Data is ubiquitous in today’s society: we track our locations with GPS; we record the number of steps each day using pedometers; we digitally record daily life with pictures and videos; etc. To interpret and to understand the data requires both technical theory to find important features of the data and application expertise to interpret those features. My research is primarily in the technical theory used to find important features of data. In particular, my research is an interdisciplinary approach to topological data analysis (TDA). TDA refers to a collection of methods for extracting topological structure from data.

My approach to TDA employs statistics and is application-oriented. In this statement, I describe two current research projects illustrating both the broad spectrum of possible applications for my research as well as a specific application. In both projects, I have made novel contributions and published in premiere peer-reviewed venues. As a faculty member, I plan to continue both developing the theoretical foundations and employing the theory in practical applications.

In the first project, I describe a statistical approach to persistent homology. We can summarize data in various ways. One way is to use persistent homology, a description of the (in nontechnical words) components, tunnels, and voids of the topological space represented by the data. However, persistent homology is cumbersome to use directly in order to compare data sets or to recognize meaningful features. I turn to statistics to overcome this problem.

In the second project, I work on evaluating road network reconstructions from GPS trajectory data based on both the geometry and the topology of the embedding. The datasets studied in this application are the maps created from trajectory data. I have helped develop the theory to measure the difference between two datasets, and have applied this to various real and hypothetical datasets.

### The Intersection of Statistics and Computational Topology

Persistent homology is a method for studying the homology (i.e., the components, the tunnels, and the voids) at multiple scales simultaneously. More precisely, it provides a framework to quantify the evolution of the homology of a parameterized family of topological spaces. For example, we can study the persistent homology of a time-varying coverage region of a mobile sensor network. We track the homological changes that occur as the (time) parameter changes. This information is encoded in the persistence diagram, a multiset of

![Figure 1](image-url)

Figure 1: We sample points from a circle (a), compute the distance function to that sample (b), and compute the persistence diagram (c) for the lower-level sets of that function. In addition, we mark the region around the diagonal where persistence points are indistinguishable from topological noise.
points in the plane, each corresponding to the birth and death of a homological feature that existed for some interval of time. As time $t$ varies, homology is born and dies. The pairing of birth $b$ and death $d$ times is called a persistence pair, and is recorded in $\mathcal{P}$ as the point $(b, d)$, as illustrated in Figure 1c for the lower-level set filtration of a distance function. Features that exist for a long interval $d - b$ can be viewed as topologically significant, while features with small intervals are indistinguishable from noise. Analyzing the persistence diagram $\mathcal{P}$ boils down to thresholding persistence pairs based on the value of $d - b$.

One of the current challenges in persistent homology is to identify the pertinent topological descriptors of a data set. For example, the persistence diagram could have hundreds or thousands of points, and a statistical method for recognizing important points has been missing for the first decade of research in this field. I investigate how statistics can enhance data analysis in computational topology in order to help recognize the important features and to compare different data sets.

**Contributions.** The fundamental contributions of my papers in this area include providing the definition and methods for computing a confidence set for $\mathcal{P}$. A $(1 - \alpha)$-confidence set for $\mathcal{P}$ is an estimated diagram $\hat{\mathcal{P}}$ along with a real value $c$ such that the distance between $\mathcal{P}$ and $\hat{\mathcal{P}}$ is at most $c$, with probability $1 - \alpha$. This distance $c$ can then be used as a threshold for distinguishing significant features in the persistence diagram. For example, if $\mathcal{P}$ is the persistence diagram for the lower-level set filtration of a distance function, then $\hat{\mathcal{P}}$ is the diagram corresponding to the lower-level set filtration of the distance to a sample $S$. Assuming that the sample $S$ does not have outliers, I can compute $c$ using the bootstrap or one of the other statistical techniques developed in [7, 11]. On the other hand, if $S$ has large outliers, then using the distance-to-a-measure (DTM) function may be more insightful than the distance function itself [6]. Currently, we are working on establishing confidence sets for DTM [8].

**Publications and Recognition.** Both the statistics and topology research communities have indicated interest in this research on statistical approaches to computational homology. Furthermore, I recently began a collaboration with the Rabdan Lab at Columbia University, focusing on biological data analysis. Additionally, this research has already resulted in several papers, including an article in the Annals of Statistics [11] and a conference paper at the Symposium on Computational Geometry [10]. We also have a paper in submission to the International Conference on Artificial Intelligence and Statistics (AISTATS) on using subsamples of data to compute stable topological descriptors of large data sets [9]. Along with Fabrizio Lecci, I created an R package that implements the methods that we have developed, making these tools accessible to statisticians and others who prefer to code in R. A companion paper to this software will be submitted to a special issue of the Journal of Symbolic Computation.

