samsi
NSF • Duke • NCSU • UNC • NISS

Topological Data Analysis Workshop
February 3-7, 2014

SPEAKER TITLES/ABSTRACTS

Yuliy Baryshnikov
University of Illinois

“What is the Dimension of the Internet?”

The large-scale structure of the Internet (or, rather of the graph of the Autonomous System Nodes) has been attracting a lot of attention by the researchers for decades. One family of attractive models for this graph stipulates that it "looks like" is if sampled from a hyperbolic plane. We discuss possible tests for the dimension of samples from manifolds, and apply them to the ANS graph.

Joint work with Yu. Mileyko (UH).

Paul Bendich
Duke University

“Persistent Local Homology: Theory, Algorithms, Applications, New Directions”

This talk fits into the general topic of 'stratification learning,' wherein one tries to make inferences about data based on some assumption that it is sampled from a mixture of manifolds glued together in some nicely-structured way. The theoretical tool of persistent local homology (PLH), now more than five years old, provides a useful way to understand the local singularity structure of the input dataset, where what one means by "local" can be thought of as a multi-scale input parameter.

In this talk, we will give an intuitive introduction to PLH, discuss recent algorithmic and implementation improvements, and show a few applications. The talk will be as non-technical as humanly possible, and will (hopefully, time permitting!) serve as a survey of the work of many people, including but not limited to the speaker, John Harer, Herbert Edelsbrunner, Dmitriy Morozov, David Cohen-Steiner, Bei Wang, Sayan Mukerjee, Primoz Skraba, Ellen Gasparovic, Fengtao Fan, Yusu Wang, and Tamal Dey.

Omer Bobrowski
Duke University

“Phase Transitions in Random Čech Complexes”

In manifold learning, one often wishes to infer geometric and topological features of an unknown manifold embedded in a d-dimensional Euclidean space from a finite (random) point cloud. One topological invariant of a considerable interest is the homology of the underlying space.

A common method for recovering the homology of a manifold from a set of random samples is to cover each point with a d-dimensional ball and study the union of these balls. By the Nerve Lemma, this method is equivalent to study the homology of the Čech complex generated from the random point cloud.

In this talk we discuss the limiting behavior of random Čech complexes as the sample size goes to infinity and the radius of the balls goes to zero. We show that the limiting behavior exhibits multiple phase transitions at different levels, depending on the rate at which the radius of the balls goes to zero. We present the different regimes and phase transitions discovered so far, and observe the nicely ordered fashion in which homology groups of different dimensions appear and vanish. One interesting consequence of this analysis is a sufficient condition for the random Čech complex to successfully recover the homology of the original manifold.

Peter Bubenik
Cleveland State University

“Statistical Topological Data Analysis using Persistence Landscapes”

In this talk I will define a topological summary for data that I call the persistence landscape. Since this summary lies in a vector space, it is easy to calculate averages of such summaries, and distances between them. Viewed as a random variable with values in a Banach space, this summary obeys a Strong Law of Large Numbers and a Central Limit Theorem. I will show how a number of standard statistical tests can be used for statistical inference using this summary.

Frederic Chazal
INRIA

“Stability and Convergence Properties of Persistence Diagrams in Topological Data Analysis”

In TDA, persistent homology appears as a fundamental tool to infer relevant topological information from data. Persistence diagrams are usually computed from filtrations built on top of data sets sampled from some unknown (metric) space. They provide "topological signatures" revealing the structure of the underlying space. To ensure the relevance of such signatures, it is necessary to prove that they come with stability properties with respect to the way data are sampled. In this talk, we will introduce the use of persistent homology in TDA and present a few results on the stability of persistence diagrams built on top of general metric spaces. We will show that the use of persistent homology can be naturally considered in general statistical frameworks and persistence diagrams can be used as statistics with interesting convergence properties.

Harish Chintakunta
North Carolina State University

"Distributed Homology using Harmonics"

The talk will have two primary themes: 1) use of harmonics in computing homology, and 2) localizing topological features and simplification of a given complex distributively. Harmonics are elements in the null space of the Laplacian and are very easy to compute. I will show how some simple properties of harmonics can be exploited to determine contractible and homologous cycles without having to reduce huge matrices. This in turn enables distributed computations to localize topological features, and to verify persistence. I will show how these techniques can be used to localize deployment failures in sensor networks, and extend the techniques for efficient computation of homology generators in general.

