14}$$

which can be also written as

$$l_W(C;\mu) = \sum_{j=1}^{k} |C_j| \mathbb{V}_j \mathcal{X}, \tag{15}$$

where $\mathbb{V}_j \mathcal{X} = \frac{1}{|C_j|} \sum_{X \in C_j} \mathcal{D}(X, \mu_j)$ is the Wasserstein graph variance within cluster C_j. The optimal cluster is found by minimizing $l_W(C)$ in (14) over every possible C. If μ is given and fixed, the identification of clusters C can be done easily by assigning each network to the closest mean. Thus the Wasserstein clustering algorithm can be written as the two-step optimization similar to the expectation maximization (EM) algorithm often used in variational inferences and likelihood methods [7]. The first step computes the cluster mean. The second step minimizes the within-cluster distance. The two-step optimization is then iterated till convergence. Such process converges locally.

Theorem 6 *The Wasserstein graph clustering algorithm converges locally.*

Proof In the **expectation step**, we compute the cluster mean. Assume $C = (C_1, \cdots, C_k)$ is estimated from the previous iteration. In the current iteration, the cluster mean μ corresponding to C is updated as

$$\mu_j \leftarrow \frac{1}{|C_j|} \sum_{X \in C_j} X.$$

for each j. From Theorem 5, the cluster mean gives the lowest bound on function $l_W(C;\nu)$ for any $\nu = (\nu_1, \cdots, \nu_k)$:

$$l_W(C;\mu) = \sum_{j=1}^{k} \sum_{X \in C_j} \mathcal{D}(X, \mu_j) \leq \sum_{j=1}^{k} \sum_{X \in C_j} \mathcal{D}(X, \nu_j) = l_W(C;\nu). \tag{16}$$

In each iteration, we check if the cluster mean μ is changed from the previous iteration. If not, the algorithm simply stops. Thus we can force $l_W(C;\nu)$ to be strictly decreasing over each iteration.

In the **minimization step**, the clusters are updated from C to $C' = (C'_{J_1}, \cdots, C'_{J_k})$ by reassigning each graph $\mathcal{X}_i$ to the closest cluster C_{J_i}, i.e.,

$$J_i = \arg\min_j D_W(\mathcal{X}_i, \mu_j).$$

Subsequently, we have

$$l_W(C';\mu) = \sum_{j=1}^{k} \sum_{X \in C'_{J_i}} D_W(\mathcal{X}_i, \mu_{J_i}) \leq \sum_{j=1}^{k} \sum_{X \in C_j} D_W(\mathcal{X}_i, \mu_j) = l_W(C;\mu). \tag{17}$$

From (16) and (17), $l_W(C;\mu)$ strictly decreases over iterations. Any bounded strictly decreasing sequence converges. □

Numerical implementation. Just like k-means clustering algorithm that converges only to local minimum, there is no guarantee the Wasserstein graph clustering converges to the global minimum [28]. This is remedied by repeating the algorithm multiple times with different random seeds .

3.4 Clustering as a linear assignment problem

Let y_i be the true cluster label for the i-th data. Let $\widehat{y}_i$ be the estimate of y_i we determined from Wasserstein graph clustering. Let $y = (y_1, \cdots, y_n)$ and $\widehat{y} = (\widehat{y}_1, \cdots, \widehat{y}_n)$. In clustering, there is no direct association between true clustering labels and predicted cluster labels. Given k clusters $C_1, \cdots, C_k$, its permutation $\pi(C_1), \cdots, \pi(C_k)$ is also a valid cluster for $\pi \in \mathbb{S}_k$, the permutation group of order k. There are $k!$ possible permutations in $\mathbb{S}_k$ [18]. The clustering accuracy $A(y, \widehat{y})$ is then given by

$$A(\widehat{y}, y) = \frac{1}{n} \max_{\pi \in \mathbb{S}_k} \sum_{i=1}^{n} \mathbf{1}(\pi(\widehat{y}) = y).$$

This a modification to an assignment problem and can be solved using the Hungarian algorithm in $O(k^3)$ run time [22]. Let $C(\widehat{y}, y)$ be the confusion matrix of size $k \times k$ tabulating the correct number of clustering in each cluster. The diagonal entries show the correct number of clustering while the off-diagonal entries show the incorrect number of clusters. To compute the clustering accuracy, we need to sum the diagonal entries. Under the permutation of cluster labels, we can get different confusion matrices. For large k, it is prohibitive expensive to search for all permutations. Thus we need to maximize the sum of diagonals of the confusion matrix under permutation with weight $C = (c_{ij})$:

$$\frac{1}{n} \max_{Q \in \mathbb{S}_k} \mathrm{tr}(QC) = \frac{1}{n} \max_{Q \in \mathbb{S}_k} \sum_{i,j} q_{ij} c_{ij}, \tag{18}$$

where $Q = (q_{ij})$ is the permutation matrix consisting of entries 0 and 1 such that there is exactly single 1 in each row and each column. This is a linear sum assignment problem (LSAP), a special case of linear assignment problem [33].

