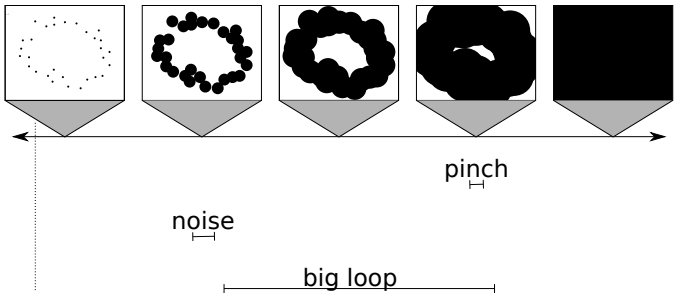


Figure 1: Given a finite point cloud, we construct a sequence of topological spaces by considering the union of balls of increasing radius, shown above from left to right. We think of the radius as increasing with time, and look at snapshots from the timeline above. The bars below the timeline are the *persistence barcode*, which illustrates how *births* and *deaths* of topological features pair.

My research is in the field of topological data analysis (TDA). One tool in TDA for summarizing data is persistent homology, which is (informally) a multi-scale way to describe the components, tunnels, and voids of the underlying set from which the data was sampled. However, a gap still exists in this field between theory and practice: how do we convey topological properties to domain experts? How do we compute (often exponential-time) topological properties over large domains? To overcome these issues, I use statistical approaches. To ground my theoretical studies, I am motivated by applications in road network analysis, prostate cancer histology images, and more. Thus, my research is both interdisciplinary and collaborative. I have made novel contributions to my research community and have published in premiere peer-reviewed venues. As I progress in my career, I plan to continue both contributing to the theoretical foundations and employing the theory in practical applications. After giving a succinct summary of persistent homology (the main tool used in my research), we discuss the main themes of my research, including my early work as a student and my current work in theory and applications.

Persistent homology is a method for studying the homology (i.e., the components, the tunnels, and the higher-dimensional '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, pairing *births* or appearances of new features with *deaths* or merging of feature classes. This information is encoded in the *persistence diagram or barcode*, a multiset of points in the plane, each corresponding to the birth-death interval of some homological feature, as illustrated in Figure 1 for the lower-level set filtration of a distance function. Features that exist for a long interval of time can be viewed as topologically significant, while features with small intervals are indistinguishable from noise. These significant features are precisely those that are far away from the diagonal $y = x$. Some of my theoretical contributions focus on understanding *what is a significant feature* and other work focuses on how to interpret persistence diagrams in different domains. For a more technical and detailed discussion of homology and persistent homology, see [16, 24].

Dissertation Work and Before

As an undergraduate student, I completed a *University Scholar* senior thesis (a year-long research project that released me from coursework). The topic of this project was algebraic topology; in particular, I worked on the classification of acyclic spaces. I've presented posters on this project at MathFest 2008 and Sigma Xi 2007 (see CV); moreover, this research has provided a foundation for my future work on *applied* algebraic topology.

As a graduate student, I proved an inequality that bounds the difference between lengths of curves by a function of the Fréchet distance between the curves and the total curvatures of the curves. This result culminated in a single-author paper [17], establishing my ability to independently conduct research. Furthermore, I have continued to work on problems related to the Fréchet distance, in the context of road network analysis; see below for a discussion of this work and [7] for a sample publication.

In another project, I studied Gaussian mixtures, a widely used–but poorly understood–data model. Gaussian mixtures (sums of Gaussian kernels) are often used to represent multi-modal distributions; however, much remains unknown about Gaussian mixtures. In fact, the number of modes, i.e., local maxima, is not well understood beyond dimension one. Given a Gaussian mixture, we associate each component with the mode closest to its center. I call any unassociated mode a *ghost mode*, as it seems to appear from nowhere. The existence of ghost modes was first shown in [8]; however, in [14, 15], we fully analyzed one case in which exactly one more mode than kernel appears. Our investigation provides a systematic way of looking for modes, as all modes, in this setting, appeared on a finite number of one-dimensional axes. In addition, our results can be extended by using Cartesian products of simplices, providing an example Gaussian mixture with a super-linear number of modes, with respect to the number of components comprising the mixture.

Since graduating with my PhD in 2012, my research has evolved to be more data-driven, but the desire to understand fundamental properties remains at the heart of most of my research.

