Comparing High Order Networks by Persistence Homology

Weiyu Huang and Alejandro Ribeiro

***Abstract*—This paper presents methods to compare high order networks using persistence homology. High order networks induce well-founded homological features and the difference between networks is measured by the difference between the homological features. This is a reasonable approximation to a valid metric in the space of high order networks modulo permutation isomorphisms. The approximations succeed in identifying collaboration patterns of engineering and math academic journals.**

I. Introduction

We consider high order proximity networks that describe relationships between elements of a tuple and address the problem of constructing valid metric distances between them. Most often, networks are defined as structures that describe interactions between pairs of nodes [1], [2]. This is an indisputable appropriate model for networks that describe binary relationships, such as communication or influence, but not so appropriate for problems in which binary, ternary, or n-ary relationships in general, have different implications. This is, e.g., true of coauthorship networks where we count the number of joint publications by groups of scholars. Papers written by pairs of authors capture information that can be used to identify important authors and study mores of research communities. However, there is extra information to be gleaned from collaborations between triplets of authors, or even single author publications. The importance of capturing tuple proximities between groups of nodes other than pairs has been recognized and exploited in multiple domains [3]–[10].

The problem of defining distances between networks, or, more loosely, the problem of determining if two networks are similar or not, is important even in the case of pairwise networks. The problem is not complicated if nodes have equal labels in both networks [11], [12] but very challenging otherwise, as we need to consider all possible mappings between nodes of each network. This complexity has motivated the use of network features as alternatives to the use of distances. Examples of features that have proved useful in particular settings are clustering coefficients [13], neighborhood topology [14], betweenness [15], motifs [16], wavelets [17], and graphlet-based heuristics [18]–[20]. Feature analysis is valuable, but it does not allow for meaningful comparisons unless application specific features are already known to be important. A different alternative is to define actual distances [21]. Because they have to consider node correspondences, network distances are computationally intractable. Their practical value is limited to small networks and to the transformation of the problem into one of building distance approximations instead of one of searching for appropriate features.

The main problem addressed in this paper is the approximation of the metric distances between high order networks defined in [21]. To achieve this, we relate high order networks to simplicial complexes [22], [23] and relate relationship functions to homological features. The difference between networks is then measured by the difference between the homological features. We justify this is a reasonable approximation. Persistence homology can be computed efficiently for very large networks [24]. We use these approximations to compare the coauthorship networks of academic journals from engineering and math communities and show that they succeed in discriminating their respective collaboration patterns.

II. High Order Networks

A network of order K over the node space X is defined as a collection of $K+1$ relationship functions $\{r_X^k : X^{k+1} \to \mathbb{R}_+\}_{k=0}^K$ from the space X^{k+1} of $(k+1)$-tuples to the nonnegative reals,

$$N_X^K = \left(X, r_X^0, r_X^1, \ldots, r_X^K\right). \tag{1}$$

For point collections $x_{0:k} := (x_0, x_1, \ldots, x_k) \in X^{k+1}$, values of their k-order relationship functions are denoted as $r_X^k(x_{0:k})$ and are intended to represent a measure of similarity or dissimilarity for members of the group. A network of order 0 is one in which only node weights are given, a network of order 1 is one in which weights and pairwise relationships are defined, a network of order 2 adds relationships between triplets and so on. We assume that relationship values are normalized so that $0 \leq r_X^k(x_{0:k}) \leq 1$ for all k and $x_{0:k}$.

We restrict attention to symmetric networks in which for all the $K+1$ functions r_X^k in (1) and $x_{0:k}$, $r_X^k(x_{[0:k]}) = r_X^k(x_{0:k})$ where $x_{[0:k]} = ([x_0], [x_1], \ldots, [x_k])$ is a reordering of $x_{0:k}$. The set of all symmetric networks of order K is denoted as $\mathcal{N}^K$. When defining a distance between networks we need to take into consideration that permutations of r_X^k amount to relabelling nodes and must not be considered as different entities. We therefore say two K-order networks N_X^K and N_Y^K are k-isomorphic if there exists a bijection $\phi : X \to Y$ such that

$$r_Y^k(\phi(x_{0:k})) = r_X^k(x_{0:k}), \tag{2}$$

for all $x_{0:k} \in X^{k+1}$ where $r_Y^k(\phi(x_{0:k})) := r_Y^k(\phi(x_0), \ldots, \phi(x_k))$. The map ϕ is called a k-isometry. When networks N_X^K and N_Y^K are k-isomorphic we write $N_X^K \cong_k N_Y^K$. The space of K-order networks modulo k-isomorphism is denoted by $\mathcal{N}^K \mod \cong_k$. For each nonnegative integer $0 \leq k \leq K$, the space $\mathcal{N}^K \mod \cong_k$ of networks of order K modulo k-isomorphism can be endowed with a pseudometric. The definition of this distance requires introducing the prerequisite notion of correspondence [25].