4 Application to functional brain networks

The proposed method is applied in the accurate estimation of state spaces in dynamically changing functional brain networks. The 479 subjects resting-state functional magnetic resonance images (rs-fMRI) used in this paper were collected on a 3T

MRI scanner (Discovery MR750, General Electric Medical Systems, Milwaukee, WI, USA) with a 32-channel RF head coil array. The 479 healthy subjects consist of 231 males and 248 females ranging in age from 13 to 25 years were used. The image acquisition and preprocessing details are given in [8]. After preprocessing, which include motion corrections and image alignment to the template, the resulting rs-fMRI consist of $91 \times 109 \times 91$ isotropic voxels at 295 time points. We parcellated the brain volume into 116 non-overlapping brain regions from a widely used atlas [53]. The fMRI data were averaged across voxels within each brain region, resulting in 116 average fMRI signals with 295 time points for each subject. The rs-fMRI signals were then scaled to fit to unit interval [0, 1] and treated as functional data in $[0, 1]$.

4.1 Weighted Fourier series representation

The most common approach in computing time-varying correlation in time series data is through SW, where correlations between brain regions are computed over the windows [3, 29, 45, 37, 28]. However, the use of discrete windows can induce unnecessary high-frequency fluctuations in dynamic correlations [38], though in some cases tapering can mitigate this effect [3]. Further, correlation computation within windows is sensitive to outliers [20].

To address these problems, we performed the Weighted Fourier series (WFS) representation that generalizes the cosine Fourier transform with the additional exponential weight that smooths out high frequency noises while reducing the Gibbs phenomenon [12, 26]. WFS further avoids using sliding windows (SW) in computing correlations over time. For persistent homology method to work robustly across different subjects and time points, such signal denoising methods are needed. Consider arbitrary noise signal $f(t), t \in [0, 1]$ which will be denoised through diffusion.

Theorem 7 *The unique solution to 1D heat diffusion:*

$$\frac{\partial}{\partial s}h(t,s) = \frac{\partial^2}{\partial t^2}h(t,s) \tag{19}$$

on unit interval $[0, 1]$ *with initial condition* $h(t, s = 0) = f(t)$ *is given by WFS:*

$$h(t,s) = \sum_{l=0}^{\infty} e^{-l^2\pi^2 s} c_{fl}\psi_l(t), \tag{20}$$

where $\psi_0(t) = 1$, $\psi_l(t) = \sqrt{2}\cos(l\pi t)$ *are the cosine basis and* $c_{fl} = \int_0^1 f(t)\psi_l(t)dt$ *are the expansion coefficients.*

The algebraic derivation is given in [12]. Note the cosine basis is orthonormal

$$\langle \psi_l, \psi_m \rangle = \int_0^1 \psi_l(t)\psi_m(t)\, dt = \delta_{lm},$$

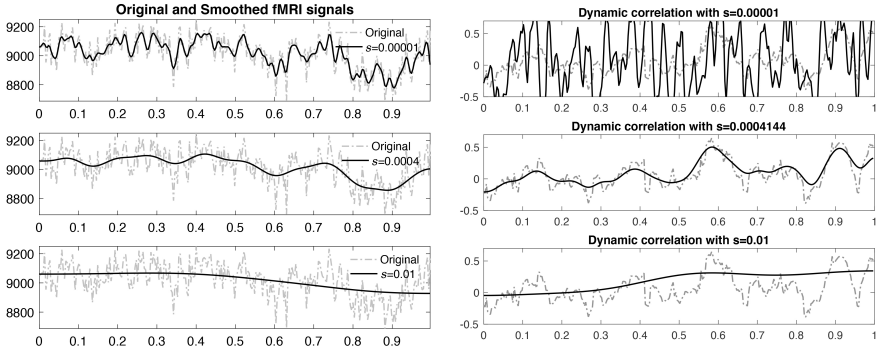

Fig. 4 Left: The original and smoothed fMRI time series using WFS with degree $L = 295$ and different heat kernel bandwidth s. The bandwidth 4.141×10^{-4} is used in this study approximately matches 20 TRs often used in the sliding window methods. Right: Doted gray lines are correlations computed over sliding windows. The solid black lines are correlations computed using WFS.

