FAST TOPOLOGICAL CLUSTERING WITH WASSERSTEIN DISTANCE

Anonymous authors
Paper under double-blind review

ABSTRACT

The topological patterns exhibited by many real-world networks motivate development of topology-based methods for assessing the similarity of networks. However, extracting topological structure is difficult, especially for large and dense networks whose node degrees range over multiple orders of magnitude. In this paper, we propose a novel and computationally practical *topological clustering* method that clusters complex networks with intricate topology using principled theory from persistent homology and optimal transport. Such networks are aggregated into clusters through a centroid-based clustering strategy based on both their topological and geometric structure, preserving correspondence between nodes in different networks. The notions of topological proximity and centroid are characterized using a novel and efficient approach to computation of the Wasserstein distance and barycenter for persistence barcodes. The proposed method is demonstrated to be effective using both simulated networks and measured functional brain networks.

1 INTRODUCTION

Network models are extremely useful representations for complex data. Significant attention has been given to cluster analysis within a single network, such as detecting community structure (Newman, 2006; Rohe et al., 2011; Yin et al., 2017). Less attention has been given to clustering of collections of network representations. Clustering approaches typically group similar networks based on comparisons of edge weights (Xu & Wunsch, 2005), not topology. Assessing similarity of networks based on topological structure offers the potential for new insight, given the inherent topological patterns exhibited by most real-world networks. However, extracting meaningful network topology is a very difficult task, especially for large and dense networks whose node degrees range over multiple orders of magnitude (Barrat et al., 2004; Bullmore & Sporns, 2009; Honey et al., 2007).

Persistent homology (Edelsbrunner et al., 2000; Wasserman, 2018) has recently emerged as a powerful tool for understanding, characterizing and quantifying complex networks (Chung et al., 2019). Persistent homology represents a network using topological features such as connected components and cycles. Many networks naturally divide into modules or connected components (Bullmore & Sporns, 2009; Honey et al., 2007). Similarly, cycle structure is ubiquitous and is often used to describe information propagation, robustness and feedback mechanisms (Keizer et al., 1995; Kwon & Cho, 2007; Ozbudak et al., 2005; Venkatesh et al., 2004; Weiner et al., 2002). Effective use of such topological descriptors requires a notion of proximity that quantifies the similarity between persistence barcodes, a convenient representation for connected components and cycles (Ghrist, 2008). Wasserstein distance, which measures the minimal effort to modify one persistence barcode to another (Rabin et al., 2011), is an excellent choice due to its appealing geometric properties (Staerman et al., 2021) and its effectiveness shown in many machine learning applications (Kolouri et al., 2017; Mi et al., 2018; Solomon et al., 2015). Importantly, Wasserstein distance can be used to interpolate networks while preserving topological structure, and the mean under the Wasserstein distance, known as Wasserstein barycenter (Agueh & Carlier, 2011), can be viewed as the topological centroid of a set of networks.

The high cost of computing persistence barcodes, Wasserstein distance and the Wasserstein barycenter limit their applications to small scale problems, see, e.g., (Clough et al., 2020; Hu et al., 2019; Kolouri et al., 2017; Mi et al., 2018). Although approximation algorithms have been developed (Cuturi, 2013; Cuturi & Doucet, 2014; Li et al., 2020; Solomon et al., 2015; Vidal et al., 2019; Xie et al., 2020; Ye

et al., 2017), it is unclear whether these approximations are effective for clustering complex networks as they inevitably limit sensitivity to subtle topological features. Indeed, more and more studies, see, e.g., (Robins & Turner, 2016; Xia & Wei, 2014) have demonstrated that such subtle topological patterns are important for the characterization of complex networks, suggesting these approximation algorithms are undesirable.

Recently, it was shown that the Wasserstein distance and barycenter for persistence barcodes have closed-form solutions that can be computed exactly and efficiently by projecting the persistence barcodes into one dimension (Songdechakraiwut et al., 2021). Motivated by this result, we present a novel and computationally practical *topological clustering* method that clusters complex networks of the same size with intricate topological characteristics. Topological information alone is effective at clustering networks with no correspondence between nodes in different networks. However, when networks have meaningful node correspondence, we perform the cluster analysis using combined topological and geometric information to preserve node correspondence. Statistical validation based on ground truth information is used to demonstrate the effectiveness of our method when discriminating subtle topological features in simulated networks. The method is further illustrated by clustering measured functional brain networks associated with different levels of arousal during general anesthesia. Our proposed method outperforms other clustering approaches in both the simulated and measured data.

The paper is organized as follows. Background on one-dimensional projection of persistence barcodes is given in section 2, while section 3 presents our topological clustering method. In sections 4 and 5, we compare the performance of our method to several baseline algorithms using simulated and measured networks. Section 6 concludes the paper with a brief discussion of the potential impact of this work.

2 Projection of Persistence Barcodes

2.1 Graph filtration

Consider a network represented as a weighted graph $G = (V, \boldsymbol{w})$ comprising a set of nodes V with symmetric adjacency matrix $\boldsymbol{w} = (w_{ij})$, with edge weight w_{ij} representing the relationship between node i and node j. The number of nodes is denoted as $|V|$. The binary graph $G_\epsilon = (V, \boldsymbol{w_\epsilon})$ of G is defined as a graph consisting of the node set V and binary edge weights $w_{\epsilon,ij} = 1$ if $w_{ij} > \epsilon$ and $w_{ij} = 0$ otherwise. We view the binary network G_ϵ as a 1-skeleton, a simplicial complex comprising only nodes and edges (Munkres, 2018). In the 1-skeleton, there are two types of topological features: connected components and cycles. The number of connected components and the number of cycles in the binary network are referred to as the 0-th Betti number $\beta_0(G_\epsilon)$ and the 1-st Betti number $\beta_1(G_\epsilon)$, respectively. A *graph filtration* of G is defined as a collection of nested binary networks (Lee et al., 2012):

$$G_{\epsilon_0} \supseteq G_{\epsilon_1} \supseteq \cdots \supseteq G_{\epsilon_k},$$

where $\epsilon_0 \leq \epsilon_1 \leq \cdots \leq \epsilon_k$ are filtration values. As ϵ increases, more and more edges are removed from the network G since we threshold the edge weights at higher connectivity. For instance, $G_{-\infty}$ has each pair of nodes connected by an edge and thus is a complete graph consisting of a single connected component, while G_∞ has no edges and represents the node set. Figure 1 illustrates the graph filtration of a four-node network and the corresponding Betti numbers.

