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 associated with connected components and cycles. 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 (Barannikov, 1994; Edelsbrunner et al., 2000; Wasserman, 2018) has recently emerged as a powerful tool for understanding, characterizing and quantifying complex networks (Carrière et al., 2020; 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, cycles are ubiquitous and are 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 (Songdechakraiwut et al., 2021), 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; Lacombe et al., 2018; 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 network graphs have closed-form solutions that can be computed exactly and efficiently (Songdechakraiwut et al., 2021) because the persistence barcodes are inherently one dimensional. 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 when there is 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 the 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 administration of general anesthesia. Our proposed method outperforms other clustering approaches in both the simulated and measured data.

The paper is organized as follows. Background on our one-dimensional representation 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, and conclude the paper with a brief discussion of the potential impact of this work.

2 ONE DIMENSIONAL 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 (Munkres, 2018), a simplicial complex comprising only nodes (0-dimensional topological features) and edges (1-dimensional topological features). There are no topological features of higher dimensions in the 1-skeleton, in contrast to well-known Rips complexes (Ghrist, 2008). 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. Note that other filtrations for analyzing graphs have been proposed, including use of descriptor functions such as heat kernels (Carrière et al., 2020) and task-specific learning (Hofer et al., 2020).

2.2 BIRTH-DEATH DECOMPOSITION

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

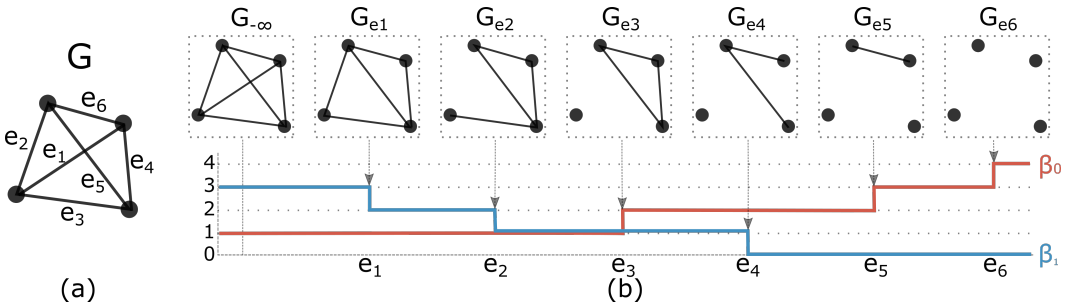

Figure 1: (a) Four-node network G. (b) As the filtration value increases, the number of connected components β_0 monotonically increases while the number of cycles β_1 monotonically decreases. Connected components are born at the edge weights e_3, e_5, e_6 while cycles die at the edge weights e_1, e_2,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

In the edge-weight threshold graph filtration defined in Section 2.1, connected components are born and cycles die as the filtration value increases (Chung et al., 2019). Specifically, β_0 is monotonically increasing from $\beta_0(G_{-\infty}) = 1$ to $\beta_0(G_{\infty}) = |V|$. There are $\beta_0(G_{\infty}) - \beta_0(G_{-\infty}) = |V| - 1$ connected components that are born over the filtration. Connected components will never die once they are born, implying that every connected 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, $G_{-\infty}$ is a complete graph containing all possible cycles; thus, all cycles 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 $\frac{|V|(|V|-1)}{2}$ edge weights and $|V| - 1$ of these weights are associated with the birth of connected components, the number of cycles in $G_{-\infty}$ is thus 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 $D(G) = \{e_1, e_2, e_4\}$. Other graph filtrations (Carrière et al., 2020; Hofer et al., 2020) do not necessarily share this monotonicity property and consequently one-dimensional barcode representations are not applicable.

