AFRL-RI-RS-TR-2020-012

CLIQUE AND INDEPENDENCE COMPLEX STRUCTURES FOR SENSOR NETWORK ANALYSIS

UNIVERSITY OF DELAWARE

FEBRUARY 2020

FINAL TECHNICAL REPORT

APPROVED FOR PUBLIC RELEASE; DISTRIBUTION UNLIMITED

STINFO COPY

AIR FORCE RESEARCH LABORATORY
INFORMATION DIRECTORATE

■ **AIR FORCE MATERIEL COMMAND** ■ **UNITED STATES AIR FORCE** ■ **ROME, NY 13441**

NOTICE AND SIGNATURE PAGE

Using Government drawings, specifications, or other data included in this document for any purpose other than Government procurement does not in any way obligate the U.S. Government. The fact that the Government formulated or supplied the drawings, specifications, or other data does not license the holder or any other person or corporation; or convey any rights or permission to manufacture, use, or sell any patented invention that may relate to them.

This report is the result of contracted fundamental research deemed exempt from public affairs security and policy review in accordance with SAF/AQR memorandum dated 10 Dec 08 and AFRL/CA policy clarification memorandum dated 16 Jan 09. This report is available to the general public, including foreign nations. Copies may be obtained from the Defense Technical Information Center (DTIC) (http://www.dtic.mil).

AFRL-RI-RS-TR-2020-012 HAS BEEN REVIEWED AND IS APPROVED FOR PUBLICATION IN ACCORDANCE WITH ASSIGNED DISTRIBUTION STATEMENT.

FOR THE CHIEF ENGINEER:

/ S /
JOHN KELLY
Work Unit Manager

/ S /
JAMES S. PERRETTA
Deputy Chief, Information Exploitation
and Operations Division
Information Directorate

This report is published in the interest of scientific and technical information exchange, and its publication does not constitute the Government's approval or disapproval of its ideas or findings.

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. REPORT DATE *(DD-MM-YYYY)*	2. REPORT TYPE	3. DATES COVERED *(From - To)*
FEBRUARY 2020	FINAL TECHNICAL REPORT	JUN 2019 – SEP 2019

4. TITLE AND SUBTITLE

CLIQUE AND INDEPENDENCE COMPLEX STRUCTURES FOR SENSOR NETWORK ANALYSIS

5a. CONTRACT NUMBER
N/A

5b. GRANT NUMBER
FA8750-18-1-0078

5c. PROGRAM ELEMENT NUMBER
62788F

6. AUTHOR(S)

Chad Giusti

5d. PROJECT NUMBER
G2TP

5e. TASK NUMBER
18

5f. WORK UNIT NUMBER
DE

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
University of Delaware
222 S. College Ave
Newark DE 19716

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)

Air Force Research Laboratory/RIGC
525 Brooks Road
Rome NY 13441-4505

10. SPONSOR/MONITOR'S ACRONYM(S)
AFRL/RI

11. SPONSOR/MONITOR'S REPORT NUMBER
AFRL-RI-RS-TR-2020-012

12. DISTRIBUTION AVAILABILITY STATEMENT
Approved for Public Release; Distribution Unlimited. This report is the result of contracted fundamental research deemed exempt from public affairs security and policy review in accordance with SAF/AQR memorandum dated 10 Dec 08 and AFRL/CA policy clarification memorandum dated 16 Jan 09

13. SUPPLEMENTARY NOTES

14. ABSTRACT

The goal of this project was the development of topological methods for use in the study of sensor networks. There were two primary tasks; investigation of whether classically known topological phenomena, which allow for locally consistent information to lead to global inconsistency can be leveraged to attack a sensor network/ and, investigation of whether persistent homology of geometric and functional network architectures is stable under subsampling. As a result of this work, the PI proposes a novel method for measuring dissimilarity and consistency of measurements in sensor networks.

15. SUBJECT TERMS
Automated network classification/characterization; geolocation refinement; high dimensional SIGINT data analysis; signal analysis; signal geolocation; signal network; signal targeting; Tactical SIGINT

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			**JOHN KELLY**
U	U	U	UU	16	19b. TELEPHONE NUMBER (*Include area code*) **N/A**

TABLE OF CONTENTS

LIST OF FIGURES

1.0 SUMMARY

The goal of this project was to apply ideas and tools from the field of topological data analysis to understand the structure of sensor networks. First, the PI investigated whether topological structure in a sensor network could be exploited to attack the network by intentionally introducing false information in a manner that would be undetectable to nearby sensors. Such attacks were determined to be infeasible in realistic sensor networks, as success relies on sparse, ring-like network structure to succeed. In addition, the PI began investigating how the persistent homology, a measure of mesoscale topological structure in sensor networks, is affected by subsampling the weights on edges of the network. Computational experiments demonstrated that complicated and inconsistent effects, but analysis of these computations was hindered by a more subtle issue: the topological signatures are highly sensitive to choice of dissimilarity measure used to construct the network from the same underlying data. The PI therefore recommends a systematic investigation of these issues, and proposes a sheaf-theoretic measure of the consistency and dissimilarity of sensor response to stimuli that may be more amenable to rigorous topological analysis.

2.0 INTRODUCTION

A wide range of scientific and engineering problems focus on complex systems of interacting agents with the ability to sense and interact with their environment locally, and to communicate a limited amount of information to one another through channels that rely on the structure of the environment. Such systems can be thought of as sensor networks embedded in the environment, with agents acting as sensors, and the network structure modeling the communication channels between them. Assuming the structure of the system is static on the time scale of interest, these various interacting components can be investigated using techniques from algebraic topology, a field of mathematics that describes qualitative mesoscale features of systems based on the interactions of local elements.