where δ_{lm} is Kroneker-detal taking value 1 if $l = m$ and 0 otherwise. We can rewrite (20) as a more convent convolution form

$$h(t,s) = \int_0^1 K_s(t,t')f(t')dt',$$

where heat kernel $K_s(t,t')$ is given by

$$K_s(t,t') = \sum_{l=0}^{\infty} e^{-l^2\pi^2 s}\psi_l(t)\psi_l(t').$$

The diffusion time s is usually referred to as the kernel bandwidth and controls the amount of smoothing. Heat kernel satisfies $\int_0^1 K_s(t,t')\ dt = 1$ for any t' and s.

To reduce unwanted boundary boundary effects near the data boundary $t = 0$ and $t = 1$ [26, 28], we project the data onto the circle C with circumference 2 by the mirror reflection:

$$g(t) = f(t) \text{ if } t \in [0,1], \quad g(t) = f(2-t) \text{ if } t \in [1,2].$$

Then perform WFS on the circle.

Theorem 8 *The unique solution to 1D heat diffusion:*

$$\frac{\partial}{\partial s}h(t,s) = \frac{\partial^2}{\partial t^2}h(t,s) \tag{21}$$

on the circle C with the initial periodic condition $h(t, s = 0) = f(t)$ if $t \in [0,1]$, $h(t, s = 0) = f(2-t)$ if $t \in [1,2]$ is given by WFS:

$$h(t,s) = \sum_{l=0}^{\infty} e^{-l^2\pi^2 s} c_{fl}\psi_l(t), \tag{22}$$

where $\psi_0(t) = 1$, $\psi_l(t) = \sqrt{2}\cos(l\pi t)$ *are the cosine basis and* $c_{fl} = \int_0^1 f(t)\psi_l(t)dt$ *are the expansion coefficients.*

Proof The cosine basis is defined on interval $[0,1]$. We extend the domain of the basis by mirror reflection $\tilde{\psi}(t) = \psi(t)/\sqrt{2}$ in $[0,1]$ and $\tilde{\psi}(t) = \psi(2-t)/\sqrt{2}$ for $t \in [1,2]$. Since $\tilde{\psi}(2) = \tilde{\psi}(0)$, the extended basis $\tilde{\psi}$ is a proper basis on circle C. The basis is scaled to have orthonormality:

$$\langle\tilde{\psi}_l, \tilde{\psi}_m\rangle = \int_0^1 \psi_l(t)\psi_m(t)\ dt + \int_1^2 \psi_l(2-t)\psi_m(2-t)\ dt = \delta_{lm}.$$

Subsequently, we can also extend the heat kernel as $\tilde{K}_s(t,t') = K_s(t,t')/2$ if $t' \in [0,1]$ and $\tilde{K}_s(t,t') = K_s(t,2-t')/2$ if $t' \in [1,2]$. The extended heat kernel satisfies

$$\int_0^2 K_s(t,t')dt' = \int_0^1 K_s(t,t')dt' + \int_1^2 K_s(t,2-t')dt' = 1.$$

Then, the solution to (21) is given by heat kernel convolution [12]

$$\begin{aligned} h(t,s) &= \int_0^2 \tilde{K}_s(t,t')h(t',s=0)\ dt' \\ &= \int_0^1 \frac{1}{2}K_s(t,t')f(t')dt' + \int_1^2 \frac{1}{2}K_s(t,2-t')f(2-t')dt' \\ &= \int_0^1 K_s(t,t')g(t')dt'. \end{aligned}$$

Hence, heat kernel smoothing on the circle with mirror reflection symmetry can be simply done by applying WFS in unit interval $[0,1]$. □

Numerical implementation. The cosine series coefficients c_{fl} are estimated using the least squares method by setting up a matrix equation [12]. We set the expansion degree to equate the number of time points, which is 295. The window size of 20 TRs were used in most sliding window methods [3, 34, 28]. We matched the full width at half maximum (FWHM) of heat kernel to the window size numerically. We used the fact that diffusion time s in heat kernel approximately matches to the kernel bandwidth of Gaussian kernel $e^{-t^2/2\sigma^2}$ as $\sigma = s^2/2$ (page 144 in [11]). 20 TRs is approximately equivalent to heat kernel bandwidth of about $4.144 \cdot 10^{-4}$ in terms of FWHM. Figure 4 displays the WFS representation of rsfMRI with different kernel bandwidths.