Finding the 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 Clustering Method

3.1 Topological Distance Simplification

Use of edge-weight threshold filtration and limiting consideration to connected components and cycles as topological features results in significant simplification of the 2-Wasserstein distance (Rabin et al., 2011) between barcode descriptors (Cohen-Steiner et al., 2010) of networks as follows. 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 an 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

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

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. 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 birth and death values, 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^2_{net} between G and H as a weighted sum of the squared geometric distance and the squared topological distance:

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

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 and is helpful when node identity is meaningful, such as in neuroscience applications. 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^2_{net} from the networks to their representatives is minimized, i.e.,

$$\min_{\mathcal{C},\mathcal{M}} L(\mathcal{C},\mathcal{M}) = \min_{\mathcal{C},\mathcal{M}} \sum_{h=1}^k \sum_{G \in C_h} d^2_{net}(M_h, G). \tag{5}$$

The *topological clustering* formulation given in (5) suggests a natural iterative relocation algorithm using coordinate descent (Banerjee et al., 2005). In particular, the algorithm alternates between 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. Minimization of L is achieved simply by assigning each observed network to the cluster whose representative M_h is the nearest in terms of the criterion d^2_{net}. In the re-estimation step, the algorithm minimizes L with respect to $\mathcal{M}$ while holding $\mathcal{C}$ fixed. In this case, L is minimized by re-estimating the representatives for each individual cluster:

$$\min_{\mathcal{M}} \sum_{h=1}^k \sum_{G \in C_h} d^2_{net}(M_h, G) = \sum_{h=1}^k \min_{M_h} \sum_{G \in C_h} d^2_{net}(M_h, G). \tag{6}$$

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 (MacQueen, 1967) since the topological term is excluded.

$\lambda = 1$ describes clustering based on pure topology and ignores correspondence of edge weights. This case is potentially useful for clustering networks whose node sets are not identical or of different size. Each representative M_h of cluster C_h minimizes the sum of the squared topological distances, i.e.,

$$\min_{M_h} \sum_{G \in C_h} d^2_{top}(M_h, G) = \min_{B(M_h), D(M_h)} \sum_{G \in C_h} \Big(\sum_{b \in B(M_h)} \big[b - \tau^*_0(b)\big]^2 + \sum_{d \in D(M_h)} \big[d - \tau^*_1(d)\big]^2 \Big). \tag{7}$$

Thus, we only need to optimize over the topology of the network, i.e., $B(M_h)$ and $D(M_h)$, instead of the original network M_h itself. The topology solving (7) is 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. The results in (Songdechakraiwut & Chung, 2020) may be used to find the centroid of different size networks.

For the most general case considering both correspondence of edge weights and topology, i.e., when $\lambda \in (0, 1)$, we can optimize the objective function given in (6) by gradient descent. Let $H = (V, \boldsymbol{u})$ be a cluster representative being estimated given C_h. The gradient of the squared topological distance $\nabla_H d^2_{top}(G, H)$ with respect to edge weights $\boldsymbol{u} = (u_{ij})$ is given as a gradient matrix whose ij-th entry is

$$\frac{\partial d^2_{top}(G, H)}{\partial u_{ij}} = \begin{cases} 2\big[u_{ij} - \tau^*_0(u_{ij})\big] & \text{if } u_{ij} \in B(H); \\ 2\big[u_{ij} - \tau^*_1(u_{ij})\big] & \text{if } u_{ij} \in D(H). \end{cases} \tag{8}$$

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 u_{ij}, we have the slight adjustment of either a birth value in $B(H)$ or a death value in $D(H)$, which slightly changes the topology of the network H. 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.

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

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 H from the previous iteration. Note that the proposed method constrains the search of cluster centroids to a space of meaningful networks through topological interpolation. In contrast, the sample mean, such as would be used in a naive application of k-means, does not necessarily represent a meaningful network.

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 relative to that of seven other methods. We use the term *topology* to denote the proposed approach with $\lambda > 0$. Clustering using the special case of $\lambda = 0$ is termed the *edge-weight* method. Recall that the graph filtration used decomposes the edge weights into two groups of birth and death values corresponding to connected components and cycles (section 2.2). In order to assess the impact of separating edge weights into birth and death sets, we evaluate an ad hoc method called *sort* that applies k-means to feature vectors obtained by simply sorting edge weights.