2.2 Birth-death decomposition

Persistent homology keeps track of birth and death of connected components and cycles over filtration values ϵ, and associates their *persistence* (life-time duration from birth to death) to them. The persistence is represented as a *persistence barcode* $PB(G)$ comprising intervals $[b_i, d_i]$ such that each tabulates the life-time of a connected component or a cycle that appears at the filtration value b_i and vanishes at d_i.

Over increasing filtration values, connected components are born while cycles die (Chung et al., 2019). Specifically, β_0 is monotonically increasing from $\beta_0(G_{-\infty}) = 1$ to $\beta_0(G_\infty) = |V|$. There are $\beta_0(G_\infty) - \beta_0(G_{-\infty}) = |V| - 1$ number of new connected components that are born over the filtration. Connected components will never die once they are born, implying that every connected

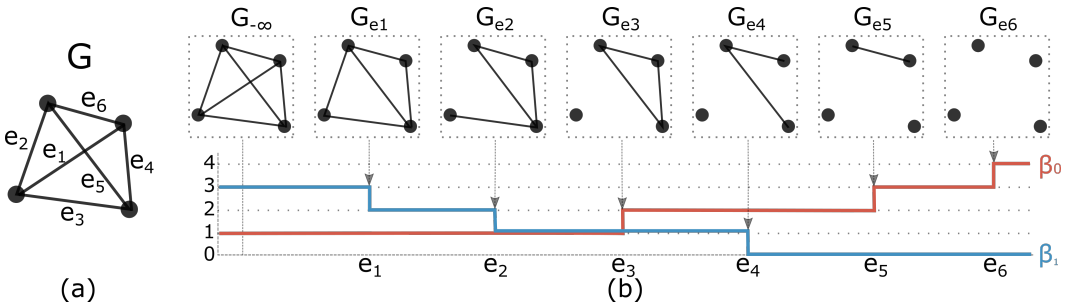

Figure 1: (a) Four-node network G. (b) As the filtration value increases, the number of connected components β_0 monotonically increases while the number of cycles β_1 monotonically decreases. Connected components are born at the edge weights e_3, e_5, e_6 while cycles die at the edge weights e_1, e_2, e_4.

component has death value at ∞. Thus, we can represent their persistence as a collection of finite birth values $B(G) = \{b_i\}_{i=1}^{|V|-1}$. On the other hand, all the cycles must be in $G_{-\infty}$, a complete graph and thus have birth values at $-\infty$. Again, we can represent the persistence of the cycles as a collection of finite death values $D(G) = \{d_i\}$. How many cycles are there? Since the deletion of an edge w_{ij} must result in either the birth of a connected component or the death of a cycle, every edge weight must be in either $B(G)$ or $D(G)$. Thus, the edge weight set $W = \{w_{ij} | i > j\}$ decomposes into the collection of birth values $B(G)$ and the collection of death values $D(G)$. Since $G_{-\infty}$ is a complete graph with $\frac{|V|(|V|-1)}{2}$ number of edges and there are $|V| - 1$ connected components, the number of cycles is then equal to $\frac{|V|(|V|-1)}{2} - (|V| - 1) = 1 + \frac{|V|(|V|-3)}{2}$. In the example of Figure 1, we have $B(G) = \{e_3, e_5, e_6\}$ and $D(G) = \{e_1, e_2, e_4\}$.

Finding birth values in $B(G)$ is *equivalent* to finding edge weights comprising the *maximum spanning tree* of G and can be done using well-known methods such as Prim's and Kruskal's algorithms (Lee et al., 2012). Once $B(G)$ is known, $D(G)$ is simply given as the remaining edge weights. Finding $B(G)$ and $D(G)$ requires only $O(n \log n)$ operations, where n is the number of edges in the network, and thus is extremely computationally efficient.

3 Method

3.1 Topological distance

Since the topology of a network is completely characterized by the persistence barcode of connected components and cycles, the topological dissimilarity between two networks can be measured using the 2-Wasserstein distance between their corresponding barcodes as follows (Cohen-Steiner et al., 2010; Rabin et al., 2011). Let G and H be two given networks that have the same number of nodes. The *topological distance* $d_{top}(G, H)$ is defined as the optimal matching cost:

$$\Big(\min_{\tau} \sum_{p \in PB(G)} ||p - \tau(p)||^2\Big)^{\frac{1}{2}} = \Big(\min_{\tau} \sum_{p=[b_p,d_p] \in PB(G)} \big[b_p - b_{\tau(p)}\big]^2 + \big[d_p - d_{\tau(p)}\big]^2\Big)^{\frac{1}{2}}, \quad (1)$$

where the optimization is over all possible bijections τ from barcode $PB(G)$ to barcode $PB(H)$. Intuitively, we can think of each interval $[b_i, d_i]$ as a point (b_i, d_i) in 2-dimensional plane and that the topological distance measures the minimal amount of work to move points in $PB(G)$ to $PB(H)$. Note this alternative representation of points in the plane is equivalent to the persistence barcode and called the *persistence diagram* (Edelsbrunner & Harer, 2008). Moving a connected component point (b_i, ∞) to a cycle point $(-\infty, d_j)$ or vice versa takes infinitely large amount of work. Thus, we only need to optimize over bijections that match the same type of topological features. Subsequently, we can equivalently rewrite d_{top} in terms of $B(G), D(G), B(H)$ and $D(H)$ as

$$d_{top}(G, H) = \Big(\min_{\tau_0} \sum_{b \in B(G)} \big[b - \tau_0(b)\big]^2 + \min_{\tau_1} \sum_{d \in D(G)} \big[d - \tau_1(d)\big]^2\Big)^{\frac{1}{2}}, \quad (2)$$

where τ_0 is a bijection from $B(G)$ to $B(H)$ and τ_1 is a bijection from $D(G)$ to $D(H)$. The first term matches connected components to connected components and the second term matches cycles to cycles. Note that matching each type of topological feature separately is commonly done in medical imaging and machine learning studies (Clough et al., 2020; Hu et al., 2019). The topological distance d_{top} has a closed-form solution that allows for efficient computation as follows (Songdechakraiwut et al., 2021).