Jessi Cisewski
Carnegie Mellon University

“Persistent Homology of the Intergalactic Medium via the Lyman-alpha Forest"

Light we observe from quasars has traveled through the intergalactic medium (IGM) to reach us, and leaves an imprint of some properties of the IGM on its spectrum. There is a particular imprint of which cosmologists are familiar, dubbed the Lyman-alpha forest. From this imprint, we can infer the distribution of neutral hydrogen along the line of sight from us to the quasar. The Sloan Digital Sky Survey Data Release 9 (SDSS – DR9) produced over 54,000 quasar spectra that can be used for analysis of the Lyman-alpha forest and, thus, aid cosmologists in further understanding the IGM along with revealing or corroborating other properties of the Universe.

With cosmological simulation output, we use local polynomial smoothing to produce a 3D map of the IGM. Describing the topological features of the IGM can aid in our understanding of the large-scale structure of the Universe, along with providing a framework for comparing cosmological simulation output with real data beyond the standard measures. I will illustrate how persistent homology can be used in this setting.

Brittany Fasy
Tulane University

“Local Homology Based Distance Between Maps”

We define a topology-based distance measure between road networks embedded in the plane. This distance measure is based on local homology, but does not require explicitly mapping one road network (or subnetwork) to another road network (or subnetwork). This work is motivated by comparing different road networks from different data sources and to access the quality of map construction algorithms. If time allows, I will demonstrate how we can overcome this hurdle by using the bootstrap to estimate this distance between the unknown ground truth and a reconstruction.

Jennifer Gamble
North Carolina State University

"Quantifying Coverage in Dynamic Sensor Networks with Zigzag Persistent Homology"

The use of homology as a tool to describe coverage in a sensor network was introduced by de Silva and Ghrist: a simplicial complex is built using local information about which sensors are in communication range of each other, and the homology of this complex can be used to make global coverage guarantees (assuming a specific coverage model, and relationship between sensing and communication radii). Extending this to the time-varying setting, a dynamic sensor network may be represented as a sequence of simplicial complexes, and zigzag persistence used to compute the lifetimes of homological features in the network over time. These lifetimes are then summarized in a barcode or persistence diagram.

We will present ways in which the barcodes/persistence diagrams can quantify information about the time-varying coverage in the network. While an exact correspondence between bars and coverage holes is not possible (for example, Adams and Carlsson show that the existence of a homology class persisting over an interval does not imply the existence of a corresponding evasion path over the same interval), the barcode can still present an overview of how well the network is covered over time. Moreover, we will discuss ways to glean geometrically-relevant information using adaptive choices of representative cycles for the homology classes, along with a hop-distance based filtration.

Giseon Heo
University of Alberta

"Topological and Statistical Data Analysis"

Persistent homology is a recently established topological technique that has been found useful in high dimensional data analysis. Persistent homology studies the history of the true features of complex data over a wide range of scales. It can distinguish between the innate properties of an unknown space and the noise. Persistent homology has three descriptors: barcodes, persistence diagrams, and persistence landscapes. The notion of persistence landscapes, introduced by Bubenik (2012), has become a useful tool in statistical inference. We illustrate how persistent homology can be incorporated in statistical analysis through an example.

Peter Kim
University of Guelph

"Pyrosequencing and Computational Topology: A tale of two homologies"

With the introduction of massively parallel sequencing (MPS) technologies, we are in a position to capture pictures of the microbiome at specific timepoints. MPS allows us to consider interactions between bacteria and their environment with an accuracy greater than achieved in vitro. We describe the process from microbiome sampling to sequence production and cleaning, and the first of two homologies. We describe some of the transformations possible, including the generation of phylogenetic trees for the bacteria in a specimen. We propose a means of applying persistent homology to phylogenetic trees, and speculate on the future of this application.

