Persistent Homology Lower Bounds on Distances in the Space of Networks

Weiyu Huang and Alejandro Ribeiro

***Abstract*—High order networks are weighted hypergraphs collecting relationships between elements of tuples, not necessarily pairs. Valid metric distances between high order networks have been defined but they are difficult to compute when the number of nodes is large. The goal here is to find tractable approximations of these network distances. The paper does so by mapping high order networks to filtrations of simplicial complexes and showing that the distance between networks can be lower bounded by the difference between the homological features of their respective filtrations. Practical implications are explored by classifying the coauthorship networks of engineering and mathematics academic journals. The persistent homology methods succeed in discriminating engineering and mathematics communities, and in differentiating engineering communities with different research interests.**

***Index Terms*—Network theory, Networked data, Topology, Network topology, Pattern recognition**

I. Introduction

High order networks describe relationships between tuples that go beyond the pairwise relationship that are more often considered [2]. High order networks are important in applications in which relationships between triplets, quadruplets, and generic n-tuples are important; e.g. coauthorship networks can be constructed by detailing the number of joint publications by pairs of scholars, but adding information on publications of individual authors and joint publications by specific triplets of authors generates a more accurate description. The relevance of expressing tuple relationships other than pairwise has been noted in various domains [3]–[5].

This paper aims to develop efficient methods to quantify differences between high order networks of possibly different sizes. Ideally, differences should be quantified using metrics in the space of high order networks. However, while distances are not difficult to define, they can be hard to compute. This difficulty has motivated the use of feature comparisons in which the difference between specific properties of the network is used as an alternative [6], [7]. While computationally simpler, the use of features is application dependent, utilizes only a small portion of the information conveyed by the networks, and may yield conflicting comparative judgements because the triangle inequality is not necessarily valid. A proper distance between pairwise and high order networks overcomes these drawbacks. This is the motivation for the definition of network metrics for pairwise [8] and high order networks [9].

The metric distances between high order networks defined in [9] have been applied to compare networks with small number of nodes and have succeeded in identifying collaboration patterns of coauthorship networks [10]. However, network distances are difficult to compute when the number of nodes in the networks is large. The goal of this paper is to develop network discrimination methods that are computable in large networks. These methods are constructed by drawing a parallel between high order networks and algebraic topology filtrations. Homological features of filtrations are then used to compare high order networks and shown to provide *lower bounds* for the actual network distances. Although distance lower bounds suffer from some of the same problems associated with feature comparisons, they nonetheless have important properties. Among them, a large lower bound entails a large distance and that we can use lower bounds to estimate distance intervals because upper bounds are easy to determine. The proposed methods succeed in discriminating engineering journals from mathematics journals. We also attempt a more challenging classification problem of three different engineering communities where we achieve a moderate success.

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. A journal version of this paper can be found in [1].

II. High Order Networks

A network of order K over the node space X is defined as a collection of $K+1$ relationships $\{r_X^k : X^{k+1} \to [0,1]\}_{k=0}^K$ from the space X^{k+1} of $(k+1)$-tuples to unit interval,

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

For point collections $x_{0:k} := (x_0, \ldots, x_k) \in X^{k+1}$, values of their k-order relationships 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. 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 $x_{[0:k]} = ([x_0], \ldots, [x_k])$ 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 should 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 $\pi : X \to Y$ such that $r_Y^k(\pi(x_{0:k})) = r_X^k(x_{0:k})$ for all $x_{0:k} \in X^{k+1}$ where $r_Y^k(\pi(x_{0:k})) := r_Y^k(\pi(x_0), \ldots, \pi(x_k))$. The map π 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 $\mathcal{N}^K \mod \cong_k$. The space $\mathcal{N}^K \mod \cong_k$ can be endowed with a pseudometric [9]. The definition of this distance requires introducing the prerequisite notion of correspondence [11].