This project was initiated with two primary aims: (1) studying how the topological structure of a network could be used to produce locally consistent but globally inconsistent representations, that is, how to attack the network in a fashion that is not locally detectable; and, (2) understanding how to subsample (edge-wise) networks without substantially altering their persistent homology.

The first aim was tractable to direct analysis, however the PI determined that such attacks are impossible in the context of real sensor networks: the necessary conditions for success require the appearance of a unique ring-like structure to the network that is common in model networks, but unlikely to occur in any real setting.

Theoretical and experimental difficulties encountered during work toward the second aim rapidly led the PI to conclude that there is substantial fundamental work needed in understanding how the selection of a dissimilarity measure used to construct a network affects the topological structure of the network. In this context, the PI formulated a sheaf-theoretic measure of compatibility between sensor responses to stimuli in a network, and proposes the use of a convex optimization problem to construct such a measure in the context of data recorded from sensors. Future work testing this method with synthetic and simulated data would lay suitable groundwork for understanding feasibility and potential applications, and should these be

promising, the PI posits that the framework will also be amenable to rigorous incorporation with existing topological data analysis methods.

3.0 METHODS, ASSUMPTIONS, AND PROCEDURES

3.1 Model Networks

We model a physical or logical network as a weighted simple graph, $G = (V, E, w)$, where V is a set of *vertices*, $E \subseteq \binom{V}{2}$, a set of *edges*, and $w: E \to [0, \infty]$ a set of *weights*. Vertices correspond to the elements of the system, such as sensors or agents. An edge between vertices models some notion of similarity, such as proximity in the environment or similarity of behavior, and the weight assigned to each edge provides a numerical representation of the *dissimilarity* of its endpoints, in this context a distance such as Euclidean distance between points or correlation distance between time series. Assigning an edge $w(e) = \infty$ can be considered equivalent to removing the edge from the graph.

In order to investigate the structure of sensor networks, we consider a variety of models for weighted networks that capture various facets of the structure of sensors in an environment or responding to data, as well as providing null models against which to compare real data.

Modeling networks from a structure -centric perspective, we consider weighted graphs for which similarity is controlled by the environment, in the form of geometric random graphs. Explicitly, we select a subspace $A \subseteq M$ of a metric space M, along with some probability distribution on A, and sample from the distribution to obtain a finite collection of points $V \subseteq A$. We construct a weighted graph from this collection of points by taking the complete graph on the vertices V, along with weights $w(\{v_1, v_2\}) = d(v_1, v_2)$. It may be useful to compose the distance with some monotone increasing function, modeling a nonlinear increase in dissimilarity as the distance between sensors increases. However, the topological methods we apply are intrinsically insensitive to such confounds (though that data can be reintroduced for finer measurements of structure) and so we omit it from our models. For the experiments described in this project, we used the standard Euclidean metric and uniform distributions on compact subspaces for all experiments.

Alternately, it is often useful to study sensor networks from the perspective of their function, that is, their response to presented stimuli. In this case, we assume that some stimulus has been presented to the sensors and assign to each sensor a vector corresponding to the recorded measurement or sequence/time series of measurements that sensor recorded. We then compute a notion of dissimilarity between sensors based on these time series. For the experiments described in this project, simulated time series were generated using the Python Timesynth package and *ad hoc* code. Functional networks were constructed by generating periodic and pseudo-periodic signals, and constant and linear gaussian processes, with small magnitude white noise, and computing dissimilarity using Peason correlation, L_2-distance, and Pearson correlation between Fourier coefficients.

3.2 Topological Measurements of Network Structure

While a broad range of measures of the structure of weighted networks are available, techniques from applied algebraic topology, specifically *persistent homology*, provide a rich and theoretically sound feature set for such systems. The intended goal of this project was the

application of these tools to study sensor networks, and so for completeness we provide a very brief summary of the pertinent ideas. A full treatment of the mathematics is beyond the scope of this document, and the PI refers readers to the broad array of standard introductory materials PI recommends for further details.

Given a set S, denote by $P(S)$ the power set of S. A *weighted simplicial complex* is a generalization of a weighted simple graph, given by a triple $\Sigma = (V, S, w)$, where $V$are the vertices, $S \subseteq P(V)$ is a set of subsets of V called the *simplices* or *faces*, and $w: S \to \mathbb{R}$ are the weights. We require that S is *downward closed*, meaning that whenever $\tau \in S$ and $\sigma \subseteq \tau$ that $\sigma \in S$, and that w is monotone increasing with respect to inclusion, so if $\sigma \subseteq \tau$, then $w(\sigma) \leq w(\tau)$. That is, our measure of dissimilarity of collections of vertices increases as we introduce more vertices. A *subcomplex* $\Sigma' = (V', S', w')$ of a weighted simplicial complex Σ, written $\Sigma' \subseteq \Sigma$, is given by taking $V' \subseteq V,\ S' \subseteq S \cap P(V')$ satisfying the downward closure condition, and $w' = w|_{S'}$. An *unweighted simplicial complex* is a weighted simplicial complex without the weights.