Fabrizio Lecci
Carnegie Mellon University

“Statistical Inference for Persistence Diagrams”

Persistent homology probes topological properties from point clouds and functions. By looking at multiple scales simultaneously, one can record the births and deaths of topological features as the scale varies. We use several statistical techniques to derive confidence sets for persistence diagrams that allow us to separate topological signal from topological noise. This is joint work with Sivaraman Balakrishnan, Brittany Terese Fasy, Alessandro Rinaldo, Aarti Singh, and Larry Wasserman.

Lek-Heng Lim
University of Chicago

“Hodge Theory in Data Analysis”

The usual "differentiable Hodge theory" on Riemannian manifolds has been very useful in physical problems like fluid dynamics and electromagnetics. However in data analytic problems, one often has only some knowledge of the proximity of data points or of the distribution of the data set, and in these cases a "continuous Hodge theory" on metric spaces and a "discrete Hodge theory" on simplicial complexes are more relevant. Furthermore, as we will argue, the discrete version can be readily appreciated by engineers and other practitioners. We will discuss two applications in ranking and game theory. Time permitting, we will also briefly describe other applications to graphics, imaging, learning, numerical analysis, robotics, and sensor networks.

J.S. Marron
University of North Carolina

“OODA of Tree-Structured Data Objects”

The field of Object Oriented Data Analysis has made a lot of progress on the statistical analysis of the variation in populations of complex objects. A particularly challenging example of this type is populations of tree-structured objects. Deep challenges arise, which involve a marriage of ideas from statistics, geometry, and numerical analysis, because the space of trees is strongly non-Euclidean in nature. These challenges, together with three completely different approaches to addressing them, are illustrated using a real data example, where each data point is the tree of blood arteries in one person's brain.

Facundo Memoli
Ohio State University

“Curvature Sets over Persistence Diagrams”

We study the structure of collections of persistence diagrams that arise from taking the Vietoris-Rips filtration of all n-tuples of points from a given metric measure space. We consider what is the induced probability measure on that collection, and study stability of this measure in the Gromov-Wasserstein sense. These ideas provide a notion of statistics over persistence diagrams which is robust to perturbations in the input metric measure spaces.

Elizabeth Munch
Institute for Mathematics and its Applications

“Categorification of Reeb Graphs”

In order to understand the properties of a real-valued function on a topological space, we can study the Reeb graph of that function. The Reeb graph is a construction which summarizes the connectivity of the level sets. Since it is efficient to compute and is a useful descriptor for the function, it has found its place in many applications. As with many other constructions in computational topology, we are interested in how to deal with this construction in the context of noise. In particular, we would like a method to "smooth out" the topology to get rid of, for example, small loops in the Reeb graph.

In this talk, we will define a generalization of a Reeb graph as a functor. Using the added structure given by category theory, we can define interleavings on Reeb graphs which can be used to compare them. This also gives an immediate method for topological smoothing and we will discuss an algorithm for computing this smoothed Reeb graph.

This is joint work with Vin de Silva and Amit Patel.

Andrew Nobel
University of North Carolina

“Some Uniformity Results for Dynamical Systems”

Uniform laws of large numbers play an important role in the theory and application of machine learning. Beginning with work of Vapnik and Chervonenkis, there is a substantial literature on uniform laws of large numbers for independent data. This talk will survey some extensions of this work to dependent data, including deterministic data generated by iteration of a fixed measure preserving map. The talk is intended to be self-contained: no prior knowledge of machine learning is assumed.

Megan Owen
Lehman College CUNY

“Mean and Variance of Metric Trees”

Data generated in such areas as medical imaging and evolutionary biology are frequently tree-shaped, and thus non-Euclidean in nature. As a result, standard techniques for analyzing data in Euclidean spaces become inappropriate, and new methods must be used. One such framework is the space of metric trees constructed by Billera, Holmes, and Vogtmann. This space is non-positively curved (hyperbolic), so there is a unique geodesic path (shortest path) between any two trees and a well-defined notion of a mean tree for a given set of trees. Furthermore, this geodesic path can be computed in polynomial time, leading to a practical algorithm for computing the mean and variance. We look at the mean and variance of distributions of phylogenetic trees that arise in tree inference, and compare with them with existing measures of consensus and variance.

This is joint work with Daniel Brown.