$$d_{top}(G,H) = \Big(\sum_{b \in B(G)} \big[b - \tau_0^*(b)\big]^2 + \sum_{d \in D(G)} \big[d - \tau_1^*(d)\big]^2 \Big)^{\frac{1}{2}}, \tag{3}$$

where τ_0^* maps the l-th smallest birth value in $B(G)$ to the l-th smallest birth value in $B(H)$ and τ_1^* maps the l-th smallest death value in $D(G)$ to the l-th smallest death value in $D(H)$ for all l.

A proof is in the supplementary material. As a result, the optimal matching cost can be computed quickly and efficiently by sorting edge weights and matching them in order. The computational cost of evaluating d_{top} is $O(n \log n)$, where n is the number of edges in networks.

3.2 Topological Clustering

Let $G = (V, \boldsymbol{w})$ and $H = (V, \boldsymbol{u})$ be two networks. We define the network dissimilarity d_{net} between G and H as a weighted sum of the squared geometric distance and the squared topological distance:

$$d_{net}\big(G,H\big) = (1-\lambda) \sum_i \sum_{j>i} \big(w_{ij} - u_{ij}\big)^2 + \lambda d_{top}^2\big(G,H\big), \tag{4}$$

where $\lambda \in [0,1]$ controls the relative weight between the geometric and topological terms. The geometric distance measures the node-by-node dissimilarity in the networks that is not captured by topology alone. Given observed networks with identical node sets, the goal is to partition the networks into k clusters $\mathcal{C} = \{C_h\}_{h=1}^k$ with corresponding cluster centroids or representatives $\mathcal{M} = \{M_h\}_{h=1}^k$ such that the sum of the network dissimilarities d_{net} from the networks to their representatives is minimized, i.e.,

$$\min_{\mathcal{C},\mathcal{M}} L(\mathcal{C},\mathcal{M}) = \min_{\mathcal{C},\mathcal{M}} \sum_{h=1}^k \sum_{G \in C_h} d_{net}(M_h, G). \tag{5}$$

The *topological clustering* formulation given in (5) suggests a natural iterative relocation algorithm using coordinate descent (Banerjee et al., 2005). In particular, the algorithm alternatively carries out two steps: an assignment step and a re-estimation step. In the assignment step, L is minimized with respect to $\mathcal{C}$ while holding $\mathcal{M}$ fixed, i.e., minimization of L can be achieved simply by assigning each observed network to the cluster whose representative M_h is the nearest in terms of the criterion d_{net}. In the re-estimation step, the algorithm minimizes L with respect to $\mathcal{M}$ while holding $\mathcal{C}$ fixed, i.e., we minimize L by re-estimating the representatives for each individual cluster:

$$\min_{\mathcal{M}} \sum_{h=1}^k \sum_{G \in C_h} d_{net}(M_h, G) = \sum_{h=1}^k \min_{M_h} \sum_{G \in C_h} d_{net}(M_h, G). \tag{6}$$

We will consider solving the objective function given in (6) for $\lambda = 0, \lambda = 1$ and $\lambda \in (0,1)$.

$\lambda = 0$ describes conventional edge clustering since the topological term is excluded (MacQueen et al., 1967).

$\lambda = 1$ describes clustering based on pure topology. Each representative M_h of cluster C_h minimizes the sum of the squared topological distances, i.e.,

$$\min_{M_h} \sum_{G \in C_h} d_{top}^2(M_h, G) = \min_{B(M_h), D(M_h)} \sum_{G \in C_h} \Big(\sum_{b \in B(M_h)} \big[b - \tau_0^*(b)\big]^2 + \sum_{d \in D(M_h)} \big[d - \tau_1^*(d)\big]^2 \Big). \tag{7}$$

Thus, we only need to optimize over topology of the network, i.e., $B(M_h)$ and $D(M_h)$, instead of the original network M_h itself. This case is useful for clustering networks whose node sets are not identical. The optimal topology solving (7) is viewed as the *topological centroid* of networks in cluster C_h. Interestingly, the topological centroid has closed-form solution and can be calculated analytically as follows.

Lemma 1. *Let $G_1, ..., G_n$ be n networks each with m nodes. Let $B(G_i) : b_{i,1} \leq \cdots \leq b_{i,m-1}$ and $D(G_i) : d_{i,1} \leq \cdots \leq d_{i,1+m(m-3)/2}$ be the topology of G_i. It follows that the l-th smallest birth value of the topological centroid of the n networks is given by the mean of all the l-th smallest birth values of such networks, i.e., $\sum_{i=1}^{n} b_{i,l}/n$. Similarly, the l-th smallest death value of the topological centroid is given by $\sum_{i=1}^{n} d_{i,l}/n$.*

Since Eq. (7) is quadratic, Lemma 1 can be proved by setting its derivative equal to zero. The complete proof is given in the supplementary material.

For the last case, if $\lambda \in (0, 1)$, we can optimize the objective function given in (6) by gradient descent. Let $\Theta = (V, \boldsymbol{\theta})$ be a cluster representative being estimated given C_h. The gradient of the squared topological distance $\nabla_\theta d^2_{top}(\Theta, G)$ with respect to edge weights $\boldsymbol{\theta} = (\theta_{ij})$ is given as a gradient matrix whose ij-th entry is

$$\frac{\partial d^2_{top}(\Theta, G)}{\partial \theta_{ij}} = \begin{cases} 2\big[\theta_{ij} - \tau_0^*(\theta_{ij})\big] & \text{if } \theta_{ij} \in B(\Theta); \\ 2\big[\theta_{ij} - \tau_1^*(\theta_{ij})\big] & \text{if } \theta_{ij} \in D(\Theta). \end{cases} \tag{8}$$

This follows because the edge weight set decomposes into the collection of births and the collection of deaths. Intuitively, by slightly adjusting the edge weight θ_{ij}, we have the slight adjustment of either a birth value in $B(\Theta)$ or a death value in $D(\Theta)$, which slightly changes the topology of the network Θ. The gradient computation consists of computing persistence barcodes and finding the optimal matching using the closed-form solution given in (3), requiring $O(n \log n)$ operations, where n is the number of edges in networks.