4.2 Dynamic correlation on weighted Fourier series

The weighted Fourier series representation provides a way to compute correlations dynamically without using sliding windows. Consider time series $x(t)$ and $y(t)$ with heat kernel $K_s(t,t')$. The mean and variance of signals with respect to the heat kernel are given by

$$\mathbb{E}x(t) = \int_0^1 K_s(t,t')x(t')\ dt'. \quad \mathbb{V}x(t) = \int_0^1 K_s(t,t')x^2(t')\ dt' - \big[\mathbb{E}x(t)\big]^2$$

Subsequently, the correlation $w(t)$ of $x(t)$ and $y(t)$ is given by

$$w(t) = \frac{\int_0^1 K_s(t,t')x(t')y(t')\ dt - \mathbb{E}x(t)\mathbb{E}y(t)}{\sqrt{\mathbb{V}x(t)}\sqrt{\mathbb{V}y(t)}}.$$

When the kernel is shaped as a sliding window, the correlation $w(t)$ exactly matches the correlation computed over the sliding window. The kernelized correlation generalizes the concept of integral correlations with the additional weighting term [27]. As $s \to \infty$, $w(t)$ converges to the Pearson correlation computed over the whole time points. Thus, the kernel bandwidth behaves like the length of sliding window.

Theorem 9 *The correlation $w(t)$ of time series $x(t)$ and $y(t)$ with respect to heat kernel $K_s(t,t')$ is given by*

$$w(t) = \frac{\sum_{l=0}^{\infty} e^{-l^2\pi^2 s} c_{xyl}\psi_l(t) - \mu_x(t)\mu_y(t)}{\sigma_x(t)\sigma_y(t)}, \tag{23}$$

with

$$\mu_x(t) = \sum_{l=0}^{\infty} e^{-l^2\pi^2 s} c_{xl}\psi_l(t), \quad \sigma_x^2(t) = \sum_{l=0}^{\infty} e^{-l^2\pi^2 s} c_{xxl}\psi_l(t) - \mu_x^2(t).$$

$$c_{xl} = \int_0^1 x(t)\psi_l(t)dt, \quad c_{yl} = \int_0^1 y(t)\psi_l(t)dt$$

are the cosine series coefficients. Similarly we expand $x(t)y(t)$, $x^2(t)$ and $y^2(t)$ using the cosine basis and obtain coefficients c_{xyl}, c_{xxl} and c_{yyl}.

The derivation follows by simply replacing all the terms with the WFS representation. Correlation (23) is the formula we used to compute the dynamic correlation in this study. Figure 4 displays the WFS-based dynamic correlation for different bandwidths. A similar weighted correlation was proposed in [40], where time varying exponential weights proportional to $e^{t/\theta}$ with exponential decay factor θ. However, our exponential weight term is related to the spectral decomposition of heat kernel in the spectral domain and invariant over time. The WFS based correlation is not related to [40].

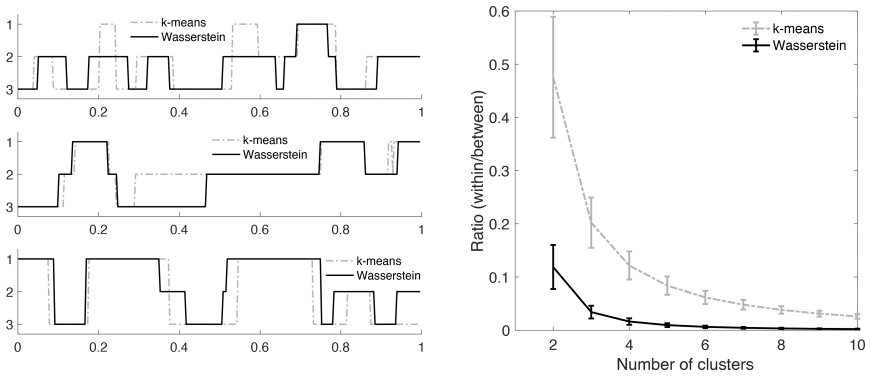

Fig. 5 Left: The time series of estimated state spaces using the Wasserstein clustering and k-means clustering for 3 subjects. The time is normalized into unit interval $[0, 1]$. Right: The ratio of within-cluster to between-cluster distances. Smaller the ratio, better the clustering fit is.