Vic Patrangenaru
Florida State University

“Neighborhood Hypothesis Testing for Mean Contour Shapes of Corpus Callosum Mid Sections”

Persistent homology as well as nonparametric statistics on manifolds techniques were applied by Heo et. al.(2012) to analyze high dimensional landmark based similarity shape data. Shapes of contours can be regarded as points on a projective space of a complex Hilbert space, that has a free cohomology algebra over the integers with a degree 2-generator, a fact that potentially limits topological data analysis methods, given the finiteness of the sample size. Classical asymptotic tests on this Hilbert manifold fail as well, given that the sample covariance matrix is always degenerate. Here we present the neighborhood hypothesis testing methodology on a Hilbert manifold as developed by Ellingson et. al. (2013), and apply it to shape analysis of contours of corpus callosum midsagittal sections data extracted from MRI images given in Fletcher (2013).

Co-authors: Leif Ellingson (Texas Tech University) and Mingfei Qiu (Florida State University)

Jose Perea
Duke University

“Persistent Homology of Time-delay Embeddings”

We present in this a talk a theoretical framework for studying the persistent homology of point clouds from time-delay (or sliding window) embeddings. We will show that maximum 1-d persistence yields a suitable measure of periodicity at the signal level, and present theorems which relate the resulting diagrams to the choices of window size, embedding dimension and field of coefficients. If time permits, we will demonstrate how this methodology can be applied to the study of periodicity on time series from gene expression data.

Katharine Turner
University of Chicago

“Can Topological Summary Statistics be Sufficient?”

Topological summary statistics can be used both as a way of summarizing an entire set of data (such as a point cloud which should be close to some density distribution we want to know) and as a way to summarize individual instances of data (such as when each data entry is a shape). Given a particular model, a statistic (or set of statistics) is considered sufficient if it provides as much information as the raw data. For some models, there is the potential for topological summary statistics to provide sufficient statistics. We will explore some possibilities.

Bei Wang
University of Utah

“Geometric Inference on Kernel Density Estimates”

We show that geometric inference of a point cloud can be calculated by examining its kernel density estimate. This intermediate step results in the inference being statically robust to noise and allows for large computational gains and scalability (e.g. on 100 million points). In particular, by first creating a coreset for the kernel density estimate, the data representing the final geometric and topological structure has size depending only on the error tolerance, not on the size of the original point set or the complexity of the structure.

To achieve this result, we study how to replace distance to a measure, as studied by Chazal, Cohen-Steiner, and Merigot, with the kernel distance. The kernel distance is monotonic with the kernel density estimate (sublevel sets of the kernel distance are superlevel sets of the kernel density estimate), thus allowing us to examine the kernel density estimate in this manner. We show it has several computational and stability advantages. Moreover, we provide an algorithm to estimate its topology using weighted Vietoris-Rips complexes.

Joint work with: Jeff M. Phillips and Yan Zheng.

Yusu Wang
Ohio State University

"Data Sparsification in Inferring Topology of Manifolds"

In recent years, a considerable progress has been made in analyzing data for inferring the topology of a space from which the data is sampled. Current popular approaches often face two major problems. One concerns with the size of the complex that needs to be built on top of the data points for topological analysis; the other involves selecting the correct parameter to build them. In this talk, I will describe some recent progress we made to address these two issues in the context of inferring homology from sample points of a smooth manifold sitting in an Euclidean space. I will describe how we sparsify the input point set and to build a complex for homology inference on top of the sparsified data, without requiring any user supplied parameter. More importantly, we show that (i) the data is sparsified at least to the level as specified by the so-called local feature size; (ii) the sparsified data is adaptive as well as locally uniform, and (iii) supports further homology inference without any scale parameter.

This is joint work with Tamal K. Dey and Dong Zhe.

Larry Wasserman
Carnegie Mellon University

"Statistical Inference for Functional Summaries of Persistent Homology"

A persistence diagram can be converted into a function, called a functional summary. A leading example is the landscape function invented by Peter Bubenik. We consider the statistical properties of functional summaries including convergence and nonparametric inference.

This is joint work with Fred Chazal, Brittany Fasy, Fabrizio Lecci and Alessandro Rinaldo.