For a weighted network $G = (V, E, \widetilde{w})$, we construct a canonical weighted simplicial complex called its *clique complex,* $X(G) = (V, S, w)$ by taking $S = P(V)$ and $w(\sigma) = \max_{\{a,b\} \subseteq \sigma} \left(\widetilde{w}(\{a, b\})\right)$. That is, we assign to each clique in the complete graph on vertices V a simplex with weight equal to its edge with maximal weight. It is easy to check that this is the minimum weight one can assign to each simplex that satisfies the requirements on w. In the case where vertices are points in a metric space and the dissimilarity measure is a distance function, cliques correspond to families of points which are pairwise within some distance given by the weight on the simplex, and is called the *Vietoris-Rips* complex of the collection of points. Another simplicial complex associated to a weighted graph is the *independence complex*, with simplices supported on the edge-complement of simplices in the graph at each weighting. The weighting on the independence complex follows the opposite monotonicity convention, and the project did not reach a stage where it was explicitly used, so further details are omitted.

Given a weighted simplicial complex Σ, let $W = w(S) \subseteq \mathbb{R}$ and write $W = \{w_1 < w_2 < \cdots < w_n\}$. We obtain a *filtration* of unweighted simplicial complexes $\Sigma_1 \subseteq \Sigma_2 \subseteq \cdots \subseteq \Sigma_n$ by taking $\Sigma_i = (V_i, S_i)$ with $S_i = \{\sigma \in S \mid w(\sigma) \leq w\_i\}$ and V_i the smallest subset of V so that $S_i \subseteq P(V_i)$. A filtration of a dissimilarity weighted simplicial complex can be thought of as a decomposition into subcomplexes which admit increasing amounts of dissimilarity.

The faces of a simplicial complex can be decomposed by the number of vertices included in each: $S = \coprod_{i=0}^{|V|} S_i$, with $S_i = \{\sigma \in S \mid \#\sigma = i + 1\}$. Note that we shift the subscript, called the *degree* of the face, by 1, so S_i consists of faces containing $(i + 1)$ vertices. This shift corresponds to the dimension of the convex hull of $(i + 1)$ points in general position, and, for example, $S_{-1} = \{\emptyset\}$, S_0=V, $S_1 = E$, and so on. Observe also that downward closure ensures that there are *boundary maps* $\partial_i: S_i \to P(S_{i-1})$, taking each simplex $\sigma \in S_i$ to the subset $\{\sigma \setminus \{a\} \mid a \in \sigma\}$, each element of which has as its convex hull one of the geometric faces of the convex hull of the points in σ.

Given an unweighted simplicial complex $\Sigma = (V, S)$, we define a sequence of $\mathbb{F}_2$-vector spaces called chain groups via $C_i(\Sigma) = \mathbb{F}_2\langle e_\sigma \mid \sigma \in S_i\rangle$. That is, the *i*th chain group of Σ is the vector space with basis corresponding to simplices in S_i. Elements of the chain groups are called *chains* and can be thought of as formal sums of simplices of the same degree.

In addition, the boundary maps described above induce linear transformations, called *differentials,* $d_i : C_i(\Sigma) \to C_{i-1}(\Sigma)$ defined on basis vectors $d_i(e_\sigma) = \sum_{\tau \in \partial_i(\sigma)} e_\tau$. Elements of the kernel of d_i are chains with "vanishing boundary" called *cycles*. For d_1, this is simply then notion of circuits in a graph, but in other dimensions one obtains the analog of *closed* geometric objects of the form that might (though not always) enclose lower-dimensional volume. It is elementary to show that $d_i \circ d_{i+1} = 0$, and so $im(d_{i+1}) \subseteq \ker(d_i)$; that is, that one way to construct a cycle is to take the boundary of a higher-dimensional chain, and that such boundaries are always closed.

The *homology* of an unweighted simplicial complex measures the "number of non-bounding cycles". Explicitly, for each $i = 0, 1, \ldots, N$, we define the quotient vector space $H_i(\Sigma) = \frac{\ker(d_i)}{im(d_{i+1})}$. Elements of $H_i(\Sigma)$ are equivalence classes of i-cycles in Σ such that two cycles are equivalent if they differ by the boundary of some collection of $(i+1)$-simplices. We consider homology classes to be *topological features* of the complex. Developing general semantics for homology classes in terms of the data underlying a simplicial complex is one of the fundamental open challenges in applied algebraic topology.

Homology is a measure of structure in unweighted complexes, but our objects of interest are weighted. Decomposing a weighted simplicial complex Σ into a filtration of unweighted complexes $\Sigma_1 \subseteq \Sigma_2 \subseteq \cdots \subseteq \Sigma_N$, we can then apply homology to each to obtain a collection of features at each successive weighting. However, one of the fundamental results in algebraic topology is that homology is a *functor*: given a homomorphism of simplicial complexes $f: \Sigma \to \Sigma'$, there is are *induced* linear transformations $f_i: H_i(\Sigma) \to H_i(\Sigma')$ for each *i*. The inclusions of simplicial complexes in a filtration are examples of such homomorphisms, inducing transformations $\iota_{k-1,k}: H_i(\Sigma_{k-1}) \to H_i(\Sigma_k)$, so we can apply the induced linear transformations to track how cycles evolve as we increase the threshold weight in the complex. This collection of vector spaces and linear maps is called the *persistent homology* of the weighted complex.

