

PERSISTENT HOMOLOGY LOWER BOUNDS ON NETWORK DISTANCES

WeiYu Huang Alejandro Ribeiro

Department of Electrical and Systems Engineering, University of Pennsylvania

ABSTRACT

High order networks are weighted complete hypergraphs collecting relationships between elements of tuples. Valid metric distances between high order networks have been defined but they are difficult to compute when the number of nodes is large. We relate high order networks to the filtrations of simplicial complexes and show 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 comparing the coauthorship networks of engineering and mathematics academic journals. The lower bounds succeed in discriminating engineering communities from mathematics and in differentiating engineering communities with different research interests.

Index Terms— High order networks, network distances, persistent homology, computational topology, pattern recognitions.

1. INTRODUCTION

We consider high order networks that describe relationships between elements of tuples and address the problem of computing an approximate value for their distance. While networks are more often defined as structures describing relationships between pairs of nodes [1, 2], there are many applications in which higher order relationships between triplets, quadruplets, and generic n -tuples are also important. For example, in coauthorship networks where relationships detail the number of joint publications by groups of scholars, apart from collaboration between pairs of authors, there is extra information to be gleaned from collaborations between triplets of authors, or even single author publications. The importance of expressing tuple relationships between groups of nodes in addition to pairs has been utilized in multiple domains [3–9].

The problem of defining distances between networks is important even in the case of pairwise networks. However, while distances are not difficult to define, they are difficult to compute. This has motivated the usage of feature comparisons in which the difference between specific properties of the network is used as a tractable alternative [10–16]. While tractable, 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 – like two networks being close to a third but far from each other – because the triangle inequality is not necessarily valid. A proper distance between pairwise [17] and high order networks [18] overcomes these drawbacks.

The metric distances between high order networks defined in [18] have been applied to compare networks with small number of nodes and have succeeded in identifying collaboration patterns of coauthorship networks. However, because they have to consider all possible node correspondences (Definition 1), network distances cannot be evaluated when the number of nodes is large. This has motivated the derivation of distance approximations relying on homological features [19]. In this paper we address the stronger problem of deriving computationally tractable lower bounds on network distances through the use of persistent homology. It has to be acknowledged that the computation of distance lower bounds suffers from some of the same problems associated with feature

comparison. Still, distance lower bounds do have important properties such as the facts that a large lower bound entails a large distance and that lower bounds can be used to estimate distance intervals because upper bounds are easy to determine using specific correspondences (see Definition 2). The paper begins with a review of important concepts in the definition of distances between high order networks (Section 2) and proceeds to the derivation of lower bounds (Section 3). The lower bounds are used in the classification of coauthorship communities (Section 4).

2. HIGH ORDER NETWORKS

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

$$N_X^K = \left(X, r_X^0, r_X^1, \dots, r_X^K \right). \quad (1)$$

For point collections $x_{0:k} := (x_0, \dots, x_k) \in X^{k+1}$, 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 assume that relationships are normalized so that $0 \leq r_X^k(x_{0:k}) \leq 1$ and 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 $x_{[0:k]} = ([x_0], \dots, [x_k])$ 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 $\pi : X \rightarrow Y$ 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 $r_Y^k(\pi(x_{0:k})) := r_Y^k(\pi(x_0), \dots, \pi(x_k))$. 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 $\mathcal{N}^K \text{ mod } \cong_k$. For each integer $0 \leq k \leq K$, the space $\mathcal{N}^K \text{ mod } \cong_k$ can be endowed with a pseudometric [18]. The definition of this distance requires introducing the prerequisite notion of correspondence [20].

Definition 1 A correspondence between two sets X and Y is a subset $C \subseteq X \times Y$ such that $\forall x \in X$, there exists $y \in Y$ such that $(x, y) \in C$ and $\forall y \in Y$ 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 the 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 defined as