Definition 1 *A correspondence between two sets X and Y is a subset $C \subset X \times Y$ such that* for all *$x \in X$, there exists $y \in Y$ such that $(x, y) \in C$ and* for all *$y \in Y$ there exists $x \in X$ such that $(x, y) \in C$. The set of all correspondences between X and Y is denoted as $\mathcal{C}(X, Y)$.*

A correspondence in the sense of Definition 1 is a map between node sets X and Y so that every element of each set has a correspondent in the other set. Correspondences include permutations as particular cases but also allow for the mapping of a single point in X to multiple correspondents in Y or, vice versa. Most importantly, this allows definition of correspondences between networks with different numbers of elements. We can now define the distance between two networks by selecting the correspondence that makes them most similar.

Definition 2 *Given networks N_X^K and N_Y^K, a correspondence C between the node spaces X and Y, and an integer $0 \leq k \leq K$ define the k-order network difference with respect to C as*

$$\Gamma_{X,Y}^k(C) := \max_{(x_{0:k}, y_{0:k}) \in C} \left| r_X^k(x_{0:k}) - r_Y^k(y_{0:k}) \right|. \tag{3}$$

The k-order network distance between networks N_X^K and N_Y^K is then defined as

$$d_{\mathcal{N}}^k(N_X^K, N_Y^K) := \min_{C \in \mathcal{C}(X,Y)} \left\{ \Gamma_{X,Y}^k(C) \right\}. \tag{4}$$

For a given correspondence C the network difference $\Gamma_{X,Y}^k(C)$ selects the maximum distance difference $|r_X^k(x_{0:k}) - r_Y^k(y_{0:k})|$ among all pairs of correspondents. The distance in (4) is defined by selecting the correspondence that minimizes these maximal differences. Observe that

Work in this paper is supported by NSF CCF-1217963 and AFOSR MURI FA9550-10-1-0567. The authors are with the Department of Electrical and Systems Engineering, University of Pennsylvania, 200 South 33rd Street, Philadelphia, PA 19104. Email: {whuang, aribeiro}@seas.upenn.edu.

since correspondences may be between networks with different number of elements, Definition 2 defines a pseudometric $d^k_{\mathcal{N}}(N^K_X, N^K_Y)$ when the node cardinalities $|X|$ and $|Y|$ are different. The distance in Definition 2 is a pseudometric in the space of high order network modulo isomorphism [21]. For future reference, the notions of metric and pseudometric are formally stated next.

Definition 3 *Given a space $\mathcal{S}$ and an isomorphism $\cong$, a function $d : \mathcal{S} \times \mathcal{S} \to \mathbb{R}$ is a metric in $\mathcal{S} \mod \cong$ if for any $a, b, c \in \mathcal{S}$ the function d satisfies:*

(i) **Nonnegativity.** $d(a, b) \geq 0$.

(ii) **Symmetry.** $d(a, b) = d(b, a)$.

(iii) **Identity.** $d(a, b) = 0$ *if and only if* $a \cong b$.

(iv) **Triangle inequality.** $d(a, b) \leq d(a, c) + d(c, b)$.

The function is a pseudometric in $\mathcal{S} \mod \cong$ if for any $a, b, c \in \mathcal{S}$ the function d satisfies (i), (ii), (iv), and

(iii') **Relaxed identity.** $d(a, b) = 0$ *if* $a \cong b$.

While different order functions r^k_X and r^l_X of a given network N^K_X need not be related, it is common to observe that adding nodes to a tuple results in decreasing or increasing relationships. This motivates the consideration of dissimilarity and proximity networks that we undertake in the following section.

A. Dissimilarity and Proximity Networks

In dissimilarity networks the function $r^k_X(x_{0:k})$ encodes a level of dissimilarity between elements of the $x_{0:k}$ tuple. In this scenario it is reasonable to assume that adding elements to a tuple makes the group more dissimilar and therefore results in a higher value in the relationship function. In proximity networks the function $r^k_X(x_{0:k})$ encodes a level of similarity or proximity between elements of the tuple. Under this circumstance it is reasonable to assume that adding elements to a tuple makes the group less similar, resulting in a lower value in the relationship function. These restrictions make up the formal definition that we introduce next.

Definition 4 *We say that the K-order network $D^K_X = \left(X, r^0_X, \ldots, r^K_X\right)$ is a dissimilarity network if order increasing property holds, i.e. for any order $1 \leq k \leq K$ and tuples $x_{0:k} \in X^{k+1}$ we have*

$$r^k_X(x_{0:k}) \geq r^{k-1}_X(x_{0:k-1}), \tag{5}$$

and the inequality (5) *equalizes if and only if the point x_k also appears in the point collection $x_{0:k-1}$. We say that the K-order network P^K_X is a proximity network if order decreasing property holds, i.e. under the same conditions we have $r^k_X(x_{0:k}) \leq r^{k-1}_X(x_{0:k-1})$ and the inequality equalizes if and only if the point x_k also appears in the point collection $x_{0:k-1}$. Denote the set of all dissimilarity networks of order K as $\mathcal{D}^K$ and the set of all proximity networks of order K as $\mathcal{P}^K$.*

To see that the order decreasing property in Definition 4 is reasonable, consider the specific case of dissimilarities $r^2_X(x, x')$ and $r^3_X(x, x, x')$, they entail same information as they both convey how different is x from x'. On the other hand, dissimilarities $r^2_X(x, x') \leq r^3_X(x, x, x'')$ and they cannot equalize unless $x = x''$ or $x' = x''$. Order increasing property generalizes this observation and requires dissimilarity $x_{0:k}$ being equal to $x_{0:k-1}$ if and only if the added point x_k is identical to some point in the point collection $x_{0:k-1}$. Further note that since we restricted attention to symmetric networks a relationship as in (5) holds if we remove an arbitrary node from the tuple $x_{0:k}$, not necessarily the last. Thus, the order increasing property implies that removing an element from a tuple can't make the set less dissimilar than it was.