Evaluating the gradient for (6) requires computing the gradients of all the observed networks. This can be computationally demanding when the size of a dataset is large. However, an equivalent minimization problem that allows faster computation is possible using the following result:

Lemma 2.

$$\sum_{h=1}^{k} \min_{\Theta_h} \sum_{G \in C_h} d_{net}(\Theta_h, G) = \sum_{h=1}^{k} \min_{\Theta_h} \Big((1-\lambda) \sum_i \sum_{j>i} \big(\theta_{h,ij} - \overline{w}_{h,ij}\big)^2 + \lambda d^2_{top}\big(\Theta_h, \widehat{M}_{top,h}\big) \Big), \tag{9}$$

where $\overline{\boldsymbol{w}}_{\boldsymbol{h}} = (\overline{w}_{h,ij})$ are edge weights in the sample mean network $\overline{M}_h = (V, \overline{\boldsymbol{w}}_{\boldsymbol{h}})$ of cluster C_h, and $\widehat{M}_{top,h}$ is the topological centroid of networks in cluster C_h.

Thus, instead of computing the gradient for every network in the set, it is sufficient to compute the gradient at the cluster sample means $\overline{M}_h$ and topological centroids $\widehat{M}_{top,h}$. Hence, one only needs to perform *topological interpolation* between the sample mean network $\overline{M}_h$ and the topological centroid $\widehat{M}_{top,h}$ of each cluster. That is, the optimal representative is the one whose geometric location is close to $\overline{M}_h$ and topology is similar to $\widehat{M}_{top,h}$. At each current iteration, we take a step in the direction of negative gradient with respect to an updated Θ from the previous iteration. As before, as λ increases, the importance of topological characteristics increases.

Furthermore, we have the following theorem:

Theorem 1. *The topological clustering algorithm monotonically decreases L in (5) and terminates in a finite number of steps at a locally optimal partition.*

The proofs for Lemma 2 and Theorem 1 are provided in the supplementary material.

4 Validation using simulated networks

Simulated networks of different topological structure are used to evaluate the clustering performance of the proposed approach and that of well-known clustering algorithms including k-means and spectral clustering (Shi & Malik, 2000). Edge weights are vectorized for implementation of k-means and spectral clustering. In addition, we evaluate k-medoids equipped with the very popular bottleneck distance known for its central stability theorem (Cohen-Steiner et al., 2007). We also evaluate k-mediods equipped with the absolute difference between modularity measures (Newman, 2006; Reichardt & Bornholdt, 2006) to exploit community structure.

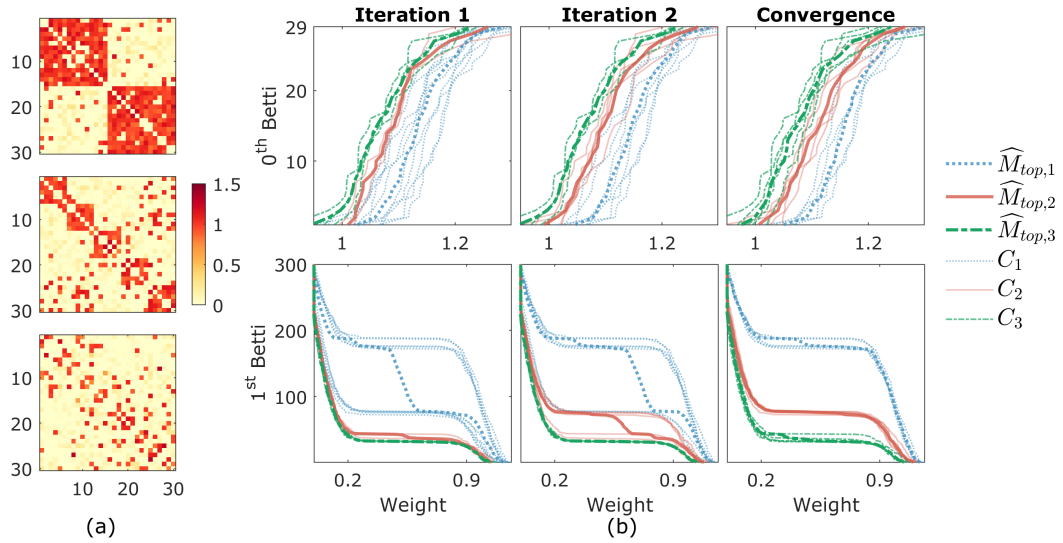

Figure 2: Toy illustration of performing topological clustering using only topological centroids ($\lambda = 1$ in (7)). Clustering is performed on 15 networks: five each with $m = 2, 5$ and 10 modules. All networks use $|V| = 30$, within module connection probability $r = 0.9$ and edge weight standard deviation $\sigma = 0.1$. (a) Network examples for $m = 2$ (top), 5 (middle) and 10 (bottom) modules. (b) Illustration of topological clustering. Topological centroids $\widehat{M}_{top,i}$ (thick lines) and individual network topologies are visualized using Betti numbers as a function of edge weights (filtration values). The algorithm converges in three iterations and the final partition (last column) perfectly matches the ground truth, i.e., C_1, C_2 and C_3 corresponds to $m = 2, 5$ and 10.