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

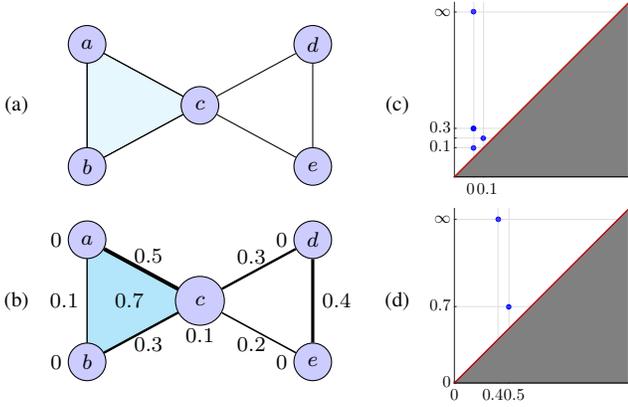


Fig. 1. (a): A bow tie as an example of a simplicial complex which consists 5 0-simplices, 6 1-simplices and 1 2-simplex. (b): A weighted high order network can be represented equivalently as a simplicial complex with weights; weights represent the times the simplices appear in the simplicial complex. (c): the zeroth-dimensional and (d): the first-dimensional persistence diagrams of the filtration induced by (b).

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, $d_N^k(N_X^K, N_Y^K)$ is well-defined when the node cardinalities $|X|$ and $|Y|$ are different. While different order functions r_X^k and r_X^l of a network 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 consideration 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 therefore results in a higher value in the relationship function. In proximity networks the function $r_X^k(x_{0:k})$ describes a level of 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. These restrictions make up the formal definition that we introduce next.

Definition 3 We say that the K -order network $D_X^K = (X, r_X^0, \dots, r_X^K)$ 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

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

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 the set 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 valid metric in the space $\mathcal{D}^K \text{ mod } \cong_k$ and $d_{\mathcal{P}}^k$ a metric in $\mathcal{P}^K \text{ mod } \cong_k$ [18]. Proximity and dissimilarity networks are related entities: $P_X^K = (X, p_X^0, \dots, p_X^K)$ and $D_X^K = (X, d_X^0, \dots, d_X^K)$ 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 distance between two networks coincides to the distance between

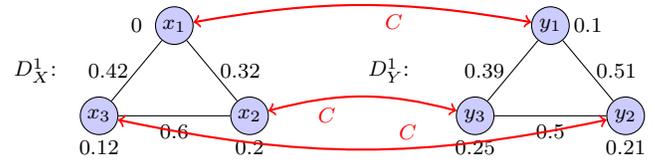


Fig. 2. An example where the bottleneck distance between the k -th dimensional persistence diagrams of the filtrations $\mathcal{L}(D_X^K)$ and $\mathcal{L}(D_Y^K)$ is the same as their k -order network distance for k being 0 and 1. The optimal correspondence C yields network distances of 0.1, the same as the bottleneck distances between the respective persistence diagrams.

their respective duals, 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 computationally tractable lower bounds. These bounds will be obtained by relating dissimilarity networks to filtrations in homology [21, 22] as we discuss next.

3. PERSISTENCE BOUNDS ON NETWORK DISTANCES

We start by establishing a relationship between simplicial complexes with dissimilarity networks. Simplicial complex is a finite collection of simplices, which is defined as the convex hull of the set of non-repeating points, see Figure 1 (a) for a bow tie as an example. 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 $\emptyset = L_{\alpha_0} \subseteq \dots \subseteq L_{\alpha_m} = L$ with $0 = \alpha_0 < \dots < \alpha_m$ an ordered sequence of real numbers is defined as a valid filtration \mathcal{L} . From Definition 3, 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 $\mathcal{L}(D_X^K)$. We formally state this as next.

Proposition 1 The filtration $\mathcal{L}(D_X^K)$ induced from a given dissimilarity network D_X^K is a well-defined filtration.

Proof: See [23] for proofs in the paper. ■

Proposition 1 also implies that $\mathcal{L}(D_X^K)$ is a well-defined simplicial complex. Figure 1 (b) presents a simplicial complex with weights, in which the numbers adjacent to simplices denote the time when simplices appears in the simplicial complex. L_0 is consisted of all vertices except $[c]$ and $L_{0.15}$ is consisted of all vertices union one 1-simplex $[a, b]$.