We say a homology class $[\sigma] \in H\_i(\Sigma_k)$ is *born* at filtration *k* if $[\sigma] \notin im(\iota_{k-1,k})$ and that it *dies* at filtration k if $[\sigma] \neq [0] \in H_i(\Sigma_k)$ but $[\sigma] \in \ker(\iota_{k,,k+1})$. We write $b_{[\sigma]} = k$ if $[\sigma]$ is born at filtration k, and $d_{[\sigma]} = k$ if $[\sigma]$ dies at filtration k. A foundational theorem in applied topology states that the collection of birth/death pairs for cycles through the filtration is well-defined and characterizes the persistent homology of a weighted complex up to isomorphism. It also provides a powerful summary statistic, called a *persistence diagram*, which is simply a scatter plot of the pairs $(b_{[\sigma]}, d_{[\sigma]})$ characterizing the persistent homology of a the complex of interest, as illustrated in Figure 1.

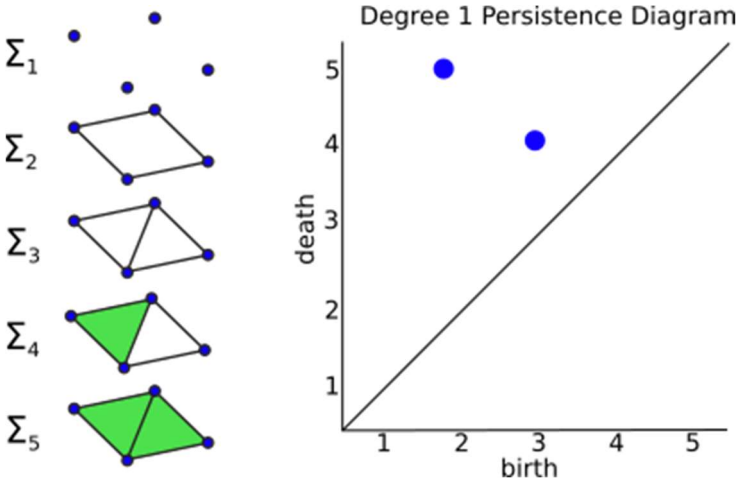

Figure 1: Filtered simplicial complex and its degree 1 persistence diagram

The computation of persistent homology is computationally expensive, but continues to improve by orders of magnitude on an annual basis in terms of both speed and resource use. Persistent homology computations for this project were performed in the Eirene software library, version 0.6,1, written in the Julia language.

4.0 RESULTS AND DISCUSSION

4.1 Infeasibility of Topological Attacks on Real Sensor Networks

One of the intended directions of research for the project was investigating whether subversion of a sensor in a distributed sensor network could be used to introduce false data that appears locally consistent to nearby sensors, thus failing to alert the system, but which is nonetheless globally inconsistent.

The intuition that this might be a possibility comes from the topological theory of vector bundles, which record ways to consistently, continuous assign of vectors in a vector space to all points in a space. If we think of sensors in a sensor network representing data as vectors, then a consistent assignment of information across all sensors can be thought of a discretization of this continuous notion of a vector bundle. In the continuous setting, the cylinder and the Möbius band are examples of 1-dimensional vector bundles over a circle. The cylinder consistently assigns a (constant) vector to all points on the circle, providing globally consistent data. On the other hand, on the Möbius band one can assign data to any local neighborhood via the constant vector, but translating the vector far enough along the band (a full rotation about the circle) results in an inversion of the vector: this assignment of data is necessarily *globally inconsistent*. Such inconsistencies can be generated by a single local inversion: a small neighborhood inside of which the requisite "twist" occurs.

In the discrete context, such as in a sensor network, vector bundles are equivalent to *cellular sheaves* of vector spaces on a simplicial complex built from the underlying graph. Given a finite simplicial complex $\Sigma = (V, S)$, a *cellular presheaf of vector spaces on* Σ is an assignment of a vector space X_σ to each face $\sigma \in S$, along with linear maps $\phi_{\tau,\sigma} : X_\sigma \to X_\tau$ whenever $\sigma \subseteq \tau$, with $\phi_{\sigma,\sigma}$ the identity map on X_σ. If, in addition, whenever $\sigma \subseteq \tau \subseteq \rho$, $\phi_{\rho,\tau} \circ \phi_{\tau,\sigma} = \phi_{\rho,\sigma}$ then the presheaf is a *sheaf*.

A sheaf is a data structure that records an assignment of local information, given by vectors $\mathrm{v_a} \in \mathrm{X_a}$, $\mathrm{a} \in \mathrm{V}$, along with rules for when the local information is globally consistent, which occurs when all possible images of the assigned vectors agree: $\phi_{\sigma,a}(v_a) = \phi_{\sigma,b}(v_b)$ for all $a, b \in \sigma$, for all $\sigma \in S$. See Figure *2*(a) for an illustration of a sheaf along with a globally consistent assignment of data. The top panel is an example of a cellular sheaf of real vector spaces over a the complete graph on three vertices. It consists of an assignment of a vector space to each simplex, not necessarily allt he same, and maps between vector spaces in the direction opposite that of the boundary maps in the complex. The bottom panel gives an example of an assignment of globally consistent data for this sheaf, given by a vector in each vector space so that the image under each linear map is correct.

In the context of a sensor network, a reasonable model simplicial complex is the clique complex of the underlying network. That is, we assume that a collection of sensors $\{a_1, \ldots a_k\}$ that are close enough so all pairs can interact can request information from their neighbors, and thus form a clique σ in the network, can perform a *local consistency check* using the maps ϕ_{σ,a_i}. This is

equivalent to querying all neighboring sensors for their current states, using predetermined rules for sensor state consistency in the form of the linear maps. (See Section 4.3 below for further discussion of these rules.)