Modular network structure Random modular networks $\mathcal{X}_i$ are simulated with $|V|$ nodes and m modules such that the nodes are evenly distributed among modules. Figure 2-a displays modular networks with $|V| = 30$ nodes and $m = 2, 5, 10$ modules such that $|V|/m = 15, 6, 3$ nodes are in each module, respectively. Edges connecting two nodes within the same module are assigned a random weight following a normal distribution $\mathcal{N}(\mu, \sigma^2)$ with probability r or otherwise Gaussian noise $\mathcal{N}(0, \sigma^2)$ with probability $1-r$. On the other hand, edges connecting nodes in different modules have probability $1-r$ of being $\mathcal{N}(\mu, \sigma^2)$ and probability r of being $\mathcal{N}(0, \sigma^2)$. The modular structure becomes more pronounced as the within-module connection probability r increases. Any negative edge weights are set to zero. This procedure yields random networks $\mathcal{X}_i$ that exhibit topological connectedness. We use $\mu = 1$ universally throughout the study. Figure 2-b illustrates a cluster analysis of the proposed approach using only topological centroids ($\lambda = 1$ in (7)) on a toy dataset.

Simulation Three groups of modular networks $L_1 = \{\mathcal{X}_i\}_{i=1}^{20}$, $L_2 = \{\mathcal{X}_i\}_{i=21}^{40}$ and $L_3 = \{\mathcal{X}_i\}_{i=41}^{60}$ are simulated resulting in 60 networks in the dataset, each of which has a group label L_1, L_2 or L_3. We consider network sizes $|V| = 60, 120$ with $r = 0.6$ and $\sigma = 0.5$. This choice of r results in relatively weak module structure, as illustrated in Figure 3. Two different sets of module sizes are also considered: 1) $m = 2, 3$ and 5, and 2) $m = 2, 5$ and 10. We also create control datasets where the number of modules is either $m = 2$ or $m = 5$ in all three groups. Initial clusters for k-means, k-medoids and the proposed topological clustering are selected at random. For our method, we employ the clustering based on topological centroid ($\lambda = 1$ in (7)) to demonstrate the discriminative power of topological distance in clustering noisy data.

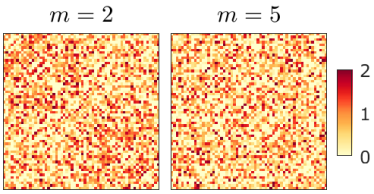

Figure 3: Example modular networks with $|V| = 60$ nodes are generated using within module connection probability $r = 0.6$ and edge weight standard deviation $\sigma = 0.5$. Left: $m = 2$ modules. Right: $m = 5$ modules.

The dataset is partitioned into three clusters C_1, C_2 and C_3 using the candidate algorithms. Clustering performance is then evaluated by first assigning each cluster to the group label that is most frequent in that cluster and then calculating the accuracy statistic s as the fraction of correctly labeled networks,

Table 1: Clustering performance comparison for simulated networks. (a) average accuracy (b) average p-values for various parameter settings of $|V|$ (number of nodes) and m (number of modules).

$\|V\|$	m	k-means	Spectral	Modularity	Bottleneck	Topology
60	2/3/5	0.46 ± 0.04	0.39 ± 0.03	0.49 ± 0.05	0.42 ± 0.04	$\mathbf{0.75 \pm 0.06}$
	2/5/10	0.44 ± 0.05	0.39 ± 0.03	0.53 ± 0.05	0.45 ± 0.05	$\mathbf{0.78 \pm 0.07}$
120	2/3/5	0.50 ± 0.08	0.39 ± 0.03	0.44 ± 0.05	0.42 ± 0.04	$\mathbf{0.95 \pm 0.03}$
	2/5/10	0.50 ± 0.07	0.40 ± 0.03	0.47 ± 0.05	0.44 ± 0.05	$\mathbf{0.86 \pm 0.12}$
120	2/2/2	0.42 ± 0.04	0.40 ± 0.03	0.42 ± 0.04	0.42 ± 0.04	0.42 ± 0.03
	5/5/5	0.41 ± 0.03	0.39 ± 0.03	0.42 ± 0.04	0.42 ± 0.04	0.42 ± 0.03

(a) Average accuracy

$\|V\|$	m	k-means	Spectral	Modularity	Bottleneck	Topology
60	2/3/5	0.28 ± 0.26	0.68 ± 0.29	0.16 ± 0.19	0.56 ± 0.32	$\mathbf{0.00 \pm 0.00}$
	2/5/10	0.37 ± 0.33	0.64 ± 0.29	0.05 ± 0.11	0.39 ± 0.33	$\mathbf{0.00 \pm 0.00}$
120	2/3/5	0.17 ± 0.27	0.61 ± 0.30	0.47 ± 0.33	0.54 ± 0.32	$\mathbf{0.00 \pm 0.00}$
	2/5/10	0.12 ± 0.22	0.58 ± 0.32	0.26 ± 0.27	0.43 ± 0.31	$\mathbf{0.00 \pm 0.00}$
120	2/2/2	0.55 ± 0.31	0.60 ± 0.29	0.60 ± 0.31	0.59 ± 0.31	0.59 ± 0.30
	5/5/5	0.60 ± 0.30	0.62 ± 0.33	0.55 ± 0.29	0.56 ± 0.30	0.56 ± 0.30

(b) Average p-values

i.e., $s = \frac{1}{60}\sum_{i=1}^{3} \max_j\{|C_i \cap L_j|\}$, where $|C_i \cap L_j|$ denotes the number of common networks in both C_i and L_j. Note this evaluation of clustering performance is called *purity*, which not only is transparent and interpretable but also works well in this simulation study where the number and size of clusters are small and balanced, respectively (Manning et al., 2008). Since the distribution of the accuracy s is unknown, a permutation test is used to determine the empirical distribution under the null hypothesis that sample networks and their group labels are independent (Ojala & Garriga, 2010). The empirical distribution is calculated by repeatedly shuffling the group labels and then re-computing the corresponding accuracy for one million random permutations. By comparing the observed accuracy to this empirical distribution, we can determine the statistical significance of the clustering performance. The p-value is calculated as the fraction of permutations that give accuracy higher than the observed accuracy s. The average p-value and average accuracy across 100 independent simulations is reported.