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 at least one correspondent in the other set. 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*, the* k*-order network distance between* N_X^K *and* N_Y^K *is then defined as*

$$d_{\mathcal{N}}^k(N_X^K, N_Y^K) := \min_{C \in \mathcal{C}(X,Y)} \left\{ \max_{(x_{0:k}, y_{0:k}) \in C} \left| r_X^k(x_{0:k}) - r_Y^k(y_{0:k}) \right| \right\}. \tag{2}$$

For a given correspondence C the network difference selects the maximum relationship difference $|r_X^k(x_{0:k}) - r_Y^k(y_{0:k})|$ among all pairs of correspondents. The distance in (2) is defined by selecting the correspondence that minimizes these maximal differences. Since correspondences may be between networks with different number of nodes, $d_{\mathcal{N}}^k(N_X^K, N_Y^K)$ is well-defined when the cardinalities $|X|$ and $|Y|$ are different. While different order functions r_X^k and r_X^l of N_X^K need not be related, it is common to observe that adding nodes to a tuple results in decreasing or increasing relationships. This motivates the study of dissimilarity and proximity networks.

In dissimilarity networks the function $r_X^k(x_{0:k})$ 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 results in a higher value in the relationship function. In proximity networks the function $r_X^k(x_{0:k})$ encodes a level of proximity between elements of the tuple. Hence it is reasonable to assume that adding elements to a tuple makes the group less similar, resulting in a lower value in the relationship. These restrictions make up the formal definition that we introduce next.

Definition 3 *The* K*-order network* $D_X^K = \left(X, r_X^0, \ldots, r_X^K\right)$ *is a dissimilarity network if order increasing property holds, i.e. for any* $1 \leq k \leq K$ *and tuples* $x_{0:k} \in X^{k+1}$ *we have*

$$r_X^k(x_{0:k}) \geq r_X^{k-1}(x_{0:k-1}), \tag{3}$$

and the inequality (3) *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_X^K *is a proximity network if order decreasing property holds, i.e. under the same conditions we have* $r_X^k(x_{0:k}) \leq r_X^{k-1}(x_{0:k-1})$ *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 of all proximity networks of order* K *as* $\mathcal{P}^K$.

When the input networks in Definition 2 are dissimilarity or proximity networks we refer to the k-order distance as the k-order dissimilarity or proximity network distances. The restrictions make $d_{\mathcal{D}}^k$ a well-defined metric in the space $\mathcal{D}^K \mod \cong_k$ and $d_{\mathcal{P}}^k$ a metric in the space $\mathcal{P}^K \mod \cong_k$ [9]. Proximity and dissimilarity networks are related entities,

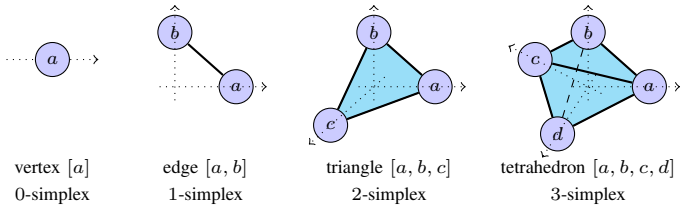


Fig. 1: Elementary k-simplices for $0 \leq k \leq 3$.

$P_X^K = (X, p_X^0, \ldots, p_X^K)$ and $D_X^K = (X, d_X^0, \ldots, d_X^K)$ are said duals when $d_X^k(x_{0:k}) = 1 - p_X^k(x_{0:k})$ for all orders $0 \leq k \leq K$ and tuples $x_{0:k}$. The k-order proximity distances and k-order dissimilarity distances between respective dual networks coincide, i.e. $d_{\mathcal{D}}^k(D_X^K, D_Y^K) = d_{\mathcal{P}}^k(P_X^K, P_Y^K)$ when D_X^K and P_X^K as well as D_Y^K and P_Y^K are duals.

The metrics defined in Definition 2 provide us well-founded methods to compare high order networks. However, the combinatorial nature in searching for the optimal correspondence in (2) makes it impossible to find the exact solution when the number of nodes in the networks is large. This motivates the development of reasonable and tractable lower bounds. These bounds will be obtained by relating dissimilarity networks to filtrations in homology [12] as we discuss next.