In addition, we evaluate several clustering algorithms that utilize network representations. Two persistent homology clustering methods based on the well-known Rips filtration (Ghrist, 2008) are evaluated. The Rips filtration results in standard two-dimensional barcodes representing connected components and cycles. Consequently, two separate Wasserstein distances are needed to match connected components to connected components and cycles to cycles. The *Wasserstein* method performs clustering using k-medoids based on the sum of the two Wasserstein distances for connected components and cycles. The computational complexity of the *Wasserstein* method is managed by utilizing an approximation algorithm (Lacombe et al., 2018) to compute Wasserstein distance. A Wasserstein barycenter clustering algorithm called *Bregman Alternating Direction Method of Multipliers* (B-ADMM) (Ye et al., 2017) is used to cluster two-dimensional barcodes for cycles. In order to make B-ADMM computationally tractable, each barcode is limited to no more than the 50 most persistent cycles.

The *Net Stats* clustering approach uses k-means to cluster three-dimensional feature vectors composed of the following network summary statistics: (1) average weighted degree over all nodes (Rubinov & Sporns, 2010), (2) average clustering coefficient over all nodes (Rubinov & Sporns, 2010) and (3) modularity (Newman, 2006; Reichardt & Bornholdt, 2006).

Lastly, we evaluate two clustering approaches based on graph kernels with kernel k-means (Dhillon et al., 2004). One is the *edge histogram* kernel (Sugiyama & Borgwardt, 2015) method. The other is the *propagation* kernel method of (Neumann et al., 2016) with each node assigned the continuous attribute of the sum of edge weights incident to the node.

Initial clusters for all methods are selected at random.

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 displays modular networks with $|V| = 60$ nodes and $m = 5$ modules such that $|V|/m = 12$ nodes are in each module. 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 $\mu = 1$ and $\sigma = 0.5$ universally throughout the study.

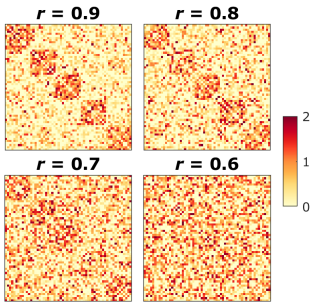

Figure 2: Example $m = 5$ module networks with $|V| = 60$ nodes and within module connection probabilities $r = 0.9, 0.8, 0.7$ and 0.6.

Simulation Three groups of modular networks $L_1 = \{\mathcal{X}_i\}_{i=1}^{20}$, $L_2 = \{\mathcal{X}_i\}_{i=21}^{40}$ and $L_3 = \{\mathcal{X}_i\}_{i=41}^{60}$ corresponding to $m = 2, 3$ and 5 modules, respectively, are simulated. This results in 60 networks in the dataset, each of which has a group label L_1, L_2 or L_3. We consider $r = 0.9, 0.8, 0.7$ and 0.6 to vary the strength of the modular structure, as illustrated in Figure 2.

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

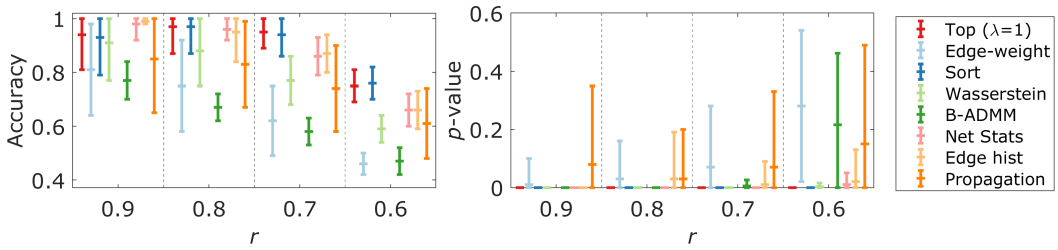

Figure 3: Clustering performance comparison for the dataset of simulated networks with $|V| = 60$ nodes and $m = 2, 3, 5$ modules with respect to average accuracy (left) and average p-values (right). Results for within-module connection probabilities $r = 0.6, 0.7, 0.8$ and 0.9 are shown. Data points (middle horizontal lines) indicate the averages over results for 100 different initial conditions, and vertical error bars indicate standard deviations.

that cluster and then calculating the accuracy statistic s as the fraction of correctly labeled networks, 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 different initial assignments is reported.