Result Table 1 summarizes the performance results. Topological clustering shows significant gains in accuracy relative to all the other methods evaluated. In the cases of differing modular structure, there is dependency between the sample networks and their group labels. Thus, the p-value indicates the degree of statistical significance to which the clustering algorithm differentiates structure (Ojala & Garriga, 2010). Large p-values indicate the algorithm is unable to differentiate network structure. The results indicate that the topological approach offers significant improvement over these existing algorithms. On the other hand, when the structure is the same throughout the dataset, topological differences should not be detected and the p-values should be large, as observed for all algorithms in the bottom two rows of Table 1-b.

5 Application to Functional Brain Networks

Dataset We evaluate our method using an extended brain network dataset from the anesthesia study reported by Banks et al. (2020) (see the supplementary material). The brain networks are based on alpha band (8-12 Hz) weighted phase lag index (Vinck et al., 2011) applied to 10-second segments of resting state intracranial electroencephalography recordings from eleven neurosurgical patients administered increasing doses of the general anesthetic propofol just prior to surgery. The network size varies from 89 to 199 nodes while the number of networks (10-second segments) per subject varies from 71 to 119. Each segment is labeled as one of the three arousal states: pre-drug wake (WA), sedated/responsive (S), or unresponsive (U); these labels are used as ground truth in the clustering

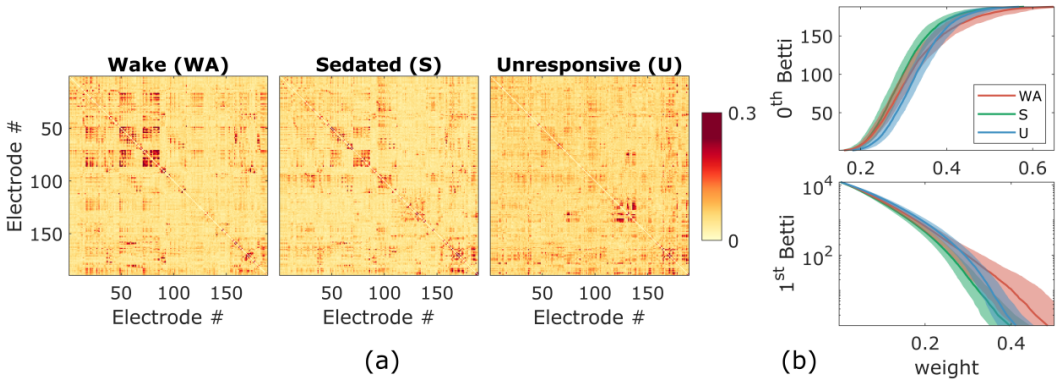

Figure 4: Representative data from a single subject. For this subject there are 36 measured networks in each of three conditions: wake, sedated, and unresponsive states for a total of 108 measured networks. (a) Sample mean networks during wake, sedated and unresponsive states of the subject computed using ground truth labels. (b) Betti plots based on ground truth labels. Each thick line represents a topological centroid of each state. Shaded areas around the centroids represent standard deviation.

analysis. Figure 4 illustrates sample mean networks, and Betti plots describing topology, for a single subject.

Performance evaluation We apply the adjusted Rand index (ARI) (Hubert & Arabie, 1985), adjusted mutual information (AMI) (Vinh et al., 2010) and Fowlkes–Mallows index (FMI) (Fowlkes & Mallows, 1983) to compare clustering performance against ground truth. For all three measures, lower scores indicate less similarity while higher scores show higher similarity between estimated and ground-truth clusters with perfect agreement scored as 1.0. For each subject, we calculate these performance metrics by running a clustering algorithm independently for 20 times, resulting in 20 scores. We then average the 10×20 scores to obtain a final score, which describes the overall quality of the output clusters across trials and subjects. We also calculated average confusion matrices for each method by assigning each cluster to the state that is most frequent in that cluster.

Cluster analysis We compare our method against existing approaches in brain network analyses: bottleneck distance (Cohen-Steiner et al., 2007); Gromov-Hausdorff (GH) (Lee et al., 2012); Kolmogorov-Smirnov (KS) (Chung et al., 2019); and the operator or spectral norm (Banks et al., 2020). Bottleneck distance is widely used in applications of persistent homology, while GH distance is often used to compare dendrogram shape differences in brain networks. KS distance was introduced to perform statistical inference on Betti numbers for large-scale brain networks. The operator or spectral norm characterizes the size of the difference between network adjacency matrices and thus is based only on geometry. We use k-medoids equipped with these metrics to perform cluster analysis on the dataset. Initial clusters are selected at random for all the baselines as well as our topological clustering. Figure 5 reports the performance of the topological approach for several predefined λ's. $\lambda = 0.4$ results in the highest average score and is used in the results that follow.

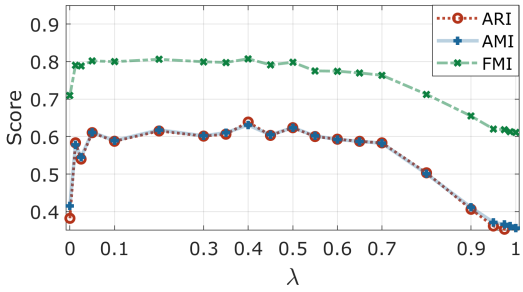

Figure 5: Average ARI, AMI and FMI performance measures across all eleven subjects as a function of λ, the relative weight given to topology.

Result Table 2 summarizes the quantitative results. Figure 6 displays the fraction of correctly and incorrectly assigned states for each true cluster for the subject presented in Figure 4; the results are similar for other subjects. Transitioning from the wake state into the state of unresponsiveness results in dramatic changes in brain connectivity (Banks et al., 2020). Still, the sedated state, in which

Table 2: Quantitative clustering results on brain network dataset.