III. Networks and Simplicial Complexes

We introduce elemental notions of topology as they apply to the study of high order networks. Begin by defining a k-simplex $\phi = [x_{0:k}]$ as the convex hull of the set of points $x_{0:k}$ (see Fig. 1) and a simplicial complex L as the collection of simplices such that for any simplex $[x_{0:k}]$, the convex hull of any subset of $x_{0:k}$ also belongs to L. Fig. 2 (a) exemplifies two triangles connected together as a simplicial complex. It is a simplicial complex because for any of its simplices, say, $[a, b, c]$, the convex hulls of the subsets of $\{a, b, c\}$ – which include the points $[a]$, $[b]$, and $[c]$ as well as the edges $[a, b]$, $[b, c]$, and $[a, c]$ – all belong to the simplicial complex as well.

An important concept in simplicial complexes is that of a hole without interior. For the simplicial complex in Fig. 2 (a), the area enclosed by $[b, d], [d, c], [c, b]$ is a hole without interior because the area is not filled. The area enclosed by $[a, b], [b, d], [d, a]$ is not because the interior is filled by the 2-simplex $[a, b, d]$. Homologies are defined to formalize this intuition and rely on the definitions of chains, cycles, and boundaries. The *k-chain* $\Phi_k = \sum_i \beta_i \phi_i$ is a summation of k-simplices ϕ_i modulated by coefficients β_i whose signs denote orientation. This definition is a generalization of the familiar definition of chains in graphs. E.g., in Fig. 2, $[a, b] + [b, d]$ is a 1-chain that we can equivalently represent as $[a, b] - [d, b]$. Further consider a given k-simplex $\phi = [x_{0:k}]$ and its border $(k-1)$-simplices defined as the ordered set of elemenets $[x_{0:\widehat{l}:k}] = \text{conv}\{x_{0:k} \backslash x_l\}$, in which each of the elements is removed in order. The *boundary* $\partial_k \phi$ of the simplex ϕ is the chain formed by its borders using alternating orientations,

$$\partial_k \phi = \sum_{l=0}^{k} (-1)^l [x_{0:\widehat{l}:k}]. \tag{4}$$

The boundary $\partial_k \phi$ of a k-simplex is the collection of $(k-1)$-simplices. Having defined chains and boundaries we consider a

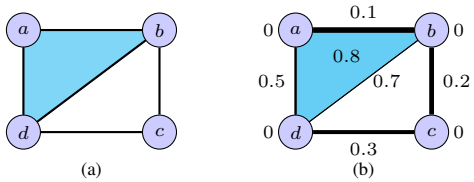


Fig. 2: (a) An example of simplicial complex L which consists of four 0-simplices, five 1-simplices, and one 2-simplex. (b) A dissimilarity network can be represented as a simplicial complexes with weights.

k-chain $\Phi_k = \sum_i \beta_i \phi_i$ and define the *chain boundary* $\partial_k \Phi_k$ as the summation of the boundaries of its component simplices,

$$\partial_k \Phi_k = \sum_i \beta_i (\partial_k \phi_i). \tag{5}$$

The boundary of the chain $\Phi = [a,b] + [b,d]$ in Fig. 2 is $\partial_k \Phi = [a] - [b] + [b] - [d] = [a] - [d]$. We can now define a *$k$-cycle* Ψ_k as a chain whose boundary is null, i.e., a k-chain for which $\partial_k \Psi = 0$. In Fig. 2, the chain $\Psi = [a,b] + [b,d] + [d,a]$ is a cycle because its boundary is $\partial_k \Psi = [a] - [b] + [b] - [d] + [d] - [a] = 0$. A k-cycle Ψ_k is a *k-homological feature* if it cannot be represented as the boundary of a $(k+1)$-chain Φ_{k+1}; otherwise Ψ_k is a *k-boundary*. The group $\mathcal{H}_k(L)$ of k-homological features of a simplicial complex L represents all k-cycles that are not k-boundaries.

To represent weighted high order networks we assign a weight to each simplex, but instead of thinking the network as a weighted simplicial complex we think of weights as parameters that indicates the time at which the simplex comes into existence. Formally, for parameters $\alpha \in [0,1]$ we define a filtration $\mathcal{L}$ as a collection of simplicial complexes L_α such that for any ordered sequence $0 = \alpha_0 < \alpha_1 < \ldots < \alpha_m = 1$ it holds $\emptyset = L_{\alpha_0} \subseteq L_{\alpha_1} \subseteq \ldots \subseteq L_{\alpha_m} = L$. The minimum time α at which a simplex becomes an element of L_α is the birth time of the simplex. Given a dissimilarity network D_X^K, we construct a filtration $\mathcal{L}(D_X^K)$ by assigning the appearing time of simplex $[x_{0:k}]$ as the relationship $r_X^k(x_{0:k})$ of the corresponding tuple,