With this formalism, the proposed attack would be performed by replacing a sensor b with a false sensor, which may report different data to different neighbors. That is, the sensor may falsify the linear maps $\phi_{\sigma,b}$ for all $\sigma \in S$ with $b \in \sigma$. See Figure 2(b) for an example of a cellular sheaf on a small "1-cycle network" that models the Möbius band. Observe that there is no globally consistent assignment of data for this sheaf. The red transformation models the half-twist in the band, and any attempt to assign consistent data starting in the bottom and working outward left fails when attempting to assign a consistent vecotr in the blue vector space, which is not local to the twist .

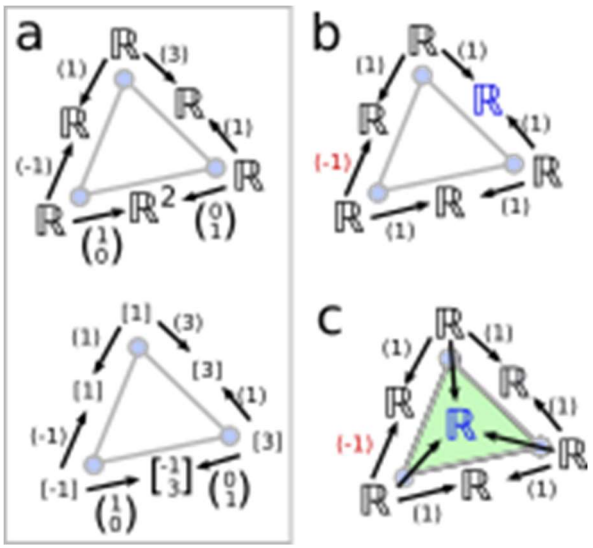

Figure 2: Sheaves provide a data structure for checking sensor consistency.

However, while such an attack might be performed in extremely simple and sparse network architectures, the PI has determined that such attacks rely fundamentally on the such unrealistic assumptions. In a large network in which the attacker might be implicated in multiple local 1-cycles such attacks must be coordinated against data reported by neighbors of neighbors, and thus would be detected unless the attacker is in possession of global data, indicating that they have more information than the network itself. On the other hand, in a dense sensor environment, where the sensors involved in a cycle can compare information through a higher simplex, the global consistency can be directly checked, resulting in a failure of the attack, per Figure 2(c). In this example, consistency issues could be checked locally if the simplicial complex is dense enough to include a 2-simplex, which aggregates data from all of the constitent sensors. However, in this example, the maps given on the edges cannot even be extended to a sheaf on the clique complex; there is no consistent choice of maps which satisfies the sheaf condition. Thus, in most real sensor networks, the proposed method of attack is essentially infeasible.

4.2 Network Persistent Homology

Based on prior work, the PI became interested in determining whether persistent homology of clique complexes for sensor networks are stable under edge subsampling. Such stability is of interest in situations involving incomplete structural or functional measurements of system structure, damaged or unreliable data, or situations where the natural dissimilarity measure is expensive to compute. This investigation was intended to follow two distinct courses: revisiting simple geometric structures to try to understand details of what certain subsampling schemes are preserving, and the experimental application of subsampling schemes to networks without explicit geometric structure, such as networks built from similarity of time series.

4.2.1 Geometric Networks

The PI investigated two structures: points drawn from a uniform distribution in a cube in $\mathbb{R}^d$ for varying values of *d*, and sampled from a thin annulus in the Euclidean plane. The former model produces a geometric random graph, for which the persistent homology is known to provide strong information about dimension, while the latter models a noisy circle, in a sense the simplest non-trivial topological space. Figure 3(a) illustrates the degree 1 persistence diagram of the clique complex of a geometric random graph built from points sampled uniformly from the Euclidean cube in five dimensions, and, in top-to-bottom order, how it deforms under various subsampling schemes, including random subsampling, subsampling by removing only small, moderate, and large distance edges. Figure 3(b) demonstrates the same diagrams for an annulus in the Euclidean plane. In Fgure 3(a), serious deformation occurs when the smallest 30% of distances are removed, while the middle and top 30% are virtually identical to the original. In Figure 3(b), the outlier point in the diagram corresponds to the topological "hole" in the annulus. This feature is preserved except under removal of shortest edges. Unlike the uniform case, there are substantial differences between the original and middle 30% removal case, but the random removal case is qualitatively quite similar to the original. Thus, even in the case of striclty geometric complexes, there are subtle questions about how subsampling affects the persistent homology.

The PI initially attempted to investigate these deformations from a theoretical perspective, by studying small examples via manual computation and attempting to generalize the observed phenomena. While they provided some intuition, these computations rapidly became intractable and the PI was unable to make progress understanding the change in the distribution of cycles from this perspective.

However, based on these investigations, the PI believes that these deformations can be understood through the lens of cycle representatives. Explicitly, each homology class is an equivalence class of closed 1-cycles, with two cycles equivalent when they differ by the boundary of a family of 2-simplices. It is reasonable to study only classes which contain connected representatives, and from these to consider only connected representatives, as these will give "local" information and dramatically reduce the complexity of the analysis.

These equivalence classes are large and have yet to be systematically studied due to their complexity. However, it is feasible that in the case of simple network models, like random graphs and Euclidean random graphs with points drawn from a simple distribution, there are statistical descriptions of how the family of representatives of different cycles evolves through the filtration. In the case of random deletion of edges, the effects of deletion would be inconsistent across classes at a low level but may have discernable patterns in terms of their statistical effect on the complete lifetime. On the other hand, deleting short, moderate, and large distance edges removes "epochs" of their evolution. Unfortunately, there are currently no computational tools available to perform the appropriate analysis in data large enough to provide real evidence: more than one software package will provide a single choice of generator at some fixed filtration level but understanding the complete equivalence class is a computationally expensive task that has not been well-studied.