When the input networks in Definition 2 are dissimilarity networks or proximity networks we refer to the k-order distance as the k-order dissimilarity or proximity network distance, respectively. We state this formally in the following definition for future reference.

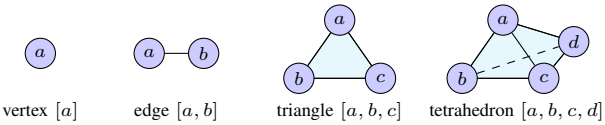

Fig. 1: k-simplices in $\mathbb{R}^{k+1}$ for $0 \leq k \leq 3$.

Definition 5 *Given dissimilarity networks $D^K_X, D^K_Y \in \mathcal{D}^K$ we say that the k-order distance $d^k_{\mathcal{D}}(D^K_X, D^K_Y) = d^k_{\mathcal{N}}(D^K_X, D^K_Y)$ of Definition 2 is the k-order dissimilarity network distance between D^K_X and D^K_Y. The k-order proximity network distance $d^k_{\mathcal{P}}(P^K_X, P^K_Y)$ is defined similarly.*

The restrictions to dissimilarities or proximities make $d^k_{\mathcal{D}}$ a well-defined metric in the space $\mathcal{D}^K \mod \cong_k$ and $d^k_{\mathcal{P}}$ a metric in the space $\mathcal{P}^K \mod \cong_k$ [21]. Proximity and dissimilarity networks have been defined separately for simplicity of presentation, but they are actually related entities. We formalize this equivalence through the introduction of dual networks in the following definition.

Definition 6 *Given a node space X, the K-order proximity and dissimilarity networks $P^K_X = \left(X, p^0_X, \ldots, p^K_X\right)$ and $D^K_X = \left(X, d^0_X, \ldots, d^K_X\right)$ are said duals if and only if*

$$d^k_X(x_{0:k}) = 1 - p^k_X(x_{0:k}), \tag{6}$$

for all orders $0 \leq k \leq K$ and tuples $x_{0:k}$.

The network distance definitions have been constructed such that given dual networks, $d^k_{\mathcal{P}}$ for proximity networks and $d^k_{\mathcal{D}}$ for dissimilarity networks are the same, as we formally state in the following proposition.

Proposition 1 *Consider two proximity networks P^K_X and P^K_Y and their corresponding dual dissimilarity networks D^K_X and D^K_Y. The k-order proximity distances $d^k_{\mathcal{P}}(P^K_X, P^K_Y)$ and k-order dissimilarity distances $d^k_{\mathcal{D}}(D^K_X, D^K_Y)$ coincide for all $0 \leq k \leq K$,*

$$d^k_{\mathcal{D}}(D^K_X, D^K_Y) = d^k_{\mathcal{P}}(P^K_X, P^K_Y). \tag{7}$$

The metrics defined in Definition 5 provides us well-founded methods to compare high order networks. However, the combinatorial nature in searching for the optimal correspondence in (3) makes it impossible to find the exact solution when the number of nodes in networks are large. For this reason, we want to find reasonable as well as computationally tractable approximations to the metrics. The structure of dissimilarity networks relates well to the concept of filtrations in computational homology [22], [23]. This motivates the consideration of using persistence homology to approximate networks distances as we start by giving a brief introduction to computational homology in the following section.

III. Interpretating Dissimilarity Networks as Filtrations

In topology, given k points $x_{0:k}$, one normally considers they live in some $\mathbb{R}^{k+1}$ space where the coordinates of point x_i are all zero except unity on the i-th axis. The k-simplex generated by the set of non-repeating points $x_{0:k}$, $\sigma = [x_{0:k}]$, is defined as the convex hull of of the set of points, $\text{conv}\{x_{0:k}\}$. See Figure 1 for examples of k-simplex with $0 \leq k \leq 3$.

For all $k \geq 1$ removing a point x_s from the set $x_{0:k}$ yields a set with $k-1$ points that we denote as $x_{0:\widehat{s}:k} := x_{0:k} \backslash x_s$. Each of the $k+1$ convex hulls $[x_{0:\widehat{s}:k}] = \text{conv}\{x_{0:k} \backslash x_s\}$ formed by removing the point x_s from the original set is a $(k-1)$-simplex we call a face of σ. For example, the set of faces for the 1-simplex $[a, b]$ and the 2-simplex $[a, b, c]$ in Figure 1 are $\{[a], [b]\}$ and $\{[a, b], [a, c], [b, c]\}$, respectively. Given the simplex $\sigma = [x_{0:k}]$, the boundary $\partial_k \sigma$ of the simplex is the collection of all faces considering orientations, which are generalizations of directed edges in graphs. The boundary of the simplex σ is $\partial_k \sigma = \sum_{s=0}^{k} (-1)^s [x_{0:\widehat{s}:k}]$. Observe that since a 0-simplex σ has no faces, $\partial_0 \sigma = 0$. For the k-simplices in Figure 1, $\partial_0[a] = 0, \partial_1[a, b] = [b] - [a], \partial_2[a, b, c] = [b, c] - [a, c] + [a, b]$ and $\partial_3[a, b, c, d] = [b, c, d] - [a, c, d] + [a, b, d] - [a, b, c]$. A simplicial complex L is a finite collection of simplices such that every face of a simplex of L is also in L and the intersection of any two simplices