$$[x_{0:k}] \in L_\alpha \iff r_X^k(x_{0:k}) \leq \alpha. \tag{6}$$

Fig. 3 (top) shows the construction of the filtration associated with the dissimilarity network in Fig. 2 (b). At time $\alpha = 0$, there are four nodes because $r_X^0(a) = r_X^0(b) = r_X^0(c) = r_X^0(d) = 0$. At time $\alpha = 0.1$, the edge $[a,b]$ starts to appear in the simplicial complex because the relationship $r_X^1(a,b) = 0.1$. As time gradually increases, more simplices get included in the simplicial complex at each time instant.

Persistent homologies examine when these homological features appear for the sequence of simplicial complexes in the filtration. Formally, given a filtration $\mathcal{L}$, its k-dimensional persistence diagram $\mathcal{P}_k\mathcal{L}$ is a collection of points of the form $\mathbf{q} = [q_b, q_d]$ where q_b and $q_d > q_b$ represent the birth and death time of a homological feature. I.e., the times q_b represent resolutions $\alpha = q_b$ at which a feature is added to the group of k-homological features $\mathcal{H}_k(L_\alpha) = \mathcal{H}_k(L_{q_b})$ and the times q_d are resolutions $\alpha = q_d$ at which a feature is removed from the group of k-homological features $\mathcal{H}_k(L_\alpha) = \mathcal{H}_k(L_{q_d})$.

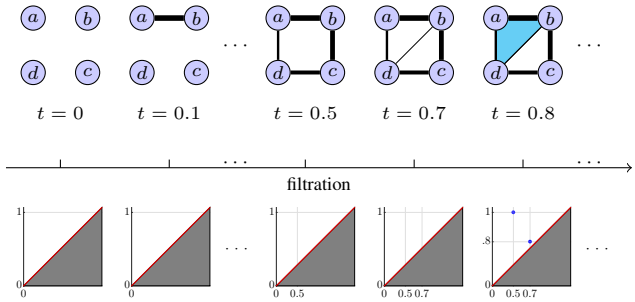


Fig. 3: Top: filtration of nested simplicial complexes of the dissimilarity network exhibited in Fig. 2 (b). The simplicial complex at each time instant are the collections of all vertices appearing before or on that time. Bottom: 1-persistence diagrams describing 1-persistent homologies for each of the simplicial complexes detailed in Top. The horizontal axis denotes the birth time of homological features and the vertical axis represents the death time.

The persistence diagram for the filtration in Fig. 3-(top) is shown in Fig. 3-(bottom). In this diagram the horizontal axis denotes the birth time of 1-homological features (when holes appear) and the vertical axis represents the death time of 1-homological features (when holes are filled). At time $\alpha = 0$, there are no 1-simplices and consequently no 1-holes. The first 1-cycle appears at time $\alpha = 0.5$ when the 1-chain $[a,b] + [b,c] + [c,d] + [d,a]$ appears. We mark this event by the addition of a vertical line in the persistence diagram. A second homological feature appears in the form of the cycle $[a,b] + [b,d] + [d,a]$ at time $\alpha = 0.7$. This event is marked by the addition of a second vertical line in the persistence diagram. This homological feature dies at time $\alpha = 0.8$ when the 2-simplex $[a,b,d]$ appears and makes the 1-cycle $[a,b] + [b,d] + [d,a]$ a 1-boundary and therefore no longer a homological feature. This is marked by the addition of the *point* $\mathbf{q}_1 = (0.7, 0.8)$ to the persistence diagram. At this time all simplices have been added to the filtration. This means that the homological feature $[a,b] + [b,c] + [c,d] + [d,a]$ never becomes a boundary of a 2-simplex. This is marked by adding the point $\mathbf{q}_2 = (0.5, 1)$ to the persistence diagram.