Results Figure 3 indicates that all methods considered perform relatively well with pronounced modularity ($r = 0.9$). As the modularity strength diminishes with decreasing r, clustering performance also decreases. The proposed topological clustering algorithm ($\lambda = 1$) performs relatively well for the more nuanced modularity associated with $r = 0.7$ and 0.6. Since the dataset is purposefully generated to exhibit dependency between the sample networks and their group labels, the p-value indicates the degree of statistical significance to which structure is differentiated (Ojala & Garriga, 2010). Small p-values indicate the algorithm is able to differentiate network structure. The proposed method has very low p-values indicating that its improved accuracy over the baseline methods is significant. The *Wasserstein* and B-ADMM methods require very high computation costs for the Rips filtration (Otter et al., 2017), and the approximations of Wasserstein distance (Lacombe et al., 2018) and barycenter (Ye et al., 2017). The ad hoc *sort* method performs nearly identically to the topology approach, showing that the separation into births of connected components and deaths of cycles has no apparent impact for these simulated networks with densely connected structure. However, we hypothesize the separation will impact performance when networks are sparse. The complete experimental results including an additional set of module sizes $m = 2, 5$ and 10 are provided in the supplementary material.

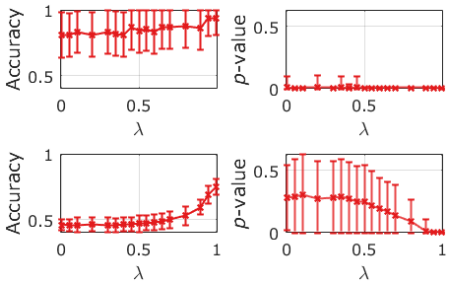

Figure 4: Clustering performance for simulated networks with $|V| = 60$ nodes and $m = 2, 3, 5$ modules as a function of λ for within-module connection probabilities $r = 0.9$ (top row) and $r = 0.6$ (bottom row).

Figure 4 displays topological clustering performance as a function of λ for the strongest ($r = 0.9$) and weakest ($r = 0.6$) degree of modularity. The performance is not very sensitive to the value of λ. When the within-module connection probability is small ($r = 0.6$), increasing the weight of the topological distance by increasing λ results in the best performance. We hypothesize that in this case the relatively strong random nature of the edge weights introduces noise into the geometric term that

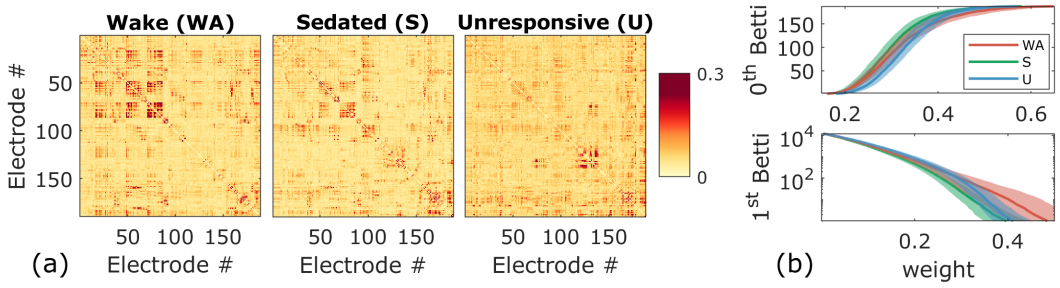

Figure 5: Representative data from a single subject (subject ID R376). For this subject there are 36 measured networks in each of three conditions: wake, sedated, and unresponsive states for a total of 108 measured networks. (a) Sample mean networks during wake, sedated and unresponsive states of the 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.

hinders performance as λ decreases. Conversely, the topological distance appears less sensitive to this particular edge weight noise, resulting in the best performance when $\lambda = 1$.