Next we illustrate the connection between relationship functions in high order networks with persistence diagrams of the filtrations of simplicial complexes. Homological features are open cycles with no enclosing simplices in the simplicial complexes. For example, $[c, d] + [d, e] + [e, c]$ forms an open cycle and represents a first-dimensional homological feature in Figure 1 (b). Consider the homological feature that exists in simplicial complex L_α for any α satisfying $\alpha_b \leq \alpha \leq \alpha_d$. This feature starts to appear in the simplicial complex from time α_b since a new cycle is formed at that time. The feature diminishes at time α_d as a result that the cycle is trivialized by a boundary. This formation and elimination of homological features in simplicial complexes in a filtration is defined as persistent homology. The interval $[\alpha_b, \alpha_d]$ is named persistence interval for the corresponding homological feature and can also be represented as

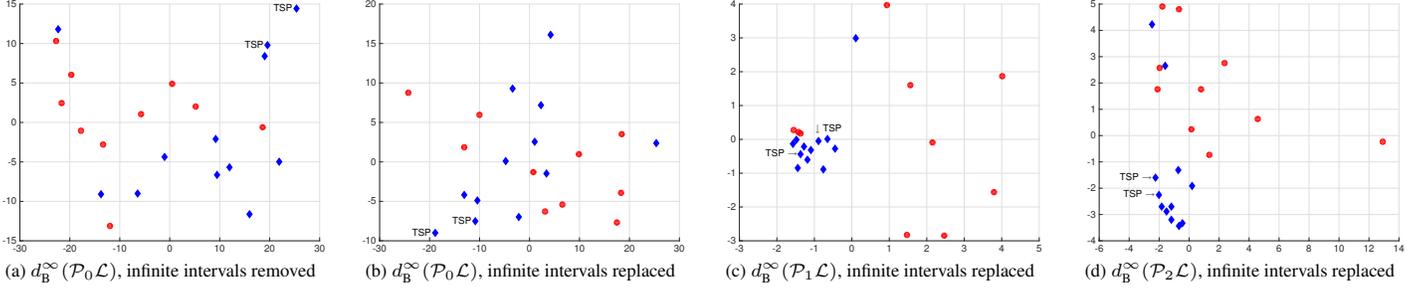


Fig. 3. Two dimensional Euclidean embeddings of the networks constructed from quinquennial publications in engineering and mathematics journals with respect to the network metric lower bounds $d_B^\infty(\mathcal{P}_0\mathcal{L})$ with infinite intervals removed or replaced by finite intervals, $d_B^\infty(\mathcal{P}_1\mathcal{L})$, and $d_B^\infty(\mathcal{P}_2\mathcal{L})$ with infinite intervals replaced. In the embeddings, red circles denote networks constructed from mathematics journals and blue diamonds represent networks from engineering journals. Networks constructed from publications of TSP are labeled.

a point (α_b, α_d) in a two-dimensional plane. The collection of all such points for the k -th dimensional homological features is called the k -th dimensional persistence diagram and denoted as $\mathcal{P}_k\mathcal{L}$.

For the example in Figure 1 (b), at time 0, L_0 consists of all 0-simplices except $[c]$. Every 0-simplex is a 0-cycle and hence there exist 4 zeroth-dimensional homological features. At time 0.1, the appearance of the 1-simplex $[a, b]$ makes the 0-cycles $[a]$ and $[b]$ dependent and one zeroth-dimensional homological feature dies, generating a persistence interval $[0, 0.1)$ and a point $(0, 0.1)$ in $\mathcal{P}_0\mathcal{L}$. Simultaneously, a new zeroth-dimensional feature represented by $[c]$ 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 $[a, c] + [c, b] + [b, a]$ appears at time 0.5 and is eliminated at time 0.7 by the appearance of the triangle $[a, b, c]$. 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 1 (c) and (d) plot the zeroth- and the first-dimensional persistence diagrams of the filtration induced from (b).

It has been shown that the persistent homology of filtrations on a simplicial complex is stable in the sense that similar filtrations yield similar persistence diagrams [24]. In future definitions, it would be important to consider the difference between persistence diagrams. We formalize this by introducing the notion of bottleneck distance as next.

Definition 4 The bottleneck distance $d_B^\infty(Q, \tilde{Q})$ between two point sets Q and \tilde{Q} with same cardinality in two dimensional space is defined as

$$d_B^\infty(Q, \tilde{Q}) = \min_{\gamma} \max_{q \in Q} \|q - \gamma(q)\|_\infty, \quad (4)$$

where γ ranges over all bijections from Q to \tilde{Q} .

For a given bijection γ between Q and \tilde{Q} , the infinity norm $\|q - \gamma(q)\|_\infty$ between a pair of points $q \in Q$ and $\gamma(q) \in \tilde{Q}$ is assigned as the larger one from the x -axis or the y -axis coordinate difference between points q and $\gamma(q)$. The maximum over infinity norm $\max_{q \in Q} \|q - \gamma(q)\|_\infty$ selects the maximum infinity norm $\|q - \gamma(q)\|_\infty$ among all pairs in the bijection. The bottleneck distance in (4) is defined by selecting the bijection that minimizes these maximal infinity norms.