IV. Persistence Bounds on Network Distances

In this section we use differences between persistence diagrams to compute lower bounds of network distances. Recall a persistence diagram $\mathcal{P}_k\mathcal{L}(D_X^K)$ is a collection of points of the form $\mathbf{q} = [q_b, q_d]$ where q_b and $q_d > q_b$ represent the birth and death time of a k-homological feature. To compare persistence diagrams $\mathcal{P}_k\mathcal{L}(D_X^K)$ and $\mathcal{P}_k\tilde{\mathcal{L}}(D_Y^K)$ of networks D_X^K and D_Y^K we begin by defining the cost of matching features $\mathbf{q} \in \mathcal{Q}$ and $\tilde{\mathbf{q}} \in \tilde{\mathcal{Q}}$ through the infinity norm of their difference,

$$\|\mathbf{q} - \tilde{\mathbf{q}}\|_\infty := \max\left[|q_b - \tilde{q}_b|, |q_d - \tilde{q}_d|\right]. \tag{7}$$

Further observe that the diagonal of a persistence diagram can be construed to represent an uncountable number of features with equal birth and death times. Thus, any feature $\mathbf{q} \in \mathcal{Q}$ can be compared to any of the artificial features of the form $\tilde{\mathbf{q}} = [\tilde{q}, \tilde{q}]$. Since this feature $\tilde{\mathbf{q}}$ can be placed anywhere in the diagonal of the persistence diagram, we choose to place it

in the point that makes the infinity norm difference smallest. This point is $\tilde{\mathbf{q}} = [(q_b + q_d)/2, (q_b + q_d)/2]$ which yields the difference $\|\mathbf{q} - \tilde{\mathbf{q}}\|_\infty = |q_d - q_b|/2$. Likewise, artificial features $\mathbf{q} = [(\tilde{q}_b + \tilde{q}_d)/2, (\tilde{q}_b + \tilde{q}_d)/2]$ can be added for any point $\tilde{\mathbf{q}} \in \tilde{\mathcal{Q}}$. We can then rephrase the comparison in (7) so that if it is more advantageous to compare with artificial diagonal features. This is formally accomplished by defining the matching cost $c(\mathbf{q}, \tilde{\mathbf{q}})$ between $\mathbf{q}$ and $\tilde{\mathbf{q}}$ as

$$c(\mathbf{q}, \tilde{\mathbf{q}}) := \min\Big[\|\mathbf{q} - \tilde{\mathbf{q}}\|_\infty, (1/2)\max\big[\,|q_d - q_b|, |\tilde{q}_d - \tilde{q}_b|\big]\Big], \tag{8}$$

The norm $\|\mathbf{q} - \tilde{\mathbf{q}}\|_\infty$ is the cost of directly matching features $\mathbf{q}$ and $\tilde{\mathbf{q}}$. The term $(1/2)\max\big[\,|q_d - q_b|, |\tilde{q}_d - \tilde{q}_b|$ is the cost of matching both, $\mathbf{q}$ and $\tilde{\mathbf{q}}$ to artificial diagonal features. The cost $c(\mathbf{q}, \tilde{\mathbf{q}})$ of matching $\mathbf{q}$ and $\tilde{\mathbf{q}}$ is the smaller of these two.

When the number of points in the persistence diagrams are different we assume without loss of generality that $\tilde{m} < m$. In this case we extend $\tilde{\mathcal{Q}}$ by adding $\tilde{m} - m$ diagonal features to define the set $\tilde{\mathcal{Q}}_e := \tilde{\mathcal{Q}} \cup \{(\tilde{q}_i, \tilde{q}_i)\}_{i=1}^{\tilde{m}-m}$. Since the sets $\mathcal{Q}$ and $\tilde{\mathcal{Q}}_e$ contain the same number of elements bijections $\pi : \mathcal{Q} \to \tilde{\mathcal{Q}}$ are well defined. We can then define a valid comparison between arbitrary $\mathcal{P}_k\mathcal{L}$ and $\mathcal{P}_k\tilde{\mathcal{L}}$ as next.