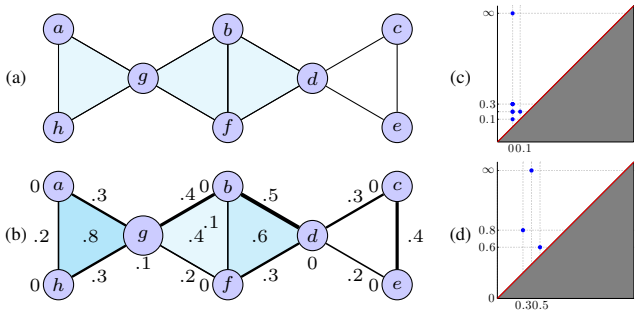

Fig. 2: (a): Two connected bow ties as an example of a simplicial complex which consists 8 0-simplices, 11 1-simplices and 3 2-simplices. The set of faces for the 2-simplex $[a, g, h]$ is $\{[a, g], [a, h], [g, h]\}$. The intersection of the two 2-simplices $[b, f, g]$ and $[b, d, f]$ is an 1-simplex $[b, f]$. (b): A weighted high order network can be represented equivalently as a simplicial complex with weights. The weight of a simplex is the time instant the simplex appears in the simplicial complex. In the original network, $r_X^1(a, g) = 0.3$; in the simplicial complex with weights, the 1-simplex $[a, g]$ appears at time 0.3. (c): the zeroth-dimensional and (d): the first-dimensional persistence diagrams of the filtration induced by (b).

is either empty or a shared face. See Figure 2 (a) for two connected bow ties as an example of a simplicial complex of dimension 2.

A k-chain is a formal sum of k-simplices of L, denoted by $c = \sum_i \beta_i \sigma_i$, where each σ_i is a k-simplex and each β_i is a coefficient. The k-chains together with the addition operation form the group of k-chains, denoted as $\mathsf{C}_k(L)$, or simply C_k. For a k-chain with $c = \sum_i \beta_i \sigma_i$, its boundary is the sum of the boundaries of its simplices, $\partial_k c = \sum_i \beta_i (\partial_k \sigma_i)$. Hence, ∂_k maps a k-chain to a $(k-1)$-chain, $\partial_k : \mathsf{C}_k \to \mathsf{C}_{k-1}$. The sequence of chain groups connected by boundary maps can be represented as

$$\cdots \xrightarrow{\partial_{k+2}} \mathsf{C}_{k+1} \xrightarrow{\partial_{k+1}} \mathsf{C}_k \xrightarrow{\partial_k} \mathsf{C}_{k-1} \xrightarrow{\partial_{k-1}} \cdots . \tag{8}$$

For the connected bow ties in Figure 2 (a), $\mathsf{C}_0 = \beta_1[a] + \cdots + \beta_8[h]$, $\mathsf{C}_2 = \beta_1'[a, g, h] + \beta_2'[b, g, f] + \beta_3'[b, d, f]$. A k-cycle is a k-chain with empty boundary, $\partial c = 0$. In the example, $[a]$ is a 0-cycle and $[a, g]+[g, h]+[a, h]$ is a 2-cycle. Z_k denotes the group of k-cycles and is the kernel of the k-th boundary map, $\mathsf{Z}_k = \ker \partial_k$. Observe that any 0-chain is a 0-cycle, therefore $\mathsf{Z}_0 = \mathsf{C}_0$. A k-boundary is a k-chain that is the boundary of a $(k+1)$-chain, $c = \partial_{k+1} d$ for some $d \in \mathsf{C}_{k+1}$. In the example, $[g]-[h]$ is a 0-boundary since $[g]-[h] = \partial_1[h, g]$ and $[h, g]$ is an 1-chain. B_k denotes the group of k-boundaries and is the image of the $(k+1)$-th boundary map, $\mathsf{B}_k = \operatorname{im} \partial_{k+1}$. The k-th homology group is the k-th cycle group modulo the k-th boundary group, $\mathsf{H}_k = \mathsf{Z}_k / \mathsf{B}_k$. The homology groups considered in this paper are of the form $\mathsf{H}_k \cong \sum_i \gamma_i \Sigma_i$ where each $\gamma_i \in \mathbb{R}$ denotes a degree of freedom and $\Sigma_i = \sum_j \beta_j \sigma_j$ with $\beta_i \in \{-1, 1\}$ is a linear combination of simplices. We say that each Σ_i represents a k-th dimensional homological feature.