Remark 1 When the number of points in $\mathcal{P}_k\mathcal{L}$ and $\mathcal{P}_k\tilde{\mathcal{L}}$ is different, we may think that the filtrations have a number of homological features being trivialized at the same time they appear, resulting in a number of points on the diagonal in the persistence diagrams. To evaluate the difference between the persistence diagrams, we can add a number of diagonal points to the persistence diagram that has fewer nodes. The points in the respective diagrams then form two sets $\{q_i\}$ and $\{\tilde{q}_i\}$ of same number of points in \mathbb{R}^2 . The bottleneck distance between $\mathcal{P}_k\mathcal{L}$ and $\mathcal{P}_k\tilde{\mathcal{L}}$ is then the

optimal result of the Linear Bottleneck Assignment Problem (LBAP): $\min_{\pi} \max_i c(q_i, \tilde{q}_{\pi(i)})$, where π ranges over all permutations and the cost $c(q, \tilde{q}) = \min \{\|q - \tilde{q}\|_\infty, 0.5 \times \max \{|q_x - \tilde{q}_x|, |q_y - \tilde{q}_y|\}\}$. In words, $c(q, \tilde{q})$ chooses the smaller one from the cost of matching q with \tilde{q} and the cost of matching both points with their respective closest point on the diagonal. Notice that $c(q, \tilde{q}) = c(q, \tilde{q}')$ for any q whenever \tilde{q}, \tilde{q}' are on the diagonal. This makes the choice of the locations of the diagonal points added in the preceding section unsubstantial. The LBAP algorithm can be solved efficiently [25, Algorithm 6.1].

The bottleneck distance between the persistence diagrams of the filtrations induced by two dissimilarity networks is a lower bound of their dissimilarity network distances as we formally state the next.

Theorem 1 Given two dissimilarity networks D_X^K and D_Y^K and an integer $0 \leq k \leq K$, 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 $d_B^\infty(D_X^K, D_Y^K)$ for any $0 \leq k' \leq k$, i.e.

$$d_B^\infty(\mathcal{P}_{k'}\mathcal{L}(D_X^K), \mathcal{P}_{k'}\mathcal{L}(D_Y^K)) \leq d_B^\infty(D_X^K, D_Y^K). \quad (5)$$

Theorem 1 provides theoretical justification that high order networks can be compared via the persistence diagrams of their induced filtrations. We emphasize that the lower bounds are tight. First, it follows from the definition of dissimilarity networks that all the k -order dissimilarities between non-repeating tuples of dissimilarity network D_X^K can be found in homological features, as we formally state in the following proposition.

Proposition 2 Given a dissimilarity network D_X^K , any of its k -order dissimilarities between non-repeating tuples 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.

Proposition 2 guarantees that nothing about dissimilarities between tuples with non-repeating elements is lost when we consider the persistent homologies of the induced filtrations. Moreover, there exist dissimilarity networks D_X^K and D_Y^K such that 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 the same as their ∞ -norm network distance for all orders $0 \leq k \leq K$. See Figure 2 for an illustration where $d_B^\infty(\mathcal{P}_k\mathcal{L}(D_X^1), \mathcal{P}_k\mathcal{L}(D_Y^1)) = d_B^\infty(D_X^1, D_Y^1) = 0.1$ for $k \in \{0, 1\}$.

Remark 2 In [26], persistence diagrams are used to compare point cloud data in three dimensional Euclidean space. It is shown that the bottleneck distance between the persistence diagrams of the Rips filtrations [27] constructed from two metric spaces is a lower bound to the Gromov-Hausdorff (GH) distance between the metric spaces. Theorem 1 can be considered as a generalization in which the filtrations induced from high order networks generalize the Rips filtrations constructed from the metric space. Similar properties of persistent homologies are established in a more universal setting.