Definition 4 *Given persistence diagrams $\mathcal{P}_k\mathcal{L}(D_X^K)$ and $\mathcal{P}_k\tilde{\mathcal{L}}(D_Y^K)$ with sets of points $\mathcal{Q}$ and $\tilde{\mathcal{Q}}$ having cardinalities $\tilde{m} < m$, define the extended set $\tilde{\mathcal{Q}}_e := \tilde{\mathcal{Q}} \cup \{(\tilde{q}_i, \tilde{q}_i)\}_{i=1}^{\tilde{m}-m}$ by adding $\tilde{m} - m$ artificial diagonal features. The bottleneck distance between the persistence diagrams $\mathcal{P}_k\mathcal{L}(D_X^K)$ and $\mathcal{P}_k\tilde{\mathcal{L}}(D_Y^K)$ of networks D_X^K and D_Y^K is defined as*

$$b^k(D_X^K, D_Y^K) := \min_{\pi:\mathcal{Q}\to\tilde{\mathcal{Q}}_e} \max_{\mathbf{q}\in\mathcal{Q}} c(\mathbf{q}, \pi(\mathbf{q})), \tag{9}$$

where π ranges over all bijections from $\mathcal{Q}$ to $\tilde{\mathcal{Q}}_e$ and the cost $c(\mathbf{q}, \pi(\mathbf{q}))$ is defined in (8)*.*

The number of bijections between two sets of points is factorial and it appears that the problem as in (9) is as difficult as the problem of finding the correspondence in evaluating the network distance. However, the problem in (9) is a instantiation of the Linear Bottleneck Assignment Problem (LBAP) that can be solved efficiently. We emphasize that $c(\mathbf{q}, \tilde{\mathbf{q}}) = c(\mathbf{q}, \tilde{\mathbf{q}}')$ for any $\mathbf{q}$ whenever $\tilde{\mathbf{q}}$ and $\tilde{\mathbf{q}}'$ are on the diagonal. Therefore, the locations of diagonal points added to construct $\mathcal{Q}_e$ are unsubstantial as per their cost.

The bottleneck distance between the persistence diagrams of the filtrations induced by two dissimilarity networks is a lower bound of their dissimilarity network distance.

Theorem 1 *Given two dissimilarity networks D_X^K and D_Y^K and $1 \le k \le K$, the bottleneck distance between the k'-th dimensional persistence diagrams of the filtrations $\mathcal{L}(D_X^K)$ and $\mathcal{L}(D_Y^K)$ is at most $d_{\mathcal{D}}^k(D_X^K, D_Y^K)$ for any $0 \le k' < k$,*

$$b^{k'}(D_X^K, D_Y^K) \le d_{\mathcal{D}}^k(D_X^K, D_Y^K). \tag{10}$$

Proof: The proof is technically rich and therefore we refer readers to [1] for details. ■

For any proximity network P_X^K with relationships $p_X^k(x_{0:k})$ we can define the dual network $\bar{D}_X^K$ with relationship functions $d_X^k(x_{0:k}) := 1 - p_X^k(x_{0:k})$. It follows that the bounds in Theorem 1 apply for proximity networks with the caveat that the filtrations $\mathcal{P}_k\mathcal{L}(D_X^K)$ and $\mathcal{P}_k\mathcal{L}(D_Y^K)$ are filtrations of the dual networks associated with the given proximity networks.