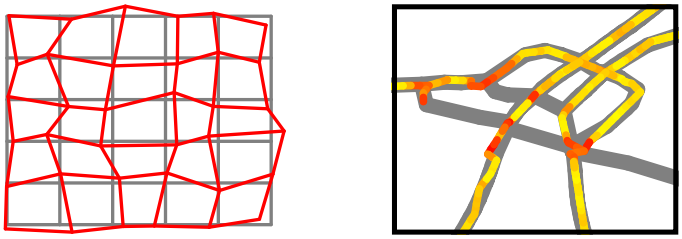
Theory: The Intersection of Statistics and Computational Topology

One of the current challenges in persistent homology is to identify the pertinent topological descriptors of a data set. I investigate how statistics can enhance data analysis in TDA. I have defined and developed algorithms to compute confidence sets for persistence diagrams using techniques such as the bootstrap [9, 11, 20], reaching a broad audience, both in the statistics and the applied topology communities. We have also developed methods for subsampling data to compute stable topological descriptors of large data sets [10]. In addition, we have made these methods available via an R package that implements our statistical methods.[1] Currently, we are investigating how to compute the power of various hypothesis tests in TDA, and will be submitting a grant to NSF DMS to support these efforts. This work is driven by applications in astronomy, comparing simulations versus observations of the distribution of matter throughout the universe [13].

Contributions. The fundamental contributions of my papers in this area include providing the definition and methods for computing a *confidence set* for a persistence diagram $\mathcal{P}$. A $(1-\alpha)$-confidence set for $\mathcal{P}$ is an estimated diagram $\widehat{\mathcal{P}}$ along with a real value c such that the distance between $\mathcal{P}$ and $\widehat{\mathcal{P}}$ 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 $\widehat{\mathcal{P}}$ 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 [9, 11, 20].

Publications and Recognition. Both the statistics and topology research communities have indicated interest in this research on statistical approaches to computational homology. Additionally, this research has already resulted in several papers, including an article in the Annals of Statistics [20] and conference papers in the Symposium on Computational Geometry [12] and the International Conference on Machine Learning [10]. Along with Fabrizio Lecci and Jisu Kim, I created an R package that implements the methods

[1] `http://cran.r-project.org/web/packages/TDA/index.html`



(a) Two Graphs (b) Local Signature

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.

that we have developed, making these tools accessible to statisticians and others who prefer to code in R.[2] A companion paper to this software will be submitted to the Journal of Statistical Software [19].

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, road networks, images, and the distribution of galaxies or other matter throughout the universe. One direction of particular interest is the move from persistence diagrams to vectorized or functional summaries of persistence diagrams. I plan to investigate general vector and functional summaries, as well as their statistical properties.

Application: 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[3] available and algorithms exist to reconstruct a road network from a set of GPS trajectories; see e.g. [22]. 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.

[2] `http://cran.r-project.org/web/packages/TDA/index.html`

[3] For example, the Open Street Map project (`http://www.openstreetmap.org`) provides free crowd-sourced data.

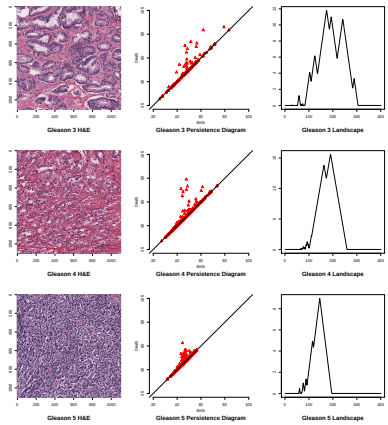


Figure 3: We compute the persistent homology of three sample ROIs, one of grades 3, 4, and 5. Notice how each sample has a very distinct persistence signature (middle) as well as functional summary (right).

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 2012, 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]. I have published both map comparison [1, 2, 6] and map construction [7] papers in respected venues (ACM SIGSPATIAL GIS, ACM's Transactions on Spatial Algorithms and Systems). Furthermore, the LH-distance has sparked interest from researchers using persistent local homology, resulting in a joint publication with Bei Wang uniting different perspectives on PLH [21].

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. In fact, along with Carola Wenk and Yusu Wang (Ohio State University), I have an NSF grant to analyze data on graphs. Over this past summer, I submitted an NSF CAREER grant, whose focus is to compare network structures across cities. A longer-term goal for me is to develop techniques for monitoring streaming, map-related data for the purpose of military applications.