We now connect computational topology with dissimilarity networks. Simplicial complexes can be considered as structures of high order networks, detailing the number and labels of vertices, edges, and higher dimensional counterparts. To incorporate relationship functions, we assign each simplex in the simplicial complex L a real value denoting the time when this simplex appears. For any $\alpha \in \mathbb{R}$, we then define $L_\alpha \subseteq L$ to be the collection of simplices appearing before or on time α. If all faces of each simplex and intersections of any simplices in L_α also appear before time or on α, L_α is a well-defined simplicial complex and the nested sequence of $\emptyset = L_{\alpha_0} \subseteq \cdots \subseteq L_{\alpha_m} = L$ with $0 = \alpha_0 < \cdots < \alpha_m$ an ordered sequence of real numbers is defined as a valid filtration $\mathcal{L}$. From Definition 4, if we assign time information based on relationship functions in a given dissimilarity network D_X^K, we have a naturally induced filtration which we denote as $\mathcal{L}(D_X^K)$. See Figure 2 (c) for an example where the numbers adjacent to simplices denote the time when simplices appears in

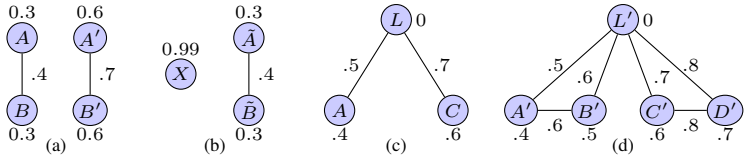

Fig. 3: Examples of dissimilarity networks where the network distances measured by persistence homology exceed the actual metric however are justifiable. Simplices without weights described are considered with the highest dissimilarity 1.

the simplicial complex. L_0 is consisted of all vertices except $[g]$ and L_2 is consisted of all vertices union three 1-simplices $\{[a, h], [b, f], [d, e]\}$.

Finally we give an intuitive definition of persistence homology. Consider the homological feature represented by Σ_i that exists in the k-th homology group $\mathsf{H}_k(L_\alpha)$ for any α satisfying $\alpha_b \le \alpha \le \alpha_d$. This feature starts to appear in the homology group at time α_b as a new independent non-trivial cycle is formed at time α_b and diminishes at time α_d since this cycle is trivialized by a boundary. This formation and diminishment in the homology group of simplicial complexes in a filtration is defined as persistence homology. The interval (α_b, α_d) is named persistence interval for the corresponding homological feature and can also be represented in a two-dimensional diagram. Denote $\mathcal{B}_k\mathcal{L}$ and $\mathcal{D}_k\mathcal{L}$ as the set of birth time and death time of the k-th dimensional homological features of the filtration $\mathcal{L}$. For the example in Figure 2 (b), at time 0, L_0 consists of all 0-simplicies except $[g]$. Since every 0-simplex is a 0-cycle, there exist 7 zeroth-dimensional homological features. At time 0.1, the appearance of the 1-simplex $[b, f]$ makes the 0-cycles $[b]$ and $[f]$ dependent and one zeroth-dimensional homological features dies, generating a zeroth-dimensional persistence interval $(0, 0.1)$. At the same time, a new zeroth-dimensional homological feature represented by $[g]$ appears. As the filtration continues and more edges appear, all zeroth-dimensional homological features disappear except one denoting the entire connected component. For the first-dimension, the homological feature represented by the cycle $[a, h] + [h, g] + [g, a]$ appears at time 0.3 and is killed at time 0.8 by the appearance of the triangle $[a, g, h]$. At the end of the filtration, we have one zeroth-dimensional homological feature born at time 0 and one first-dimensional homological feature represented by $[c, d] + [d, e] + [e, c]$ born at time 0.4. Figure 2 (c) and (d) plot the zeroth-dimensional and the first-dimensional persistence diagrams of the filtration induced by (b). Persistence homologies can be computed with very low cost [24], [26], and we are going to use them to approximate network distances that we undertake in the following section. We will focus on the analysis of dissimilarity networks but it generalizes easily to proximity networks as a direct result of Proposition 1.

IV. Approximating Network Distances Using Persistence

The main challenge in exactly computing network distances as Definition 5 is that we need to search the optimal correspondence between vertices while minimizing the maximal differences for the k-order dissimilarities. It compares k-order functions while considers the complete structures in the networks. A relaxation can be made by comparing k-order functions with no consideration about the network structures and searching the optimal correspondence between k-simplices. More explicitly, given dissimilarity networks D_X^K and D_Y^K, the relaxed network distance can be defined as

$$e_{\mathcal{D}}^k(D_X^K, D_Y^K) = \min_{C^k \in \mathcal{C}(X^{k+1}, Y^{k+1})} \left\{ \max_{(x_{0:k}, y_{0:k}) \in C^k} \left| r_X^k(x_{0:k}) - r_Y^k(y_{0:k}) \right| \right\}, \tag{9}$$

where C^k is a k-correspondence between the powered node spaces X^{k+1} and Y^{k+1} with each element in C^k being a pair of $k+1$ tuples $(x_{0:k}, y_{0,k})$. $e_{\mathcal{D}}^k(D_X^K, D_Y^K)$ can be computed with very low cost and $e_{\mathcal{D}}^k(D_X^K, D_Y^K) \le d_{\mathcal{D}}^k(D_X^K, D_Y^K)$ follows easily.