4.3 Estimation of distinct state space in dynamic connectivity

For p brain regions, we estimated $p \times p$ dynamically changing correlation matrices $C_i(t)$ for the i-th subject using WFS. Let $\mathbf{C}_{ij}$ denote the vectorization of the upper triangle of $p \times p$ matrix $C_i(t_j)$ at time point t_j into $p^2 \times 1$ vector. The collection of $\mathbf{C}_{ij}$ over $T = 295$ time points and $n = 479$ subjects is then feed into Wasserstein clustering in identifying the recurring brain connectivity states that is common across subjects at the group level. We compared the proposed Wassertein clustering against the k-means clustering, which has been often used baseline method in the state space modeling [3, 26, 28]. After clustering, each correlation matrix $C_i(t_j)$ is assigned integers between 1 and k. These discrete states serve as the basis of investigating the dynamic pattern brain connectivity [51]. For the convergence of both Wasserstein and k-means clustering, the clusterings were repeated 10 times with different initial centroids and the average results are reported. Figure 5-left displays the result of the Wasserstein clustering against the k-means in few brain regions for a subject. 295 time points are rescaled to fit into unit interval $[0, 1]$.

The optimal number of cluster k was determined by the *elbow method* [3, 42, 51, 28]. For each value of k, we computed the ratio of the within-cluster to between-cluster distances. The ratio shows the goodness-of-fit of the cluster model. The optimal number of clusters were determined by the elbow method, which gives the largest slope change in the ratio. $k = 3$ gives the largest slope in the both methods (Figure 5-right). At $k = 3$, the ratio is 0.034 ± 0.012 for 479 subjects for Wasserstein while it is 0.202 ± 0.047 for the k-means. The *six* times smaller ratio for the Wasserstein clustering demonstrates the superior model fit over k-means. Figure 6 shows the results of clustering. The dynamic change of states can be viewed as a Markov chain [24, 4, 28]. For the Wasserstein clustering, the probabilities of staying in the states 1, 2 and 3 are 0.97, 0.96 and 0.97 respectively. for the k-means clustering,

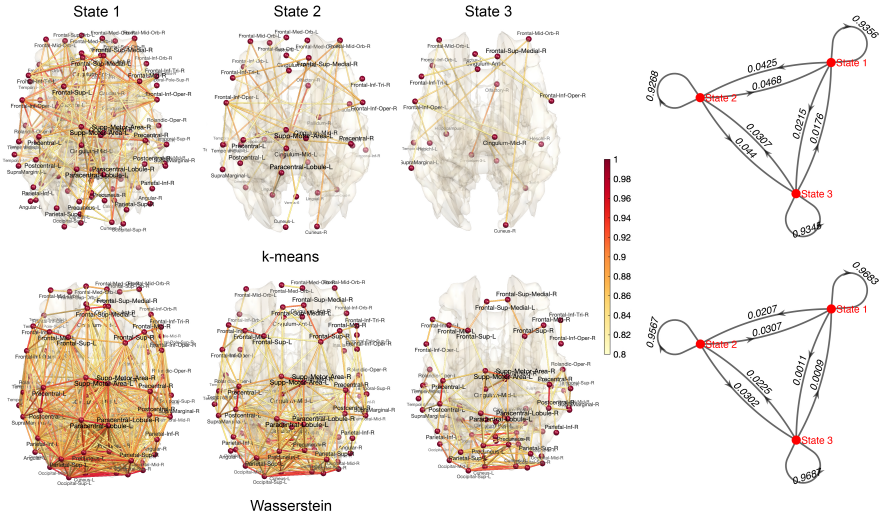

Fig. 6 Left: The average estimated state spaces of dynamically changing brain networks. Right: The change of state spaces modeled as a Markov chain with transition probabilities. The Wasserstein clustering results significantly increases the transition probability to stay in the same state.

the probabilities of staying in the states 1, 2 and 3 are 0.94, 0.93 and 0.93 respectively. These few percentage differences significantly increases the Wasserstein clustering model fit. We believe our method provides more accurate results.

5 Conclusion

In this study, the proposed the Wasserstein graph clustering for estimation and quantification of dynamic state changes in time varying networks. We developed a coherent statistical theory based on persistent homology and presented how such method is applied to the resting state fMRI data. The resting-state brain networks tend to remain in the same state for a long period before the transition to another state [3, 45, 10]. The average brain network in each state (Figure 6) does not follow similar connectivity patterns observed in the previous studies [9]. But further research is needed for independent validation.