Measure	Spectral	Bottleneck	GH	KS	Topology ($\lambda = 0.4$)
ARI	0.44 ± 0.30	0.30 ± 0.32	0.35 ± 0.22	0.33 ± 0.32	$\mathbf{0.64 \pm 0.23}$
AMI	0.46 ± 0.28	0.27 ± 0.29	0.32 ± 0.22	0.34 ± 0.31	$\mathbf{0.63 \pm 0.22}$
FMI	0.67 ± 0.18	0.58 ± 0.21	0.61 ± 0.16	0.60 ± 0.20	$\mathbf{0.81 \pm 0.13}$

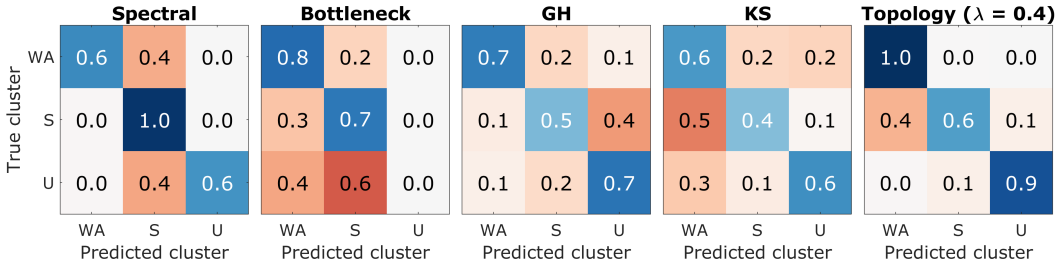

Figure 6: Average confusion matrices from clustering the subject data illustrated in Figure 4. We independently run each of the algorithms 20 times and then average the resulting confusion matrices.

subjects have been administered propofol but are still conscious, is expected to be more like the wake state than the unresponsive state. Thus, errors are expected to be more likely in differentiating sedated and wake states than sedated and unresponsive states. The majority of cluster errors from the proposed topological clustering approach are associated with the natural overlap between wake and sedated states, as illustrated in Figure 6. This suggests that our method is consistent with such biological expectations. Quantitatively, Table 2 and Figure 6 indicate that topological clustering offers significantly better performance than the previously proposed methods in brain network analyses.

6 Impact

The demonstrated effectiveness and computational elegance of our approach to clustering networks based on topological similarity will have a high impact on the analysis of large and complex network representations. In the study of brain networks, algorithms that can demonstrate correlates of behavioral states are of considerable interest. The results presented here, in which biomarkers of changes in arousal state are derived from data obtained during general anesthesia, demonstrate potential for addressing the important clinical problem of passively assessing arousal state in clinical settings, e.g., monitoring depth of anesthesia and in establishing diagnosis and prognosis for patients with traumatic brain injury and other disorders of consciousness. More broadly, the algorithm presented here will contribute to elucidating the neural basis of consciousness, one of the most important open problems in biomedical science.

References

Martial Agueh and Guillaume Carlier. Barycenters in the Wasserstein space. *SIAM Journal on Mathematical Analysis*, 43(2):904–924, 2011.

Arindam Banerjee, Srujana Merugu, Inderjit S Dhillon, Joydeep Ghosh, and John Lafferty. Clustering with bregman divergences. *Journal of Machine Learning Research*, 6(10), 2005.

Matthew I Banks, Bryan M Krause, Christopher M Endemann, Declan I Campbell, Christopher K Kovach, Mark Eric Dyken, Hiroto Kawasaki, and Kirill V Nourski. Cortical functional connectivity indexes arousal state during sleep and anesthesia. *NeuroImage*, 211:116627, 2020.

Alain Barrat, Marc Barthelemy, Romualdo Pastor-Satorras, and Alessandro Vespignani. The architecture of complex weighted networks. *Proceedings of the National Academy of Sciences*, 101(11): 3747–3752, 2004.

Ed Bullmore and Olaf Sporns. Complex brain networks: graph theoretical analysis of structural and functional systems. *Nature Reviews Neuroscience*, 10(3):186–198, 2009.

Moo K Chung, Hyekyoung Lee, Alex DiChristofano, Hernando Ombao, and Victor Solo. Exact topological inference of the resting-state brain networks in twins. *Network Neuroscience*, 3(3): 674–694, 2019.

James Clough, Nicholas Byrne, Ilkay Oksuz, Veronika A Zimmer, Julia A Schnabel, and Andrew King. A topological loss function for deep-learning based image segmentation using persistent homology. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020.

David Cohen-Steiner, Herbert Edelsbrunner, and John Harer. Stability of persistence diagrams. *Discrete & Computational Geometry*, 37(1):103–120, 2007.

David Cohen-Steiner, Herbert Edelsbrunner, John Harer, and Yuriy Mileyko. Lipschitz functions have Lp-stable persistence. *Foundations of Computational Mathematics*, 2010.

Marco Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. *Advances in Neural Information Processing Systems*, 26:2292–2300, 2013.

Marco Cuturi and Arnaud Doucet. Fast computation of Wasserstein barycenters. In *International Conference on Machine Learning*, pp. 685–693, 2014.

Herbert Edelsbrunner and John Harer. Persistent homology-a survey. *Contemporary Mathematics*, 453:257–282, 2008.

Herbert Edelsbrunner, David Letscher, and Afra Zomorodian. Topological persistence and simplification. In *Proceedings 41st Annual Symposium on Foundations of Computer Science*, pp. 454–463. IEEE, 2000.

Edward B Fowlkes and Colin L Mallows. A method for comparing two hierarchical clusterings. *Journal of the American Statistical Association*, 78(383):553–569, 1983.

Robert Ghrist. Barcodes: the persistent topology of data. *Bulletin of the American Mathematical Society*, 45(1):61–75, 2008.

Christopher J Honey, Rolf Kötter, Michael Breakspear, and Olaf Sporns. Network structure of cerebral cortex shapes functional connectivity on multiple time scales. *Proceedings of the National Academy of Sciences*, 104(24):10240–10245, 2007.

Xiaoling Hu, Fuxin Li, Dimitris Samaras, and Chao Chen. Topology-preserving deep image segmentation. *Advances in Neural Information Processing Systems*, 32, 2019.

Lawrence Hubert and Phipps Arabie. Comparing partitions. *Journal of Classification*, 2(1):193–218, 1985.

Joel Keizer, Yue-Xian Li, Stanko Stojilković, and John Rinzel. Insp3-induced ca2+ excitability of the endoplasmic reticulum. *Molecular Biology of the Cell*, 6(8):945–951, 1995.