We want to find an approximation of $d_{\mathcal{D}}^k$ that is more intricate than $e_{\mathcal{D}}^k$ and partly considers the underlying network structures. We are going to

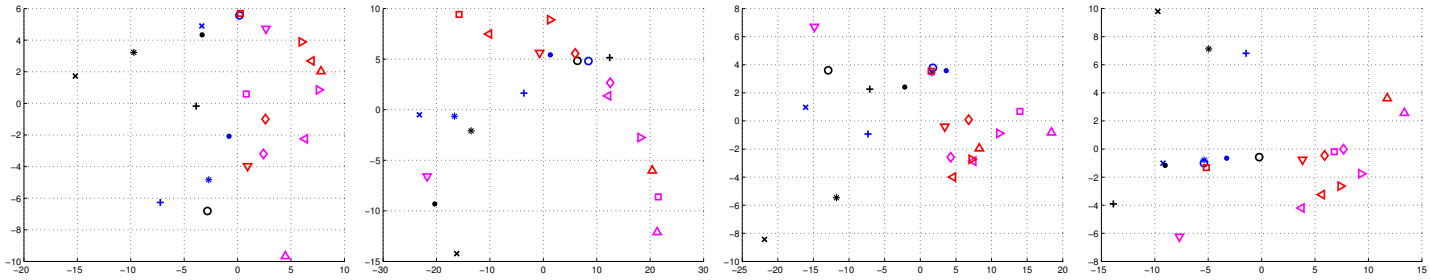

Fig. 4: Multidimensional scale of academic journals based on distances computed using homological features. The distances used are $f^0_{\mathcal{D},b}$, $f^1_{\mathcal{D},d}$, $f^1_{\mathcal{D},b}$, and $f^2_{\mathcal{D},d}$, from left to right respectively. In the figures, blue points denote 2004 - 2008 and black 2009 - 2013 math networks. Red points denote 2004 - 2008 and magenta 2009 - 2013 engineering networks. Points with same symbol represent networks from same journal. For engineering networks: squares represent SPM, diamond TAC, upper triangle TIT, lower triangle TPAMI, left triangle TSP, and right triangle TWC.

achieve this using persistent homology. First observe that follows from the dissimilarity network definition, all the non-trivial k-order dissimilarities of any given dissimilarity network D^K_X can be found in homological features, as we formally state in the following proposition.

Proposition 2 *Given a dissimilarity network D^K_X, any of its k-order dissimilarities between tuples of non-repeating nodes appear either in the death time of the $(k-1)$-th dimensional homological features or the birth time of the k-th dimensional homological features.*

Proof: See [27]. ■

Proposition 2 guarantees that nothing about dissimilarities between non-repeating tuples is lost when we consider the persistence homology of the induced filtration. This motivates us to compare networks by comparing their respective persistence intervals as we formally state next.

Definition 7 *Given dissimilarity networks D^K_X and D^K_Y, the k-order network distances between D^K_X and D^K_Y measured by the death time of the $(k-1)$-th and by the birth time of the k-th dimensional homological features are defined as*

$$f^k_{\mathcal{D},d}(D^K_X, D^K_Y) = \min_{C \in \mathcal{C}(\mathcal{D}_{k-1}\mathcal{L}(D^K_X), \mathcal{D}_{k-1}\mathcal{L}(D^K_Y))} \left\{ \max_{(\alpha_X, \alpha_Y) \in C} |\alpha_X - \alpha_Y| \right\}. \quad (10)$$

$$f^k_{\mathcal{D},b}(D^K_X, D^K_Y) = \min_{C \in \mathcal{C}(\mathcal{B}_k\mathcal{L}(D^K_X), \mathcal{B}_k\mathcal{L}(D^K_Y))} \left\{ \max_{(\alpha_X, \alpha_Y) \in C} |\alpha_X - \alpha_Y| \right\}. \quad (11)$$

In words, $f^k_{\mathcal{D},d}$ considers the maximal differences between the sets $\mathcal{D}_{k-1}\mathcal{L}(D^K_X)$ and $\mathcal{D}_{k-1}\mathcal{L}(D^K_Y)$ of death time for the $(k-1)$-th dimensional homological features and $f^k_{\mathcal{D},b}$ considers the maximal differences between the sets $\mathcal{B}_k\mathcal{L}(D^K_X)$ and $\mathcal{B}_k\mathcal{L}(D^K_Y)$ of birth time for the k-th dimensional homological features. They are related to (9) via $\max\{f^k_{\mathcal{D},d}, f^k_{\mathcal{D},b}\} \geq e^k_{\mathcal{D}}$ since the latter has more freedom in choosing the correspondence. Both $f^k_{\mathcal{D},b}$ and $f^k_{\mathcal{D},b}$ are approximations to $d^k_{\mathcal{D}}$ and there are situations where $f^k_{\mathcal{D},b} > d^k_{\mathcal{D}}$ or $f^k_{\mathcal{D},d} > d^k_{\mathcal{D}}$. We now argue it is reasonable to use the approximations in Definition 7 to compare networks by claiming: (i) the decomposition of dissimilarities into two sets, one representing death time and the other representing birth time, is reasonable, and (ii) it is reasonable that under some scenarios approximation based on persistence homology would exceed the network distance $d^k_{\mathcal{D}}$.