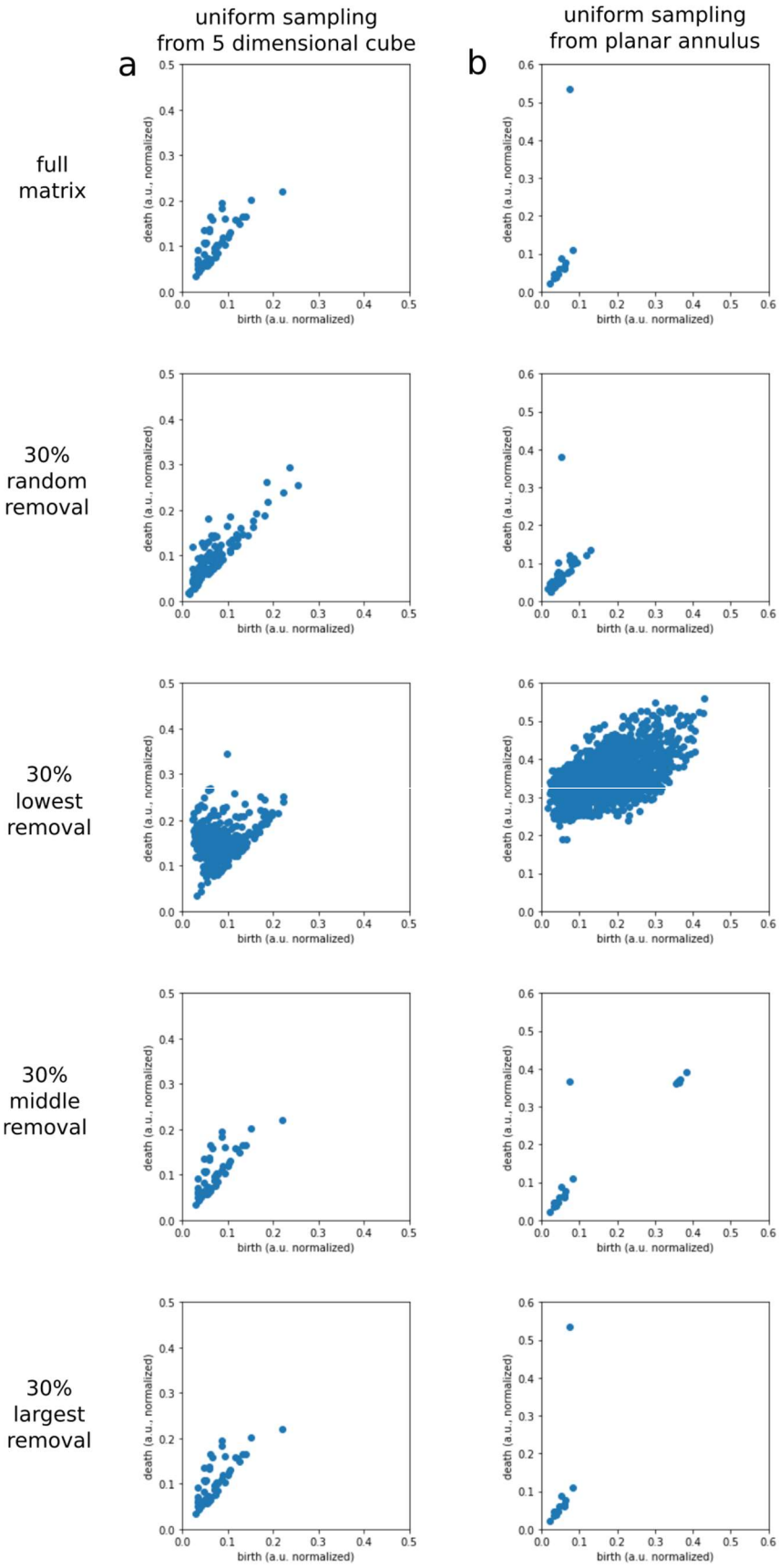

Figure 3: Degree 1 persistence diagrams for subsampled geometric complexes.

4.2.2. Time Series Dissimilarity Networks

A standard method for constructing a network based on observations of time series is to consider the complete weighted network with weights given by pairwise dissimilarity. While the PI initially intended to investigate how subsampling affected these networks, it was necessary to choose a notion of dissimilarity before proceeding. It rapidly became clear that this choice might have substantial impact on the results of the experiments, and as the PI could locate no discussion of this choice in the literature, the PI refocused effort on this problem. Figure 4 gives the persistence diagrams for dissimilarity networks built from a single sample population of pseudo-periodic time series with small magnitude white noise, with dissimilarity given by (a) correlation distance, and (b) Euclidean L2 distance. These panels demonstrate that there is substantial dependence of topological structure on the choice of dissimilarity measure. Therefore, determining which properties of the underlying signals are reflected in the persistence diagram, and thus understanding how subsampling affects the diagrams, first requires some understanding of this dependency.