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 measured 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 cluster analysis. Figure 5 illustrates sample mean networks and Betti plots describing topology for a representative subject.

Performance evaluation We apply the adjusted Rand index (ARI) (Hubert & Arabie, 1985) to compare clustering performance against ground truth. Lower scores indicate less similarity while higher scores show higher similarity between estimated and ground-truth clusters. Perfect agreement is scored as 1.0. For each subject, we calculate these performance metrics by running the algorithm for 20 different initial conditions, resulting in 20 scores which are then averaged. We also average across subjects (11×20 scores) to obtain a final overall score, which describes the overall output clusters across trials and subjects. We calculated average confusion matrices for each method by assigning each cluster to the state that is most frequent in that cluster.

Cluster analysis All baselines used in the simulation study are evaluated on the brain network dataset. In addition, we consider clustering using k-medoids and three previously proposed network distance measures for brain network analyses: *Gromov-Hausdorff* (GH) (Lee et al., 2012); *Kolmogorov-Smirnov* (KS) (Chung et al., 2019); and the *spectral* norm (Banks et al., 2020). Initial clusters are selected at random for all methods. Figure 6 reports the ARI of the topological approach for multiple predefined λ's including $\lambda = 0$ and $\lambda = 1$. The relative performance of topological clustering is reported in Figures 7 and 8 assuming $\lambda = 0.5$, which results in equal weighting of topological and geometric distances.

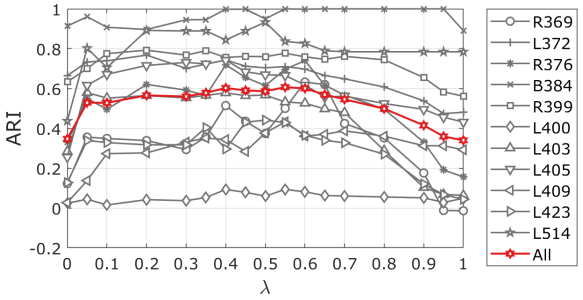

Figure 6: Average ARI performance measures per subject (gray lines) and across all eleven subjects (red line) as a function of λ.

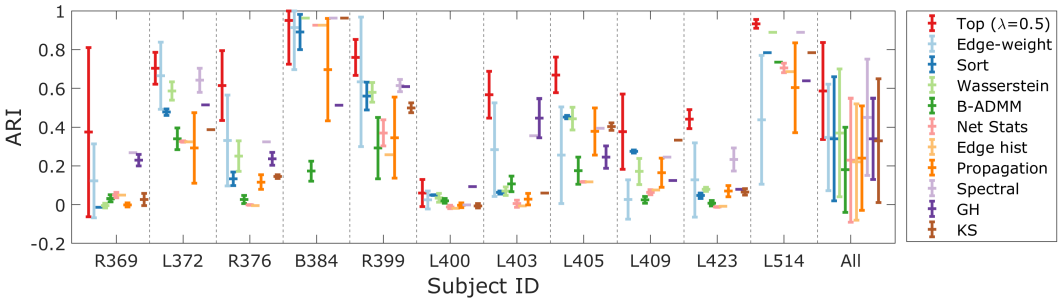

Figure 7: ARI performance for the brain network dataset. Data points (middle horizontal lines) indicate averages over 20 random initial conditions, and vertical error bars indicate standard deviations. Complete experimental results are provided in the supplementary material.

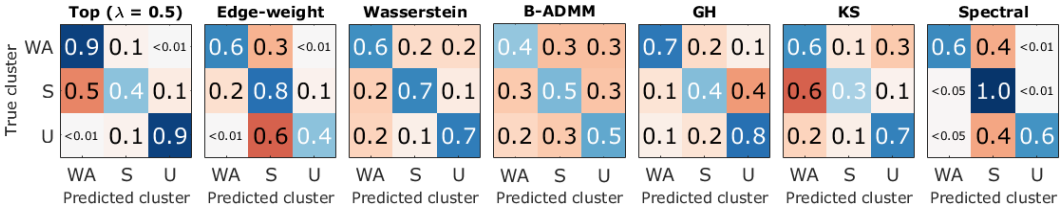

Figure 8: Average confusion matrices over 20 random initializations for select methods for subject R376 (see Figure 5). Results for the other methods are provided in the supplementary material.