**Future Work.** The techniques in my aforementioned papers are developed independent of an application. To date, most problems I have applied these techniques to are embedded in $\mathbb{R}^3$: circles, spheres, tori, 3D scanned objects, and the distribution of galaxies or other matter throughout the universe. My first step as a faculty member will be to submit an NSF proposal to apply these techniques to other areas. In particular, I would like to establish connections with faculty in the other departments. Furthermore, along with my collaborators at Carnegie Mellon, we are pursing several research directions, including the computation of the power of various hypothesis tests.

**Analysis of Road Networks and GPS Trajectories**

Digital maps are an invaluable resource today, and much effort goes into keeping these maps current. In addition to detecting change, comparing maps can help evaluate reconstruction
algorithms. GPS trajectory data is readily available and algorithms exist to reconstruct a road network from a set of GPS trajectories; see e.g. [12]. A desirable distance measure to evaluate the accuracy of the reconstruction against the true map is needed. The current distance measures (e.g., the measures presented in [5]) are mostly heuristic in nature and fail to provide theoretical guarantees. I provide theoretical guarantees by defining new distance measures that explicitly use embeddings of the maps.

**Contributions.** A road network map is a description of all ways that goods or people can be transported from one place to another. One approach to comparing the maps is to compare the sets of paths defined by the maps. This approach is taken in [1], in which we use the Fréchet distance to evaluate the distance between two given paths to define the path-based (PB) distance. The Fréchet distance is informally refereed to as the dog-leash distance and can be thought of as the shortest leash that allows a man to walk forward on one path as a dog walks forward on the second path. There are several advantages to using the PB-distance. First, if the PB-distance is small, then there exists a meaningful correspondence between the vertices of the two graphs. Second, we can approximate this distance using a (relatively) small number of paths. Moreover, the computation of this distance only has one parameter: the maximum length of the paths to consider (as opposed to some of the heuristic approaches that use multiple tuning parameters). Finally, the PB-distance is directed, so a reconstructed map of a particular bus route can be compared against a map of an entire city, without penalizing the reconstruction for not recovering streets untraveled by the buses.

In a second approach, I evaluate the distance between road networks by first creating a local signature [2]. A local signature is a function that takes a point and quantifies how similar or different the two graphs are in a neighborhood of that given point. We define this distance using a concept called local persistent homology; see [4]. Both local and global topological structures are accounted for by varying the size of the neighborhood of the local homology (LH) distance signature. Moreover, this signature can be used to visualize the distances between the graphs. Figure 2b illustrates the signature: a large (red) distance is observed when roads at an intersection are missing. The next step in this project is to use the information in the signature in order to make informed decisions: for example, clustering can help find regions of high discrepancy.

**Publications and Recognition.** The research community is receptive to these ideas. In early October, I presented an overview of map comparison techniques to an audience of a few hundred people at the annual Grace Hopper Celebration of Women in Computing [3] (the acceptance rate was only 22%). The PB-distance is tentatively accepted to ACM’s Transactions on Spatial Algorithms and Systems [2] and the LH-distance will be presented at the SIGSPATIAL workshop next week [1]. Furthermore, the LH-distance has

---

2 For example, the Open Street Map project (http://www.openstreetmap.org) provides free crowdsourced data.

Figure 2: (a): If two maps have similar embeddings, the distance between the maps should be proportional to the distance between corresponding vertices. (b): One graph is drawn in gray, and the second is colored according to the local distance between the graphs.
sparked interest from researchers using local persistent homology, for example, I was invited to present the LH-distance at a Duke University Data Science Seminar.

**Future Work.** In the upcoming months, I will be working on employing different models for the road networks, including one that would allow for bridges, tunnels, and directed streets. A longer-term goal for me is to develop techniques for monitoring streaming, map-related data for the purpose of military applications. In fact, along with Carola Wenk and Yusu Wang (Ohio State University), I have submitted an NSF grant focused on annotating a map using topological and geometric information gleaned from time-stamped GPS trajectories.

**References**