The PI attempted to derive some comparisons for correlation and L2 distances in simple cases (small networks computed from sinusoids and square waves) by hand, but was unable to derive any explicit information about their persistent homology. Following these inconclusive attempts to develop a theoretical foundation for understanding the effect on persistent homology of the choice of dissimilarity measure for a given data set, the PI has come to believe that one of the principle open questions in applied topology is understanding, for fixed data, the effect of the choice of dissimilarity measure on persistent homology. That is, what is the collection of persistence modules/persistence diagrams that can be obtained from a fixed data set under all choices of dissimilarity measure from some reasonable class? *Ad hoc* analyses may eventually provide specific examples of classes of data on which topological features induced by particular traditional dissimilarity measures can be understood without a general theory, but the PI believes that a more methodical study that begins with a study of the structure of (an appropriately chosen) space of dissimilarity measures will be necessary to gain a real grasp on these problems.

4.3 Cellular Sheaves for Learning Sensor Dissimilarity

In the course of pursuing these questions, the PI invested substantial effort in understanding what "(dis)similarity of sensors in a network" should mean. While a wide range of advanced methods for comparing sensor measurements, in many applications they are understood to be time series represented by vectors in $\mathbb{R}^n$, and

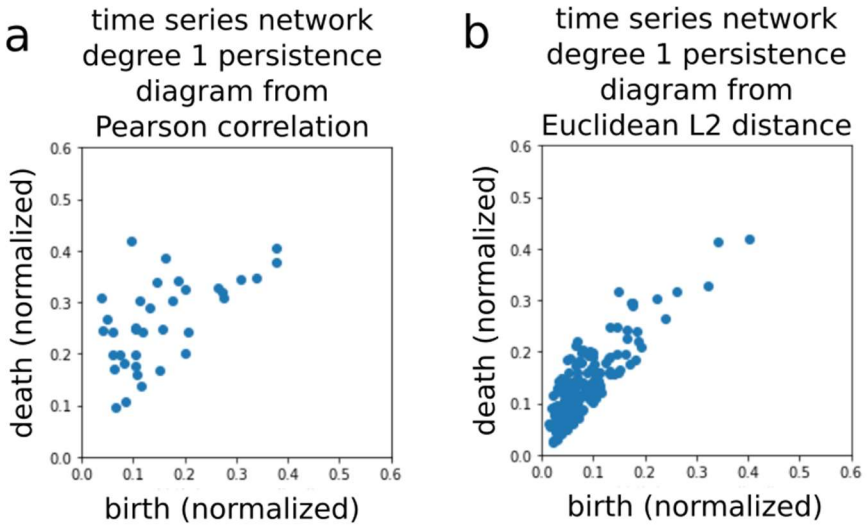

Figure 4: Degree 1 persistence diagrams for two dissimilarity networks constructed from the same synthetic time series data using (a) correlation distance and (b) Euclidean L2 distance.

(dis)similarity is measured using some variant of correlation distance, Euclidean distance, coherence of or Euclidean distance between Fourier/wavelet coefficients, or the comparison of coefficients in auto-regressive models. While each of these has advantages in different domains, they make assumptions about the structure of the recorded sensor readings, and in particular are insensitive to the structure of the underlying network or sensor receptive fields.

To meet this need, the PI proposes a novel (to the PI's knowledge) method for characterizing consistency of sensor readings utilizing the language of cellular sheaves of vector spaces, as described in Section 4.2.

Before describing the construction in general, we build intuition by considering a sensor network consisting of a pair of sensors a and b, which record responses to stimuli y presented to the network as vectors $x_a(y) \in X_a$ and $x_b(y) \in X_b$ respectively.

First, construct a sheaf over the simplicial complex $\Sigma = (\{a, b\}, \{a, b, ab\})$ by fixing a vector space X and taking $X_a = X_b = X_{ab} = X$ with $\phi_{ab,a}$ and $\phi_{ab,b}$ the identity maps. This sheaf represents a sensor network for which the only globally consistent data for the network occurs when $x_a(y) = x_b(y)$. That is, the two sensors record identical responses to each stimulus. If the two sensors disagree, we can use a measurement like correlation distance to obtain a notion of how far from consistent the data is. That is, we could define the (a, b)-in*consistency* of a measurement to be

$$I^0_{ab}\big(x_a(y), x_b(y)\big) = \sqrt{1 - Corr\left(\phi_{ab,a}\big(x_a(y)\big), \phi_{ab,b}\big(x_b(y)\big)\right)} \tag{1}$$

On the other extreme, we can take any two F-vector spaces X_a, X_b and let $X_{ab} = X_a \oplus X_b$, with linear maps $\phi_{ab,a}$ and $\phi_{ab,b}$ the standard inclusion maps. In this sheaf, for a given stimulus y, any choice of vectors $x_a(y)$ and $x_b(y)$ is globally consistent. That is, the two sensors are understood to be making completely independent measurements, as they would if they have disjoint receptive fields or measure independent features of the stimulus. However, in this sheaf all pairs of measurements are orthogonal in X_{ab} and so the proposed definition in equation (1) is uniformly 1. To rectify this, we need to exclude portions of the sensor measurements which are independent from the computation by restricting the correlation computation to the intersection of the images of $\phi_{ab,a}$ and $\phi_{ab,b}$. We define $\pi_{ab}: X_{ab} \to X_{ab}$ to be the projection onto the subspace $im\big(\phi_{ab,a}\big) \cap im(\phi_{ab,b})$, and take

$$I_{ab}\big(x_a(y), x_b(y)\big) = \sqrt{1 - Corr\big(\pi_{ab}(\phi_{ab,a}\big(x_a(y)\big)), \pi_{ab}(\,\phi_{ab,b}\big(x_b(y)\big))\big)} \tag{2}$$

whenever $im\big(\phi_{ab,a}\big) \cap im\big(\phi_{ab,b}\big) \neq \emptyset$ and $I_{ab}\big(x_a(y), x_b(y)\big)$ uniformly zero otherwise. It is reasonable to record this latter case by excluding the edge from the graph.