Results Figure 6 indicates that the performance on the brain network data set varies significantly across subjects, but in general varies stably with λ. The best performance occurs with λ between 0.4 and 0.7, suggesting that a combination of geometric and topological distances gives the best result and that performance is not sensitive to the exact value chosen for λ. Note that different regions of the brain are functionally differentiated, so it is plausible that topological changes due to changing arousal level vary with location, giving geometric distance a role in the clustering of brain networks. Prior expectations of this sort can be used to determine whether to cluster based only on topology ($\lambda = 1$) or a combination of geometry and topology.

Figure 7 compares the ARI performance metric for individual subjects and the combination of all the subjects. All methods demonstrate significant variability in individual subject performance. This is expected due to the variations in effective signal to noise ratio, the number of nodes, and the number of networks in each underlying state. Consequently the combined performance across subjects has high variability. The topological method with $\lambda = 0.5$ performs relatively well across all subjects. Figure 8 illustrates that proposed approach is particularly effective at separating wake (WA) and unresponsive (U) states. Transitioning from the wake state to the unresponsive state results in dramatic changes in brain connectivity (Banks et al., 2020). The majority of errors from the proposed topological clustering approach are associated with the natural overlap between wake and sedated states (S). The sedated brain, in which subjects have been administered propofol but are still conscious, is expected to be more like the wake brain than the unresponsive brain. Thus, errors are expected to be more likely in differentiating sedated and wake states than sedated and unresponsive states. This suggests that the types of errors observed in the proposed method are consistent with prior biological expectations.

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 derived biomarkers of changes in arousal state presented here 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.

Serguei Barannikov. The framed Morse complex and its invariants. *Advances in Soviet Mathematics, American Mathematical Society*, 1994.

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.

Mathieu Carrière, Frédéric Chazal, Yuichi Ike, Théo Lacombe, Martin Royer, and Yuhei Umeda. Perslay: A neural network layer for persistence diagrams and new graph topological signatures. In *International Conference on Artificial Intelligence and Statistics*, pp. 2786–2796. PMLR, 2020.

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, 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.

Inderjit S Dhillon, Yuqiang Guan, and Brian Kulis. Kernel k-means: Spectral clustering and normalized cuts. In *Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 551–556, 2004.

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.

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

Christoph Hofer, Florian Graf, Bastian Rieck, Marc Niethammer, and Roland Kwitt. Graph filtration learning. In *International Conference on Machine Learning*, pp. 4314–4323. PMLR, 2020.

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.

Théo Lacombe, Marco Cuturi, and Steve Oudot. Large scale computation of means and clusters for persistence diagrams using optimal transport. *Advances in Neural Information Processing Systems*, 2018.

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. 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.

Marion Neumann, Roman Garnett, Christian Bauckhage, and Kristian Kersting. Propagation kernels: Efficient graph kernels from propagated information. *Machine Learning*, 102(2):209–245, 2016.

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.

Nina Otter, Mason A Porter, Ulrike Tillmann, Peter Grindrod, and Heather A Harrington. A roadmap for the computation of persistent homology. *EPJ Data Science*, 6:1–38, 2017.

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, and Bin Yu. Spectral clustering and the high-dimensional stochastic blockmodel. *The Annals of Statistics*, 39(4):1878–1915, 2011.

Mikail Rubinov and Olaf Sporns. Complex network measures of brain connectivity: Uses and interpretations. *NeuroImage*, 52(3):1059–1069, 2010.

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 and Moo K Chung. Topological learning for brain networks. *arXiv preprint arXiv:2012.00675*, 2020.

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.

Mahito Sugiyama and Karsten Borgwardt. Halting in random walk kernels. *Advances in Neural Information Processing Systems*, 28:1639–1647, 2015.

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.

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.