**Summer@ICERM 2017: Topological Data Analysis**

**Organizers:** Jeff Brock (Brown), Katherine Kinnaird (Brown), Facundo Mémoli (Ohio State), José Perea (Michigan State).

**Lecturers:** Sara Kališnik Verovšek (Brown), Henry Adams (Colorado State).

**TAs:** Leo Betthauser (UF), Samir Chowdhury (OSU), Greg Henselman (UPenn), Melissa McGuirl (Brown), Yitzchak Solomon (Brown), Christopher Tralie (Duke), Yang Xiao (Brown).

**Focus:** Topological Data Analysis. The proposed sample projects revolve around the exploratory analysis and classification of datasets using ideas from applied algebraic topology and metric geometry, and emphasize the computational aspect of the ideas.

# 1 Background

Many datasets can be modeled as finite metric spaces. Given a finite metric space $\mathcal{X} = (X, d_X)$, for a given *scale parameter* $\varepsilon \geq 0$ one considers $R_\varepsilon(\mathcal{X})$, the Vietoris-Rips simplicial complex associated to $\mathcal{X}$. Its vertex set is $X$ and its simplices are all non-empty $\sigma \subset X$ s.t. $\mathrm{diam}(\sigma) \leq \varepsilon$. Given a field $\mathbb{F}$ and a non-negative integer $k$ one can then associate to $\mathcal{X}$ the vector space $\mathrm{H}_k(R_\varepsilon(\mathcal{X}), \mathbb{F})$ produced by the homology functor applied to the Vietoris-Rips complex of $\mathcal{X}$ at scale $\varepsilon$. Some indication about the number of connected components, loops, etc in $\mathcal{X}$ at scale $\epsilon$ can be obtained in this way. This methodology has the drawback that these numbers of connected components and so forth can change dramatically with the scale $\varepsilon$. It then becomes apparent that a more suitable idea is to try to track homology generators accross scales, which is the main idea behind *Persistent Homology*.

## Persistent Homology

If $0 = \varepsilon_0 < \varepsilon_1 < \ldots \varepsilon_m$ are the different values attained by the distance function $d_X$, then we have the inclusions: $R_{\varepsilon_0}(\mathcal{X}) \subseteq R_{\varepsilon_1}(\mathcal{X}) \subseteq \cdots \subseteq R_{\varepsilon_m}(\mathcal{X})$, i.e. a simplicial filtration which we denote by $\mathcal{R}(\mathcal{X})$. Applying the homology functor $\mathrm{H}_k(\cdot, \mathbb{F})$ to the above diagram one obtains a diagram of vector spaces and linear maps

$$\mathbb{V} = \quad V_{\varepsilon_0} \to V_{\varepsilon_1} \to \cdots \to V_{\varepsilon_m} \tag{1}$$

where each $V_{\varepsilon_i} := \mathrm{H}_k(R_{\varepsilon_i}(\mathcal{X}), \mathbb{F})$. Any such diagram of vector spaces is called a *persistence module* and can be decomposed [5] up to isomorphism as the direct sum of *interval persistence modules*: diagrams of the form

$$\mathbb{I}(b, d) = \quad 0 \to \cdots \to 0 \to \mathbb{F} \xrightarrow{1} \cdots \xrightarrow{1} \mathbb{F} \xrightarrow{0} 0 \to \cdots \to 0,$$

where $b \leq d$ represent the first and last index of appearance of $\mathbb{F}$ in the sequence. The upshot is that to any finite metric space $\mathcal{X}$ and non-negative integer $k$ one can assign a *persistence diagram* $\mathrm{D}_k(\mathcal{X})$ consisting precisely of a finite index set $A$ together with pairs $(b_\alpha, d_\alpha) \in \mathbb{R}^2_+$, $b_\alpha \leq d_\alpha$ for each $\alpha \in A$, such that the persistent module given by (1) is isomorphic to $\bigoplus_{\alpha \in A} \mathbb{I}(b_\alpha, d_\alpha)$. Persistence diagrams of finite metric spaces can be computed in polynomial time [8]. Furthermore, for fixed $k$, the map $\mathcal{X} \mapsto \mathrm{D}_k(\mathcal{X})$ is 2-Lipschitz: on the one hand one endows the collection $\mathfrak{M}$ of all finite metric spaces with the Gromov-Hausdorff distance $d_{GH}$ and on the other one considers $\mathfrak{D}$ the collection of all finite multisets of pairs of points $(b, d) \in \mathbb{R}^2_+$, $b \leq d$, and endows this collection with the so called *bottleneck distance* $d_B$. Then,

**Theorem 1** ([6]). *For any non-negative integer $k$ and all $\mathcal{X}, \mathcal{Y} \in \mathfrak{M}$,*

$$2\, d_{GH}(\mathcal{X}, \mathcal{Y}) \geq d_B(\mathrm{D}_k(\mathcal{X}), \mathrm{D}_k(\mathcal{Y})).$$

The computation of the Gromov-Hausdorff distance between finite metric spaces is NP-hard in general [12]. In contrast, computing the bottleneck distance can be done in polynomial time [8]. Thus, the stability theorem above suggests that using persistent diagrams for data classification is a promising approach.