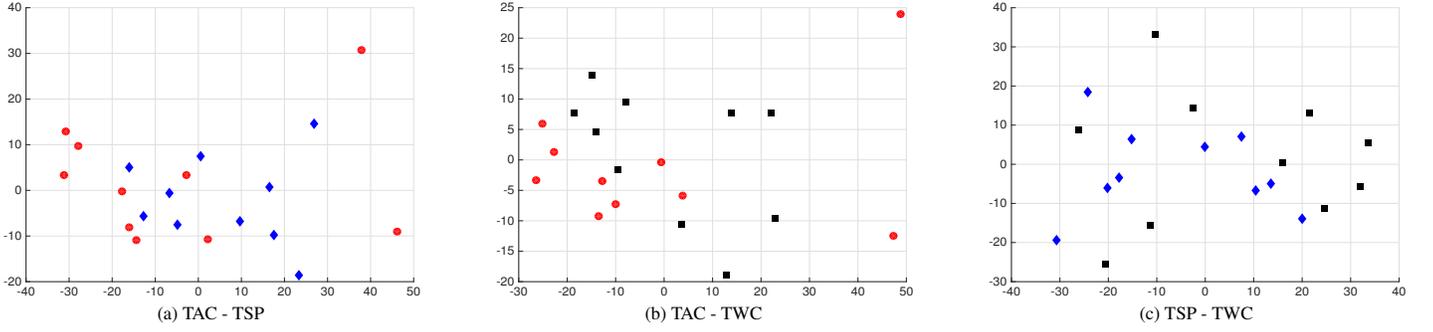


Fig. 4. Two dimensional Euclidean embeddings of the networks constructed from annual publications in TAC, TSP, and TWC with respect to the summation of the metric lower bounds $d_B^\infty(\mathcal{P}_0\mathcal{L})$, $d_B^\infty(\mathcal{P}_1\mathcal{L})$, and $d_B^\infty(\mathcal{P}_2\mathcal{L})$ with infinite intervals replaced by finite intervals. In the embeddings, red circles represent TAC, blue diamonds TSP, and black squares TWC.

The relationship between the dissimilarity network distance and the bottleneck distance can be generalized to proximity networks via the duality established. We state this formally in the following corollary.

Corollary 1 Consider proximity networks P_X^K and P_Y^K and their respective dual networks D_X^K and D_Y^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 $d_P^k(P_X^K, P_Y^K)$ for any $0 \leq k' \leq k$, i.e.

$$d_B^\infty(\mathcal{P}_{k'}\mathcal{L}(D_X^K), \mathcal{P}_{k'}\mathcal{L}(D_Y^K)) \leq d_P^k(P_X^K, P_Y^K). \quad (6)$$

4. COMPARISON OF COAUTHORSHIP NETWORKS

We apply the lower bounds established in Corollary 1 to compare second order coauthorship networks where relationship functions denote the number of publications of single authors, pairs of authors, and triplets. These coauthorship networks are proximity networks because they satisfy the order decreasing property. We consider publications in 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, Trans. Automatic Control (TAC), Trans. Pattern Analysis and Machine Intelligence, 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 individual year from 2004 to 2013. Lists of publications are queried from [28].

For each of these journals we consider all publications in the period of interest and construct proximity networks where the node space 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 comparable we normalize all relationships such that the maximum proximity is 1 in any networks. There are papers with more than three coauthors but we don't record proximities of order higher than 2. The persistent homologies of the induced filtrations of networks are computed using JavaPlex [29]. There exist persistence intervals that never diminish and these would dominate the bottleneck distance between persistence diagrams as in Definition 4. To solve this problem, we can either remove these persistence intervals or replace these intervals, e.g. $[\alpha_b, \infty)$, by persistence intervals $[\alpha_b, 1)$ which are born at the same time but disappear at time 1 – the highest possible finite value. We note that the bottleneck distance between the persistence diagrams with infinite intervals either removed or replaced is still lower bounds of the network distances. By assuming that networks from the

same community or constructed from the same journal have similar collaboration patterns, we show here that metric lower bounds succeed in identifying these patterns and in distinguishing networks from communities with different interests.