Application: Data Descriptors in Pathology

The widespread availability of digital pathology images opens up new possibilities to use computational approaches to leverage the information inherent within them for diagnosis, prognosis, and precision medicine. A collaboration with researchers at Tulane University aims to discover new quantitative image-based prognostic biomarkers (data descriptors) for prostate cancer, focusing on an investigation of novel concepts from TDA applied to prostate cancer glandular architecture. Since the structures used for prostate cancer grading

are geometric and topological in nature, we plan to use persistent homology to assist with grading; see Figure 3.

Publications and Recognition. We have a paper in progress, and have presented two posters [18, 23], winning an honorable mention at the SPIE digital pathology. Two software products have been development and will be released open source on the project website after the publication of the paper in progress: (1) a collaborative tool that pathologists can use to annotate pathology slides and (2) a tool to simulate pathology slides.

Future Work. This work was initially funded by a join NSF-NIH planning grant, and we have recently been awarded a full QuBBD follow-up grant,allowing us to expand our project team to include John Sheppard (MSU) in machine learning and Brian Summa (Tulane) in visualization. Co-PI Brown at Tulane is developing 3D imaging technology for the prostate, and we recently submitted a grant to the DoD in order to extend our analysis of 2D images to 3D histology images.

Integrating Research, Education, and Service

While passionate about my core research described above, I am also interested in data science education. In particular, I am interested in encouraging a diverse group of students to enter STEM fields. Research shows that the middle school years are formative in eventual career choice. Thus, in an NSF-funded project, we are developing lesson plans for middle schools throughout the state that incorporate CS and computational thinking. In particular, this project will develop and research storytelling as a culturally responsive way to engage middle school American Indian and rural Montana students in learning computer science and computing skills. Instead of creating a new curriculum, the project will infuse computer science across the curriculum, which will help students understand that computing skills are relevant across disciplines and are important for a wide variety of professions in the workforce. The project will use Alice, an object-based educational programming environment, that has been successful by encouraging storytelling in engaging middle school students and others who are not normally exposed to programming. Using Alice, students can tell stories by placing objects in virtual worlds they have created, and then they can program by dragging and dropping tiles that represent logical structures. By integrating these computational skills, without multiplying the number of topics to be taught, the project will promote a more diverse and comprehensive understanding of the opportunities available to students with an ability to think computationally. The project will develop resources for teachers to meet the requirements of Montana's Indian Education for All (IEFA) Act, which was mandated by the state legislature in 1999 and remains a difficult requirement for many middle school teachers to incorporate in their classrooms. The project will serve over 300 students when piloting curriculum materials and will engage 50 teachers in professional development workshops on the integration of computer science and computational thinking across middle school curriculum using a storytelling approach.

The project will use a culturally responsive approach to infuse the use of storytelling (using Alice) in the curriculum, guided by Tribal Critical Theory (TribCrit), which maintains that cultural knowledge and academic knowledge are not mutually exclusive but complement each other. The project tools, which will enable middle school teachers to integrate computer science and computational thinking throughout the curriculum, will be developed using a research-driven, iterative way to be culturally responsive to the communities served. Project research will address two complementary research questions: (1) Do storytelling and storymaking serve as effective means for engaging middle-school students in computer science?; and (2) Does the integration of computing skills into the core middle-school curriculum increase instruction and student learning of these skills? In addition, the project will document the processes and evaluate the effectiveness of the TribCrit culturally responsive approach taken in integrating computer science and computational thinking into the middle school curriculum. A mixed-method approach will be used in the research, including focus

groups, small group instructional diagnosis, surveys, and pre/post measures of computational thinking/computer science knowledge. Project results will be disseminated through professional journals and conference presentations. Selected student-created artifacts (i.e., Alice virtual worlds and stories) will be presented in Montana museums.

References

[1] Ahmed, M., Fasy, B. T., Hickmann, K., and Wenk, C. A path-based distance for street map comparison, 2015. Trans. Spatial Alg. Sys. (TSAS) 2015. Preprint available at arXiv:1309.6131.

[2] Ahmed, M., Fasy, B. T., and Wenk, C. Local persistent homology based distance between maps. In *SIGSPATIAL* (Nov. 2014), ACM.

[3] Ahmed, M., Fasy, B. T., and Wenk, C. New techniques in road network comparison. In *Grace Hopper Celebr. Women Comput.* (Oct. 2014). Online proceedings.

[4] Bendich, P., Wang, B., and Mukherjee, S. Local homology transfer and stratification learning. *ACM-SIAM Symp. Discrete Algorithms* (2012).

[5] Biagioni, J., and Eriksson, J. Inferring road maps from global positioning system traces. *Transportation Research Record: Journal of the Transportation Research Board 2291*, 1 (2012), 61–71.