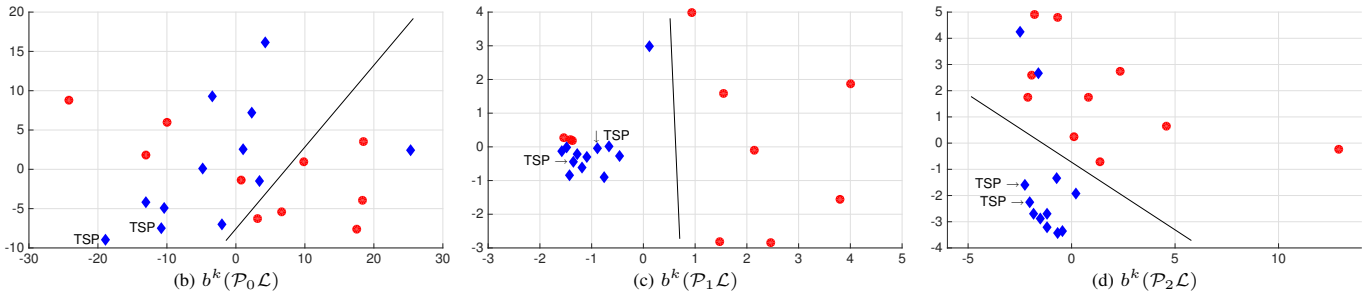
V. Applications

We apply the lower bounds to compare 2-order coauthorship networks where relationships denote the number of publications of single authors, pairs of authors, and triplets. We consider publications in 5 journals from mathematics: Computational Geometry (CG), Discrete Computational Geometry (DCG), J. of Applied Probability, (JAP) J. of Mathematical Analysis and Applications (JMAA), SIAM J. on Numerical Analysis (SJNA), and 6 journals from engineering, all from IEEE: Signal Processing Magazine (SPM), Trans. Automatic Control (TAC), Trans. Pattern Analysis and Machine Intelligence (TPAMI), Trans. Information Theory, Trans. Signal Processing (TSP), Trans. Wireless Communication (TWC). For each journal, we construct networks for the 2004-2008 and 2009-2013 quinquennia. For TAC, TSP, and TWC, we also construct networks for each annual from 2004 to 2013. Lists of publications are queried from [13].

For each of these journals we consider all publications in the period of interest and construct proximity networks where the node set X is formed by all authors of the publications. Zeroth order proximities are defined as the total number of publications of each member of the network, first order proximities as the number of papers coauthored by pairs, and second order proximities as the number of papers coauthored by triplets. To make networks with different numbers of papers comparable we normalize all relationships by the total number of papers in the network. There are papers with more than three coauthors but we don't record proximities of order higher than 2. By assuming that networks from the same community have similar collaboration patterns, we show here that network metric lower bounds succeed in identifying these patterns.

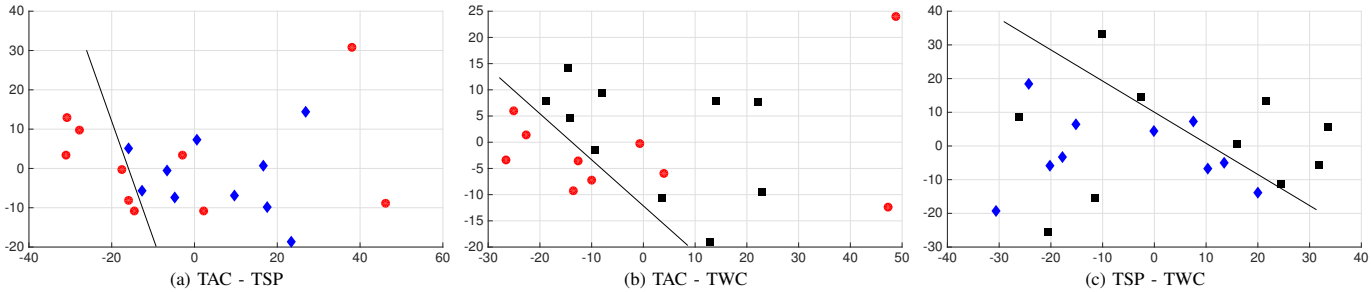
Fig. 4 shows the two dimensional Euclidean embeddings of the network metric lower bounds b^0, b^1 and b^2. The 12 engineering networks (blue diamonds) separate clearly from the 10 mathematics networks (red circles) in b^1 and b^2. The clustering is not that clear in b^0 but still networks from same community tend to be similar to each other. An unsupervised classification with one linear boundary run across the embeddings would generate errors of 2 (9.09%) to 5 (22.73%) out of 22 networks. Networks constructed from the same journal tend to be close in the lower bounds; e.g. networks of TSP with different quinquennia are marked in the embeddings and it is clear that their differences in homologies are considerably low. Similar observations hold for several other journals as well.

We analyze the persistent homology of these networks to investigate why persistent homologies succeed in discrimination. Compared to networks from mathematics communities, networks from engineering communities in general would yield 0-dimensional persistent homologies with smaller birth time but larger death time, 1-dimensional homologies with larger birth and death time, and 2-dimensional homologies with larger birth time. An interpretation would be that in engineering, there exist more small communities that never collaborate with each other and it is uncommon to have a "club" of 3 to 5 authors in engineering that a strong



(b) $b^k(\mathcal{P}_0\mathcal{L})$ (c) $b^k(\mathcal{P}_1\mathcal{L})$ (d) $b^k(\mathcal{P}_2\mathcal{L})$

Fig. 4: Two dimensional Euclidean embeddings of the networks constructed from quinquennial publications in engineering and mathematics journals w.r.t network metric lower bounds $b^k(\mathcal{P}_0\mathcal{L})$, $b^k(\mathcal{P}_1\mathcal{L})$, and $b^k(\mathcal{P}_2\mathcal{L})$. Red circles denote networks constructed from mathematics journals and blue diamonds represent networks from engineering journals. TSP are labeled.