For (i), first observe that as filtration continues, simplices representing more distant relationships are included. When $k = 1$, $f^1_{\mathcal{D},d}$ considers the time when 0-cycles are trivialized and $f^1_{\mathcal{D},b}$ considers the time when 1-cycles are formed. This separation is reasonable since 0-cycles represent connected components or communities and 0-cycles are trivialized when two isolated communities are merged together. The formation of 1-cycles represents the construction of a closed two-way path through at least three nodes that are now pairwise highly similar. When $k = 2$, $f^2_{\mathcal{D},d}$ represents the time when three nodes are not only pairwise highly similar but are highly similar as a single entity; $f^2_{\mathcal{D},b}$ denotes the establishment of a set of at least 4 highly similar nodes, i.e. given any pairs within the set, we can find a third node in the same set such that the triplet as a whole is highly similar. This analysis can be generalized to higher orders.

For (ii), we focus the cases with $k = 1$ due to simplicity and consider examples of dissimilarity networks in Figure 3. The network distance between (a) and (b) and between (c) and (d) measured by persistence homology exceed the actual metric. We here give justifications that the comparisons based on persistence are more reasonable. In comparing (a) and (b), the network metric is 0.19 with the correspondence of pairs $(A, \tilde{A}), (B, \tilde{B}), (A', X), (B', X)$. However notice the dissimilarity of X to itself is 0.99, very close to the maximal dissimilarity, meaning X is likely a noisy observation and therefore mapping both A' and B' to X may not be appropriate. The correspondence using persistence will map both (A, B) and (A', B') to $(\tilde{A}, \tilde{B})$, preserving the structure of pairs. In comparing (c) and (d), persistence will yield high network distance due to the fact that in (d) we have non-trivial first dimensional homological features represented by the cycles L', A', B' and L', C', D'. This is reasonable since they represent three nodes that are pairwise highly similar to each other and we can not find such three nodes in (c).

V. Comparison of Coauthorship Networks

In this section, we exemplify the usage of persistence to compare coauthorship networks constructed from engineering or math academic journals. We used the publicly available database of academic journals from Engineering Village [28] and and selected 5 journals from mathematics community: Computational Geometry, Discrete Computational Geometry, Journal of Applied Probability, Journal of Mathematical Analysis and Applications, SIAM Journal on Numerical Analysis, and 6 journals from engineering community, all from IEEE: Signal Processing Magazine (SPM), Trans. Automatic Control (TAC), Trans. Pattern Analysis and Machine Intelligence (TPAMI), Trans. Information Theory (TIT), Trans. Signal Processing (TSP), Trans. Wireless Communication (TWC). For each journal, we construct two coauthorship networks for quinquennia 2004 - 2008 and 2009 - 2013. These are proximity networks and we transform them into dissimilarity network by Definition 6 and compare the network distance using Definition 7 where persistence homologies are computed using JavaPlex [29]. For visualization, we plot multidimensional scale (MDS) [30] based on the computed distances in Figure 4. It can be seen that network distances truly reflect the difference in community formation of engineering and math journals. A linear classifier on the MDS results would give errors of 1 (4.55%) to 5 (22.73%) out of 22 networks. Moreover, networks constructed from same engineering journal with different quinquenniums tend to be close to each other. This is most conspicuous when the distance used is $f^2_{\mathcal{D},d}$ where networks constructed from same engineering journal form clear clusters.

VI. Conclusion

We relate high order networks to simplicial complexes and use the differences between the induced homological features to measure the differences between networks. We justify that this is a reasonable approximation to a valid metric in the space of high order networks modulo permutation isomorphisms. We use these approximations to successfully identify collaboration patterns of engineering and math academic journals.

REFERENCES

[1] R. K. Ahuja, T. L. Magnanti, and J. B. Orlin, *Network Flows: Theory, Algorithms, and Applications*. Prentice-Hall, Inc., 1993.

[2] S. Wasserman and K. Faust, *Social Network Analysis: Methods and Applications*, ser. Structural Analysis in the Social Sciences. Cambridge University Press, 1994.

[3] R. Ghrist and A. Muhammad, "Coverage and hole-detection in sensor networks via homology," in *International Symposium on Information Processing in Sensor Networks*, vol. 00, no. 1, 2005, pp. 254–260.

[4] V. de Silva and R. Ghrist, "Coordinate-free Coverage in Sensor Networks with Controlled Boundaries via Homology," *The International Journal of Robotics Research*, vol. 25, no. 12, pp. 1205–1222, Dec. 2006.

[5] B. Zhang, "Hypernetworks: A molecular evolutionary architecture for cognitive learning and memory," *Computational Intelligence Magazine, IEEE*, vol. 3, no. 3, pp. 49–63, 2008.

[6] H. Chintakunta and H. Krim, "Divide and Conquer: Localizing Coverage Holes in Sensor Networks," in *2010 7th Annual IEEE Communications Society Conference on Sensor, Mesh and Ad Hoc Communications and Networks (SECON)*, Jun. 2010, pp. 1–8.