Acknowledgement

This study was supported by NIH grants EB022856, MH101504, P30HD003352, U54HD09025, UL1TR002373 and NSF grant MDS-2010778. We would like to thank Chee-Ming Ting and Hernando Ombao of KAUST for discussion on *k*-

means clustering. We also like to thank Tananun Songdechakraiwut of University of Wisconsin-Madison and Botao Wang of Xi'an Jiaotong University for discussion on Wasserstein graph clustering.

References

1. Adler, R., Bobrowski, O., Borman, M., Subag, E. and Weinberger, S. (2010). Persistent homology for random fields and complexes. *In: Borrowing strength: theory powering applications–a Festschrift for Lawrence D. Brown*Institute of Mathematical Statistics. 124–143.
2. Agueh, M. and Carlier, G. (2011). Barycenters in the wasserstein space. *SIAM Journal on Mathematical Analysis* **43** 904–924.
3. Allen, E., Damaraju, E., Plis, S., Erhardt, E., Eichele, T. and Calhoun, V. (2014). Tracking whole-brain connectivity dynamics in the resting state. *Cerebral cortex* **24** 663–676.
4. Baker, A., Brookes, M., Rezek, I., Smith, S., Behrens, T., Smith, P. and Woolrich, M. (2014). Fast transient networks in spontaneous human brain activity. *Elife* **3** e01867.
5. Bassett, D. and Sporns, O. (2017). Network neuroscience. *Nature neuroscience* **20** 353–364.
6. Becker, C., Pequito, S., Pappas, G., Miller, M., Grafton, S., Bassett, D. and Preciado, V. (2018). Spectral mapping of brain functional connectivity from diffusion imaging. *Scientific reports* **8** 1–15.
7. Bishop, C. (2006). *Pattern recognition and machine learning*. springer.
8. Burghy, C., Fox, M., Cornejo, M., Stodola, D., Sommerfeldt, S., Westbrook, C., Van Hulle, C., Schmidt, N., Goldsmith, H., Davidson, R. et al. (2016). Experience-driven differences in childhood cortisol predict affect-relevant brain function and coping in adolescent Monozygotic twins. *Scientific Reports* **6** 37081.
9. Cai, B., Zille, P., Stephen, J., Wilson, T., Calhoun, V. and Wang, Y. (2018). Estimation of dynamic sparse connectivity patterns from resting state fMRI. *IEEE Transactions on Medical Imaging* **37** 1224–1234.
10. Calhoun, V. and Adali, T. (2016). Time-varying brain connectivity in fMRI data: whole-brain data-driven approaches for capturing and characterizing dynamic states. *IEEE Signal Processing Magazine* **33** 52–66.
11. Chung, M. (2012). *Computational Neuroanatomy: The Methods*. World Scientific, Singapore.
12. Chung, M., Dalton, K., Shen, L., Evans, A. and Davidson, R. (2007). Weighted Fourier representation and its application to quantifying the amount of gray matter. *IEEE Transactions on Medical Imaging* **26** 566–581.
13. Chung, M., Hanson, J., Lee, H., Adluru, N., Alexander, A. L., Davidson, R. and Pollak, S. (2013). Persistent homological sparse network approach to detecting white matter abnormality in maltreated children: MRI and DTI multimodal study. *MICCAI, Lecture Notes in Computer Science (LNCS)* **8149** 300–307.
14. Chung, M., Hanson, J., Adluru, L., Alexander, A., Davidson, R. and Pollak, S. (2017a). Integrative structural brain network analysis in diffusion tensor imaging. *Brain Connectivity* **7** 331–346.
15. Chung, M., Lee, H., Solo, V., Davidson, R. and Pollak, S. (2017b). Topological distances between brain networks. *International Workshop on Connectomics in Neuroimaging* **10511** 161–170.
16. Chung, M., Luo, Z., Leow, A., Adluru, A., Alexander, A., Richard, D. and Goldsmith, H. (2018). Exact combinatorial inference for brain images. *International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)* **11070** 629–637.
17. Chung, M., Huang, S.-G., Gritsenko, A., Shen, L. and Lee, H. (2019a). Statistical inference on the number of cycles in brain networks. *In: 2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019)*. IEEE 113–116.