Soheil Kolouri, Se Rim Park, Matthew Thorpe, Dejan Slepcev, and Gustavo K Rohde. Optimal mass transport: Signal processing and machine-learning applications. *IEEE Signal Processing Magazine*, 34(4):43–59, 2017.

Yung-Keun Kwon and Kwang-Hyun Cho. Analysis of feedback loops and robustness in network evolution based on boolean models. *BMC Bioinformatics*, 8(1):1–9, 2007.

Hyekyoung Lee, Hyejin Kang, Moo K Chung, Bung-Nyun Kim, and Dong Soo Lee. Persistent brain network homology from the perspective of dendrogram. *IEEE Transactions on Medical Imaging*, 31(12):2267–2277, 2012.

Lingxiao Li, Aude Genevay, Mikhail Yurochkin, and Justin M Solomon. Continuous regularized Wasserstein barycenters. *Advances in Neural Information Processing Systems*, 33, 2020.

James MacQueen et al. Some methods for classification and analysis of multivariate observations. In *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, volume 1, pp. 281–297. Oakland, CA, USA, 1967.

Christopher D Manning, Prabhakar Raghavan, and Hinrich Schütze. *Introduction to Information Retrieval*. Cambridge University Press, 2008.

Liang Mi, Wen Zhang, Xianfeng Gu, and Yalin Wang. Variational Wasserstein clustering. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 322–337, 2018.

James R Munkres. *Elements of Algebraic Topology*. CRC press, 2018.

Mark EJ Newman. Modularity and community structure in networks. *Proceedings of the National Academy of Sciences*, 103(23):8577–8582, 2006.

Markus Ojala and Gemma C Garriga. Permutation tests for studying classifier performance. *Journal of Machine Learning Research*, 11(6), 2010.

Ertugrul M Ozbudak, Attila Becskei, and Alexander Van Oudenaarden. A system of counteracting feedback loops regulates cdc42p activity during spontaneous cell polarization. *Developmental Cell*, 9(4):565–571, 2005.

Julien Rabin, Gabriel Peyré, Julie Delon, and Marc Bernot. Wasserstein barycenter and its application to texture mixing. In *International Conference on Scale Space and Variational Methods in Computer Vision*, pp. 435–446. Springer, 2011.

Jörg Reichardt and Stefan Bornholdt. Statistical mechanics of community detection. *Physical Review E*, 74(1):016110, 2006.

Vanessa Robins and Katharine Turner. Principal component analysis of persistent homology rank functions with case studies of spatial point patterns, sphere packing and colloids. *Physica D: Nonlinear Phenomena*, 334:99–117, 2016.

Karl Rohe, Sourav Chatterjee, Bin Yu, et al. Spectral clustering and the high-dimensional stochastic blockmodel. *Annals of Statistics*, 39(4):1878–1915, 2011.

Jianbo Shi and Jitendra Malik. Normalized cuts and image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(8):888–905, 2000.

Justin Solomon, Fernando De Goes, Gabriel Peyré, Marco Cuturi, Adrian Butscher, Andy Nguyen, Tao Du, and Leonidas Guibas. Convolutional Wasserstein distances: Efficient optimal transportation on geometric domains. *ACM Transactions on Graphics (TOG)*, 34(4):1–11, 2015.

Tananun Songdechakraiwut, Li Shen, and Moo Chung. Topological learning and its application to multimodal brain network integration. *24th International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, 2021.

Guillaume Staerman, Pierre Laforgue, Pavlo Mozharovskyi, and Florence d'Alché Buc. When OT meets MoM: Robust estimation of Wasserstein distance. In *International Conference on Artificial Intelligence and Statistics*, pp. 136–144, 2021.

KV Venkatesh, Sharad Bhartiya, and Anurag Ruhela. Multiple feedback loops are key to a robust dynamic performance of tryptophan regulation in escherichia coli. *FEBS Letters*, 563(1-3):234–240, 2004.

Jules Vidal, Joseph Budin, and Julien Tierny. Progressive Wasserstein barycenters of persistence diagrams. *IEEE Transactions on Visualization and Computer Graphics*, 26(1):151–161, 2019.

Martin Vinck, Robert Oostenveld, Marijn Van Wingerden, Franscesco Battaglia, and Cyriel MA Pennartz. An improved index of phase-synchronization for electrophysiological data in the presence of volume-conduction, noise and sample-size bias. *NeuroImage*, 55(4):1548–1565, 2011.

Nguyen Xuan Vinh, Julien Epps, and James Bailey. Information theoretic measures for clusterings comparison: Variants, properties, normalization and correction for chance. *The Journal of Machine Learning Research*, 11:2837–2854, 2010.

Larry Wasserman. Topological data analysis. *Annual Review of Statistics and Its Application*, 5: 501–532, 2018.

Orion D Weiner, Paul O Neilsen, Glenn D Prestwich, Marc W Kirschner, Lewis C Cantley, and Henry R Bourne. A ptdinsp 3-and rho gtpase-mediated positive feedback loop regulates neutrophil polarity. *Nature Cell Biology*, 4(7):509–513, 2002.

Kelin Xia and Guo-Wei Wei. Persistent homology analysis of protein structure, flexibility, and folding. *International Journal for Numerical Methods in Biomedical Engineering*, 30(8):814–844, 2014.

Yujia Xie, Xiangfeng Wang, Ruijia Wang, and Hongyuan Zha. A fast proximal point method for computing exact Wasserstein distance. In *Uncertainty in Artificial Intelligence*, pp. 433–453, 2020.

Rui Xu and Donald Wunsch. Survey of clustering algorithms. *IEEE Transactions on Neural Networks*, 16(3):645–678, 2005.

Jianbo Ye, Panruo Wu, James Z Wang, and Jia Li. Fast discrete distribution clustering using Wasserstein barycenter with sparse support. *IEEE Transactions on Signal Processing*, 65(9): 2317–2332, 2017.

Hao Yin, Austin R Benson, Jure Leskovec, and David F Gleich. Local higher-order graph clustering. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 555–564, 2017.