[6] Bittner, A., Fasy, B. T., Grudzien, M., Ghosh, S., Huang, J., Pelatt, K., Thatcher, C., Tumurbaatar, A., and Wenk, C. Comparing directed and weighted road maps. In *Research in Computational Topology*, E. Chambers, B. T. Fasy, and L. Ziegelmeier, Eds., AWM and IMA Series. Springer. to appear.

[7] Buchin, K., Buchin, M., Duran, D., Fasy, B. T., Jacobs, R., Sacristan, V., Silveira, R. I., Staals, F., and Wenk, C. Clustering trajectories for map construction. In *ACM SIGSPATIAL GIS* (Nov. 2017).

[8] Carreira-Perpinán, M., and Williams, C. An isotropic Gaussian mixture can have more modes than components. Informatics Research Report EDI-INF-RR-0185, Institute for Adaptive and Neural Computation, University of Edinbugh, Dec. 2003.

[9] Chazal, F., Fasy, B., Lecci, F., Rinaldo, A., Singh, A., and Wasserman, L. On the bootstrap for persistence diagrams and landscapes. *Modeling and Analysis of Information Systems 20*, 6 (2013), 96–105. Also available at arXiv:1311.0376.

[10] Chazal, F., Fasy, B. T., Lecci, F., Michel, B., Rinaldo, A., and Wasserman, L. Subsampling methods for persistent homology, 2014. ICML. Preprint available at arXiv:1406.1901.

[11] Chazal, F., Fasy, B. T., Lecci, F., Michel, B., Rinaldo, A., and Wasserman, L. Robust topological inference: Distance to a measure and kernel distance. *J. Mach. Learn. Res.* (2017). To appear.

[12] Chazal, F., Fasy, B. T., Lecci, F., Rinaldo, A., and Wasserman, L. Stochastic convergence of persistence landscapes and silhouettes. In *Proc. of the* 30^{th} *Annu. Symp. Comput. Geom.* (Jun. 2014).

[13] Cisewski, J., Fasy, B. T., Hellwing, M. W., Lovell, M., Rinaldo, A., Wasserman, L., and Wu, M. Topological hypothesis tests for the large-scale structure of the Universe. In preparation.

[14] Edelsbrunner, H., Fasy, B. T., and Rote, G. Add isotropic Gaussian mixtures at own risk: More and more resilient modes in higher dimensions. In *Proc. of the* 27^{th} *Annu. Symp. Comput. Geom.* (Jun. 2012), ACM. Symposium held in Chapel Hill, NC.

[15] Edelsbrunner, H., Fasy, B. T., and Rote, G. Add isotropic Gaussian mixtures at own risk: More and more resilient modes in higher dimensions. *Discrete Comput. Geom.* (Jun. 2013), 797–822.

[16] Edelsbrunner, H., and Harer, J. *Computational Topology: An Introduction.* AMS, Providence, RI, 2010.

[17] Fasy, B. T. The difference of length in curves in $\mathbb{R}^n$. *Acta Sci. Math. (Szeged) 77* (2011), 359–367.

[18] Fasy, B. T., Brown, J. Q., Wenk, C., Lawson, P., and Miller, C. Towards an automated quantitative diagnosis of prostate cancer, 2016. Poster presentation, BD2K All-hands Meeting and Open Data Science Symposium.

[19] Fasy, B. T., Kim, J., Lecci, F., and Maria, C. Introduction to the R package TDA. Package available on CRAN. Preprint available at ArXiv:1411.1830.

[20] Fasy, B. T., Lecci, F., Rinaldo, A., Wasserman, L., Balakrishnan, S., and Singh, A. Confidence sets for persistence diagrams. *Annals of Statistics 42*, 6 (2014), 2301–39. Preprint available at ArXiv:1303.7117.

[21] Fasy, B. T., and Wang, B. Exploring persistent local homology in topological data analysis. In *International Conf. on Acoustics, Speech and Signal Processing (ICASSP)* (2016).

[22] Karagiorgou, S., and Pfoser, D. On vehicle tracking data-based road network generation. SIGSPATIAL '12, ACM, pp. 89–98.

[23] Lawson, P., Berry, E., Brown, J. Q., Fasy, B. T., and Wenk, C. Topological descriptors for quantitative prostate cancer morphology analysis. In *Conf. Digital Pathology, SPIE Medical Imaging* (2017).

[24] Munkres, J. R. *Algebraic Topology.* Prentice Hall, Upper Saddle River, NJ, 1964.