18. Chung, M., Xie, L., Huang, S.-G., Wang, Y., Yan, J. and Shen, L. (2019b). *Rapid acceleration of the permutation test via transpositions* **11848** 42–53.
19. Cuturi, M. and Doucet, A. (2014). Fast computation of Wasserstein barycenters. *In: International conference on machine learning*. PMLR 685–693.
20. Devlin, S., Gnanadesikan, R. and Kettenring, J. (1975). Robust estimation and outlier detection with correlation coefficients. *Biometrika* **62** 531–545.
21. Edelsbrunner, H. and Harer, J. (2010). *Computational topology: An introduction*. American Mathematical Society.
22. Edmonds, J. and Karp, R. (1972). Theoretical improvements in algorithmic efficiency for network flow problems. *Journal of the ACM (JACM)* **19** 248–264.
23. Ghrist, R. (2008). Barcodes: The persistent topology of data. *Bulletin of the American Mathematical Society* **45** 61–75.
24. Gilks, W., Richardson, S. and Spiegelhalter, D. (1995). *Markov chain Monte Carlo in practice*. Chapman and Hall/CRC.
25. Guo, X. and Srivastava, A. (2020). Representations, metrics and statistics for shape analysis of elastic graphs. *In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops* 832–833.
26. Huang, S.-G., Chung, M. K., Carroll, I. C. and Goldsmith, H. H. (2019a). Dynamic functional connectivity using heat kernel. *In: 2019 IEEE Data Science Workshop (DSW)* 222–226. , DOI 10.1109/DSW.2019.8755550
27. Huang, S.-G., Gritsenko, A., Lindquist, M. and Chung, M. (2019b). Circular pearson correlation using cosine series expansion. *In: IEEE 16th International Symposium on Biomedical Imaging (ISBI)* 1774–1777.
28. Huang, S.-G., Samdin, S.-T., Ting, C., Ombao, H. and Chung, M. (2020). Statistical model for dynamically-changing correlation matrices with application to brain connectivity. *Journal of Neuroscience Methods* **331** 108480.
29. Hutchison, R., Womelsdorf, T., Allen, E., Bandettini, P. and Calhoun, V. e. a. (2013). Dynamic functional connectivity: promise, issues, and interpretations. *NeuroImage* **80** 360–378.
30. Le, H. and Kume, A. (2000). The Fréchet mean shape and the shape of the means. *Advances in Applied Probability* **32** 101–113.
31. Lee, H., Chung, M., Kang, H., Kim, B.-N. and Lee, D. (2011). Computing the shape of brain networks using graph filtration and Gromov-Hausdorff metric. *MICCAI, Lecture Notes in Computer Science* **6892** 302–309.
32. Lee, H., Kang, H., Chung, M., Kim, B.-N. and Lee, D. (2012). Persistent brain network homology from the perspective of dendrogram. *IEEE Transactions on Medical Imaging* **31** 2267–2277.
33. Lee, M., Xiong, Y., Yu, G. and Li, G. Y. (2018). Deep neural networks for linear sum assignment problems. *IEEE Wireless Communications Letters* **7** 962–965.
34. Lindquist, M. (2014). Statistical and computational methods in brain image analysis. by Moo K. Chung. Boca Raton, Florida: CRC press. 2013. *Journal of the American Statistical Association* **109** 1334–1335.
35. Lord, E., Willems, M., Lapointe, F.-J. and Makarenkov, V. (2017). Using the stability of objects to determine the number of clusters in datasets. *Information Sciences* **393** 29–46.
36. Mi, L., Zhang, W., Gu, X. and Wang, Y. (2018). Variational wasserstein clustering. *In: Proceedings of the European Conference on Computer Vision (ECCV)* 322–337.
37. Mokhtari, F., Akhlaghi, M., Simpson, S., Wu, G. and Laurienti, P. (2019). Sliding window correlation analysis: Modulating window shape for dynamic brain connectivity in resting state. *NeuroImage* **189** 655–666.
38. Oppenheim, A., Schafer, R. and Buck, J. (1999). *Discrete-time signal processing*. Upper Saddle River, NJ: Prentice Hall.
39. Petri, G., Expert, P., Turkheimer, F., Carhart-Harris, R., Nutt, D., Hellyer, P. and Vaccarino, F. (2014). Homological scaffolds of brain functional networks. *Journal of The Royal Society Interface* **11** 20140873.