Suppose now that we have two sensors with unknown consistency condition, but that we are capable of testing the system by introducing known stimuli $Y = \{y_1, y_2, \dots, y_n\}$ and recording the responses $x_a(y_i),\ x_b(y_i)$. We can use this formalism to pose the determination of the a linear coherence condition as an optimization problem. Let $X^0_{ab} = X_a \oplus X_b$ and use a norm on X^0_{ab} to define a loss function

$$L(\phi_{ab,a}, \phi_{ab,b}) = \sum_{y \in Y} \left\| \phi_{ab,a}(x_a(y)) - \phi_{ab,b}(x_b(y)) \right\| \quad (3)$$

for pairs of linear maps into X_{ab}^0, possibly including a regularization term. Finding linear maps $\tilde{\phi}_{ab,a}$, $\tilde{\phi}_{ab,b}$ which optimize such a loss function would then induce a notion of (in)consistency of measurements using the formalism described above, based solely on the observed sensor responses to stimuli. The codimension of the subspace $im(\tilde{\phi}_{ab,a}) + im(\tilde{\phi}_{ab,b})$ provides a measure of the independence of the two sensors that could be used as a coarse notion of dissimilarity, possibly under some suitable normalization. This framework can be directly extended to the case of multiple sensors, including the study of triple-wise and higher notions of sensor consistency, to produce a sheaf over an appropriate simplicial complex.

Due to the close of the period of performance, the PI was unable to theoretically investigate or implement and experimentally test the proposed method. However, there should be few substantial technical challenges to initial exploration, as the proposed optimization problem is convex optimization (in the linear map case; a similar problem can, of course, be posed in a nonlinear environment.) Further, because sheaves have a long history in algebraic topology and to homology in particular, there is substantive reason to believe that this method will be amenable to interface with existing methods in applied topology, and may admit provable relationships between the dissimilarity measure as encoded in the sheaf, and the statistics of the persistent homology derived from the resulting weighted network.

5.0 CONCLUSIONS

The initial impetus for this project was a strong trend in prior work indicating that methods from algebraic topology are likely to be useful for the study of sensor networks, as the role of sensor networks, as topological spaces, is to aggregate local data into global information. However, many of the methods of algebraic topology have classically only been developed and applied in the continuous setting, of which the discrete sensor networks we propose to study are approximations. Thus, while the intuition stands, there remains substantial work to be done bridging the gap between theory and practice.

During this project, the PI was repeatedly surprised by the complexity of technical issues that arose on both the theoretical and experimental fronts. Various phenomena that appear to have been more easily resolved in the classical continuous case, such as understanding how deformation of spaces or metrics modify the result of homology computations, do not appear amenable to the same approaches in the discrete case. While the highly nonlinear nature of the homology computation was expected to cause difficulty in understanding these relationships, it is now clear to the PI (and, with this context, it is clear that it likely has been to others before him) that developing new mathematical methods for approaching such problems is a fundamental problem for the field of applied topology. As the push to bring topological methods more broadly into scientific and engineering research continues, it is likely that many researchers will begin to encounter problems with the interface between the mathematical methods and real data. The PI believes it would be of great benefit to the applied topology community to develop a large-scale roadmap indicating where such fundamental technical difficulties are likely to lie but have yet to be carefully investigated.

In addition, since the beginning of the field, members of the applied topology community have been very interested in the use of cellular sheaves and cosheaves as data structures. Due to their perceived technical complexity, inherited from their similar, deserved reputation in the continuous case, adoption has been slow outside of a few research groups. However, many real systems, particularly sensor networks and complex systems whose constituent elements can be described in similar language, are clearly amenable to description using sheaf-theoretic language, and the PI believes that this is likely to be a fruitful and more accessible direction of inquiry as more researchers are exposed to these tools. Importing and upgrading tools from linear algebra and quiver theory to the setting of cellular sheaves will provide powerful methods for analysis of networked systems, and computational methods already on the horizon will make those methods accessible even for large networks.

LIST OF SYMBOLS, ABBREVIATIONS AND ACRONYMS

G	weighted or unweighted simple graph
V	set of vertices in a graph or simplicial complex
E	set of edges in a graph or simplicial complex
$\binom{A}{2}$	set of unordered pairs of elements from a set A
Σ	weighted or unweighted simplicial complex
S, S_i	set of simplices or i-simplices in a simplicial complex
w	weight function on edges of a weighted graph or simplices of a weighted simplicial complex
σ, τ, ρ	simplices in a simplicial complex
X, X_σ	vector space, vector space affiliated to a simplex σ
$\mathbb{R}, \mathbb{R}^d$	real numbers, real Euclidean space of dimension d
$\phi, \phi_{\sigma,\tau}$	linear transformation between vector spaces
P(A)	power set (set of all subsets) of a set A
∂_i	boundary map in a simplicial complex
$\mathbb{F}_2$	field with two elements
$\mathbb{F}_2\langle A\rangle$	$\mathbb{F}_2$-vector space with basis A
$C_i(\Sigma)$	degree *i* $\mathbb{F}_2$-chain group of simplicial complex Σ
ker (ϕ)	kernel of linear transformation ϕ
$im(\phi)$	image of linear transformation ϕ
$H_i(\Sigma)$	degree i simplicial homology of simplicial complex Σ with $\mathbb{F}_2$-coefficients
$X \oplus X'$	direct sum of vector spaces X, X'