(a) TAC - TSP (b) TAC - TWC (c) TSP - TWC

Fig. 5: Embeddings of the networks constructed from annual publications in TAC, TSP, and TWC w.r.t. the summation of the lower bounds $b^k(\mathcal{P}_0\mathcal{L})$, $b^k(\mathcal{P}_1\mathcal{L})$, and $b^k(\mathcal{P}_2\mathcal{L})$. Red circles represent TAC, blue diamonds TSP, and black squares TWC.

collaboration exists between any pairs of the authors in the "club"; such scenarios are absent for mathematician.

As a comparison, we applied some simple and reasonable methods to compare the coauthorship networks [7]. Methods to compare networks via features give us similar observations as those based on the persistent methods, but generally yield more errors (6 to 8) in classifications.

As a further analysis, we consider the networks constructed from annual publications of TAC, TSP, and TWC. Fig. 5 shows the two dimensional Euclidean embeddings of the networks with respect to the summation of the lower bounds b^0, b^1, and b^2. We expect more variations in annual networks because the time for averaging behavior is reduced. Besides, it is hard to argue that intrinsic differences exist in the collaboration patterns in automatic control, signal processing, and wireless communication communities. Still, networks constructed from the same journal but different annuals tend to be close to each other and form clustering structures. An unsupervised classification with one linear boundary in the embeddings run across the summation of lower bounds would generate 4 (20%) errors out of 20 networks in all three classification problems considered. The less obvious clustering structure formed by networks from different journals in Fig. 5 (c) compared to (a) and (b) also suggests that the collaboration patterns in research communities of signal processing and wireless communication are more similar compared to that of automatic control.

VI. Conclusion

We establish connections between high order networks and simplicial complexes and use the differences between the induced homological features to evaluate the differences between networks. We justify that this is a lower bound to a valid metrics in the space of high order networks modulo permutation isomorphisms. These lower bounds succeed in distinguishing the collaboration patterns of engineering communities from mathematics, and in discriminating engineering communities with different research interests.

References

[1] W. Huang and A. Ribeiro, "Persistent homology lower bounds on high order network distances," *IEEE Trans. Signal Process.*, vol. (submitted), 2016.

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

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

[4] H. Chintakunta and H. Krim, "Divide and Conquer: Localizing Coverage Holes in Sensor Networks," in *Conf. on Sensor, Mesh and Ad Hoc Communications and Networks (SECON)*, 2010, pp. 1–8.

[5] ——, "Distributed localization of coverage holes using topological persistence," *IEEE Trans. Signal Process.*, vol. 62, no. 10, pp. 2531–2541, 2014.

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

[7] S. Choobdar, P. Ribeiro, S. Bugla, and F. Silva, "Comparison of co-authorship networks across scientific fields using motifs," in *Advances in Social Networks Analysis and Mining (ASONAM)*, 2012, pp. 147–152.

[8] G. Carlsson, F. Memoli, A. Ribeiro, and S. Segarra, "Axiomatic construction of hierarchical clustering in asymmetric networks," 2014. [Online]. Available: http://arxiv.org/abs/1301.7724

[9] W. Huang and A. Ribeiro, "Metrics in the space of high order networks," *IEEE Trans. Signal Process.*, vol. 64, no. 3, pp. 615–629, February 2016.

[10] ——, "Persistent Homology Approximations to Network Distances," in *Proc. Global Conf. Signal Info. Process.*, Orlando FA, 2015, pp. 1002–1006.

[11] D. Burago, Y. Burago, and S. Ivanov, *A Course In Metric Geometry*. American Mathematical Society Providence, 2001, vol. 33.

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

[13] "Engineering Village." [Online]. Available: http://www.engineeringvillage.com/search/quick.url