[7] W. Ren, Q. Zhao, R. Ramanathan, J. Gao, A. Swami, A. Bar-Noy, M. P. Johnson, and P. Basu, "Broadcasting in multi-radio multi-channel wireless networks using simplicial complexes," in *Wireless Networks*, vol. 19, no. 6, Nov. 2012, pp. 1121–1133.

[8] J. Xu and V. Singh, "Unified Hypergraph for Image Ranking in a Multimodal Context," in *Acoustics, Speech and Signal Processing (ICASSP), 2012 IEEE International Conference on*, 2012, pp. 2333–2336.

[9] Y. Gao, M. Wang, D. Tao, R. Ji, and Q. Dai, "3-D object retrieval and recognition with hypergraph analysis." *IEEE Transactions on Image Processing*, vol. 21, no. 9, pp. 4290–303, Sep. 2012.

[10] A. Wilkerson, T. Moore, A. Swami, and H. Krim, "Simplifying the Homology Of Networks via Strong Collapses," in *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*, 2013, pp. 5258–5262.

[11] S. Segarra, M. Eisen, and A. Ribeiro, "Authorship attribution using function words adjacency networks," in *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*, no. 2, 2013, pp. 5563–5567.

[12] D. Khmelev and F. Tweedie, "Using Markov Chains for Identification of Writer," *Literary and linguistic computing*, vol. 16, no. 3, 2001.

[13] T. Wang and H. Krim, "Statistical classification of social networks," in *Acoustics, Speech and Signal Processing (ICASSP), 2012 IEEE International Conference on*, 2012, pp. 3977–3980.

[14] R. Singh, J. Xu, and B. Berger, "Global alignment of multiple protein interaction networks with application to functional orthology detection." *Proceedings of the National Academy of Sciences*, vol. 105, no. 35, pp. 12 763–12 768, 2008.

[15] L. Peng, L. Liu, S. Chen, and Q. Sheng, "A network comparison algorithm for predicting the conservative interaction regions in protein-protein interaction network," in *2010 IEEE Fifth International Conference on Bio-Inspired Computing: Theories and Applications (BIC-TA)*, Sep. 2010, pp. 34–39.

[16] S. Choobdar, P. Ribeiro, S. Bugla, and F. Silva, "Comparison of Co-authorship Networks across Scientific Fields Using Motifs," *2012 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, pp. 147–152, Aug. 2012.

[17] L. Yong, Z. Yan, and C. Lei, "Protein-protein interaction network comparison based on wavelet and principal component analysis," in *2010 IEEE International Conference on Bioinformatics and Biomedicine Workshops (BIBMW)*, 2010, pp. 430–437.

[18] N. Pržulj, "Biological network comparison using graphlet degree distribution." *Bioinformatics*, vol. 23, no. 2, pp. e177–183, Jan. 2007.

[19] T. Milenković and N. Pržulj, "Uncovering biological network function via graphlet degree signatures." *Cancer informatics*, vol. 6, p. 257, Jan. 2008.

[20] N. Shervashidze, S. Vishwanathan, T. H. Petri, K. Mehlhorn, and K. M. Borgwardt, "Efficient graphlet kernels for large graph comparison," in *International Conference on Artificial Intelligence and Statistics*, vol. 5, 2009, pp. 488–495.

[21] W. Huang and A. Ribeiro, "Metrics in the Space of High Order Networks," *Signal Processing, IEEE Transactions on*, vol. (revised), 2015.

[22] H. Edelsbrunner, D. Letscher, and A. Zomorodian, "Topological Persistence and Simplification," in *Proceedings of the 41st Annual Symposium on Foundations of Computer Science*, 2000, pp. 454–463.

[23] A. Zomorodian and G. Carlsson, "Computing persistent homology," *Discrete Comput. Geo*, vol. 33, pp. 249–274, 2005.

[24] K. Mischaikow and V. Nanda, "Morse theory for filtrations and efficient computation of persistent homology," *Discrete & Computational Geometry*, vol. 50, no. 2, pp. 330–353, 2013.

[25] D. Burago, Y. Burago, and S. Ivanov, *A course in metric geometry*. American Mathematical Society Providence, 2001, vol. 33.

[26] S. Harker, K. Mischaikow, M. Mrozek, and V. Nanda, "Discrete Morse Theoretic Algorithms for Computing Homology of Complexes and Maps," *Foundations of Computational Mathematics*, vol. 14, no. 1, pp. 151–184, 2014.

[27] W. Huang and A. Ribeiro, "Comparing High Order Networks using Persistence Homology," *Signal Processing, IEEE Transactions on*, vol. (submitted), 2015.

[28] "Engineering Village: the place to find answers to engineering questions." [Online]. Available: http://www.engineeringvillage.com/search/quick.url

[29] "JPLEX: Persistent Homology Computations Library." [Online]. Available: http://comptop.stanford.edu/u/programs/jplex/index.html

[30] M. A. A. Cox and T. F. Cox, "Multidimensional scaling," in *Handbook of Data Visualization*, ser. Springer Handbooks Comp.Statistics. Springer Berlin Heidelberg, 2008, pp. 315–347.