We start to use the distance lower bounds to distinguish collaboration patterns in engineering and mathematics communities. Figure 3 shows the two dimensional Euclidean embeddings [30] of the network metric lower bound $d_B^\infty(\mathcal{P}_0\mathcal{L})$ with infinite persistence intervals removed or replaced by finite counterparts and of the lower bounds $d_B^\infty(\mathcal{P}_1\mathcal{L})$ and $d_B^\infty(\mathcal{P}_2\mathcal{L})$ with infinite intervals replaced. The 12 engineering networks (blue diamonds) separate clearly from the 10 mathematics networks (red circles) in all network metric lower bounds considered. An unsupervised classification with one linear boundary run across all four lower bounds would generate errors of 2 (9.09%) to 5 (22.73%) out of 22 networks. Other lower bounds not presented yield similar results. We emphasize that networks constructed from the same journal with different quinquennia tend to be close to each other with respect to the metric lower bounds. As an illustration, the networks constructed from TSP with different quinquennia are marked and it is clear that their differences in persistent homologies are considerably low. Such scenarios are observed for several other journals as well.

Next we consider the networks constructed from annual publications of TAC, TSP, and TWC. Figure 4 shows the two dimensional Euclidean embeddings of the networks with respect to the summation of the network metric lower bounds $d_B^\infty(\mathcal{P}_0\mathcal{L})$, $d_B^\infty(\mathcal{P}_1\mathcal{L})$, and $d_B^\infty(\mathcal{P}_2\mathcal{L})$ with infinite intervals replaced by finite counterparts. 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 these research 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 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 Figure 4 (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.

5. CONCLUSION

We related high order networks to simplicial complexes and used the differences between the induced homological features to evaluate the differences between networks. We justified that this is a lower bound to a valid metric in the space of high order networks modulo permutation isomorphisms. These lower bounds succeeded in distinguishing the collaboration patterns of engineering communities from mathematics and in discriminating engineering communities with different interests.

6. 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] 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.
- [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] 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.
- [7] 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.
- [8] 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.
- [9] 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.
- [10] 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.
- [11] 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.
- [12] 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.
- [13] 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.
- [14] 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.
- [15] N. Pržulj, “Biological network comparison using graphlet degree distribution,” *Bioinformatics*, vol. 23, no. 2, pp. e177–183, Jan. 2007.
- [16] T. Milenković and N. Pržulj, “Uncovering biological network function via graphlet degree signatures,” *Cancer informatics*, vol. 6, p. 257, Jan. 2008.
- [17] 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>
- [18] W. Huang and A. Ribeiro, “Metrics in the Space of High Order Networks,” *Signal Processing, IEEE Transactions on*, vol. (to appear), 2015. [Online]. Available: <http://arxiv.org/abs/1507.01994>
- [19] —, “Persistent Homology Approximations of Network Distances,” in *Proc. Global Conf. Signal Info. Process.*, vol. (to appear), Orlando FA, Dec 14–16 2015.
- [20] D. Burago, Y. Burago, and S. Ivanov, *A course in metric geometry*. American Mathematical Society Providence, 2001, vol. 33.
- [21] 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.
- [22] A. Zomorodian and G. Carlsson, “Computing persistent homology,” *Discrete Comput. Geo.*, vol. 33, pp. 249–274, 2005.
- [23] W. Huang and A. Ribeiro, “Persistent homology lower bounds on high order network distances,” *Signal Processing, IEEE Transactions on*, vol. (submitted), 2015. [Online]. Available: <http://arxiv.org/abs/1507.03044>
- [24] F. Chazal, D. Cohen-Steiner, M. Glisse, L. J. Guibas, and S. Y. Oudot, “Proximity of Persistence Modules and Their Diagrams,” in *Proceedings of the Twenty-fifth Annual Symposium on Computational Geometry*, 2009, pp. 237–246.
- [25] R. Burkard, M. Dell’Amico, and S. Martello, *Assignment problems*. Society for Industrial and Applied Mathematics, 2009.
- [26] F. Chazal, D. Cohen-Steiner, L. J. Guibas, F. Mémoli, and S. Y. Oudot, “GromovHausdorff Stable Signatures for Shapes using Persistence,” *Eurographics Symposium on Geometry Processing*, vol. 28, no. 5, pp. 1393–1403, 2009.
- [27] E. Carlsson, G. Carlsson, and V. D. Silva, “An algebraic topological method for feature identification,” *Intl. J. Computational Geometry and Applications*, vol. 16, pp. 291–314, 2003.
- [28] “Engineering Village: the place to find answers to engineering questions.” [Online]. Available: <http://www.engineeringvillage.com/search/quick.url>
- [29] A. Tausz, M. Vejdemo-Johansson, and H. Adams, “JavaPlex: A research software package for persistent (co)homology,” in *Proceedings of ICMS 2014*, 2014, pp. 129–136.
- [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.