40. Pozzi, F., Di Matteo, T. and Aste, T. (2012). Exponential smoothing weighted correlations. *The European Physical Journal B* **85** 1–21.
41. Rabin, J., Peyré, G., Delon, J. and Bernot, M. (2011). Wasserstein barycenter and its application to texture mixing. *In: International Conference on Scale Space and Variational Methods in Computer Vision*. Springer 435–446.
42. Rashid, B., Damaraju, E., Pearlson, G. and Calhoun, V. (2014). Dynamic connectivity states estimated from resting fMRI identify differences among schizophrenia, bipolar disorder, and healthy control subjects. *Frontiers in Human Neuroscience* **8** 897.
43. Sabbagh, D., Ablin, P., Varoquaux, G., Gramfort, A. and Engemann, D. (2019). Manifold-regression to predict from meg/eeg brain signals without source modeling. .
44. Santos, F., Raposo, E., Coutinho-Filho, M., Copelli, M., Stam, C. and Douw, L. (2019). Topological phase transitions in functional brain networks. *Physical Review E* **100** 032414.
45. Shakil, S., Lee, C.-H. and Keilholz, S. (2016). Evaluation of sliding window correlation performance for characterizing dynamic functional connectivity and brain states. *NeuroImage* **133** 111–128.
46. Shi, J., Zhang, W. and Wang, Y. (2016). Shape analysis with hyperbolic wasserstein distance. *In: Proceedings of the IEEE conference on computer vision and pattern recognition* 5051–5061.
47. Sizemore, A., Giusti, C., Kahn, A., Vettel, J., Betzel, R. and Bassett, D. (2018). Cliques and cavities in the human connectome. *Journal of computational neuroscience* **44** 115–145.
48. Songdechakraiwut, T. and Chung, M. (2020). Dynamic topological data analysis for functional brain signals. 1–4. .
49. Songdechakraiwut, T., Shen, L. and Chung, M. (2021). Topological learning and its application to multimodal brain network integration. *Medical Image Computing and Computer Assisted Intervention (MICCAI)* **12902** 166–176.
50. Sporns, O. (2003). *Graph Theory Methods for the Analysis of Neural Connectivity Patterns*. Springer US, Boston, MA. 171–185.
51. Ting, C.-M., Ombao, H., Samdin, S. and Salleh, S.-H. (2018). Estimating dynamic connectivity states in fMRI using regime-switching factor models. *IEEE transactions on Medical imaging* **37** 1011–1023.
52. Turner, K., Mileyko, Y., Mukherjee, S. and Harer, J. (2014). Fréchet means for distributions of persistence diagrams. *Discrete & Computational Geometry* **52** 44–70.
53. Tzourio-Mazoyer, N., Landeau, B., Papathanassiou, D., Crivello, F., Etard, O., Delcroix, N., Mazoyer, B. and Joliot, M. (2002). Automated anatomical labeling of activations in spm using a macroscopic anatomical parcellation of the MNI MRI single-subject brain. *NeuroImage* **15** 273–289.
54. Vallender, S. (1974). Calculation of the Wasserstein distance between probability distributions on the line. *Theory of Probability & Its Applications* **18** 784–786.
55. Wang, Y., Chung, M., Dentico, D., Lutz, A. and Davidson, R. (2017). Topological network analysis of electroencephalographic power maps. *In: International Workshop on Connectomics in NeuroImaging, Lecture Notes in Computer Science (LNCS)*. vol 10511 134–142.
56. Wang, Y., Ombao, H. and Chung, M. (2018). Topological data analysis of single-trial electroencephalographic signals. *Annals of Applied Statistics* **12** 1506–1534.
57. Xu, M., Sanz, D. L., Garces, P., Maestu, F., Li, Q. and Pantazis, D. (2021). A graph Gaussian embedding method for predicting Alzheimer's disease progression with MEG brain networks. *IEEE Transactions on Biomedical Engineering* **68** 1579–1588.
58. Yang, Z., Wen, J. and Davatzikos, C. (2020). Smile-GANs: Semi-supervised clustering via GANs for dissecting brain disease heterogeneity from medical images. .
59. Yoo, J., Kim, E., Ahn, Y. and Ye, J. (2016). Topological persistence vineyard for dynamic functional brain connectivity during resting and gaming stages. *Journal of neuroscience methods* **267** 1–13.
60. Yoo, K., Lee, P., Chung, M., Sohn, W., Chung, S., Na, D., Ju, D. and Jeong, Y. (2017). Degree-based statistic and center persistency for brain connectivity analysis. *Human Brain Mapping* **38** 165–181.
61. Zomorodian, A. (2009). *Topology for computing*. Cambridge University Press, Cambridge.