# 2 Activities

**Mini-course: Persistent Homology from the computational viewpoint.** The main ideas of PH will be introduced via a tutorial combining theoretical concepts with their software implementation. The Matlab frontend of `javaplex` will be used for all demostrations, tutorials, and all applications. The package is freely available from `http://appliedtopology.github.io/javaplex/` and it is readily distributed together with an excellent tutorial that guides students both though the landscape of theoretical ideas and their immediate computational implementation. We envision covering both simplicial homology (with field coefficients) and its computation at the same time: we will be complementing standard definitions with the basic methods from linear algebra for reducing matrices and effectively finding bases for homology. These methods are readily implemented by `javaplex`. We will then introduce functoriality and progress to the concept of persistent homology and its computation.

**Mini-course: Distances between Metric Spaces and applications.** The main ideas behind the construction and applications of the Gromov-Hausdorff distance will be presented in a series of lectures.

**Mini-course: Topological Time Series Analysis.** Time series are ubiquitous in today's data rich world, so naturally their analysis is a fundamental object of study. In recent years, tools from the growing field of topological data analysis have been adapted to the analysis of time series data. In short, time series can be transformed into high-dimensional point clouds (via delay-embeddings) and their shape can be probed via persistent homology to quantify characteristics such as periodicity, quasiperiodicity, existence of motifs, presence of dynamic chaos, etc [13, 14]. This mini-course we will cover some of the theory behind topological time series analysis, and will explore applications ranging from biology to music analysis.

# 3 Projects

Projects are designed to blend theoretical elements with the computational exploration of real datasets. Students will explore PH ideas from different angles related to the classification of datasets into different categories.

**Shape Classification.** In the problem of shape categorization under deformations one is presented with a database of 2D or 3D shapes which need to be clustered into "geometrically similar" groups. For instance, the shapes in the database could comprise 3D scans of humans, furniture, animals, where the same subject may be present in the database in different poses. This situation is often dealt with by modeling the shapes as finite metric spaces: each shape $X$ is endowed with the path-length distance arising from a suitable geometric graph induced the point cloud representing the shape. This choice of the metric ensures that different poses of the same object will in practice be *quasi-isometric*.

Students will implement the whole pipeline leading to the computation of the bottleneck distance between persistence diagrams associated to shapes. Then several different databases of shapes (such as [15]) will be classified using these algorithms.

**Action recognition.** Algorithms for the automatic detection and recognition of human actions, such as walking, running, jumping, from real time data, are useful in applications spanning security, sports, healthcare, amongst others. This project will deal with the problem of action recognition from video data. By studying the periodicity structure of sequences of images one can classify video sequences into different activities. Students will use techniques from topological time series analysis to classify video sequences such as `http://www.nada.kth.se/cvap/actions/`.

**Feature Selection from Persistent Diagrams.** A further step in the direction of classifying shapes is related to *coordinatizing* the space $(\mathfrak{D}, d_B)$ of persistence diagrams [1]. This permits invoking powerful machine learning techniques in order to build *classifiers*.

Students will explore a choice of coordinates proposed by Sara Kalisnik in [9] which are stable under bottleneck distance. The goal is to find salient features/coordinates that permit efficient classification of shapes/datasets into groups reflecting different "hypotheses" (e.g. the different type of objects in the case of a database of shapes).

**Local Persistence Diagrams.** The construction of the Rips filtration of a finite metric space and subsequent persistence diagrams operated at the global level. A question of great interest is to localize such constructions while retaining stability. One construction that students will study can be described as follows: Given a dataset $\mathcal{X} = (X, d_X)$, and $\lambda \geq 0$ to each point $x \in X$ assign the filtration $\mathcal{R}_x^\lambda(X)$ where for each $\varepsilon \geq 0$, $\sigma \subseteq X$ is in the $\varepsilon$-slice of the filtration if and only if $\max\left(\mathrm{diam}(\sigma), \lambda \cdot \max_{p\in\sigma} d_X(x,p)\right) \leq \varepsilon$; when $\lambda = 0$ this reduces to the (global) Rips filtration $\mathcal{R}(\mathcal{X})$ of $\mathcal{X}$. Students will study the stability properties of this construction and variants, together with its computational applications and performance for classifying shapes in a database.

**Configuration spaces.** Given a compact metric space $(X, d_X)$ and a natural number $n$ Gromov considers the $n$-th curvature set $K_n(X)$ of $X$: the set comprising all $n \times n$ matrices arising from the restriction of the metric of $X$ to all possible $n$-tuples (possibly with repetition). These invariants are strong: It is known that the sequence $K_1(X), K_2(X), K_3(X), \ldots$ determines $(X, d_X)$ up to isometry. However, very little else is currently known about these invariants. One known example is $K_3$ of spheres: whereas $K_3(\mathbb{S}^1) \simeq \mathbb{S}^2$ [12], one has $K_3(\mathbb{S}^n) \simeq *$ for all $n \geq 2$. But more information is available since each $K_n(X)$ can itself be regarded as a metric space when endowed with the metric coming from the $\ell_\infty$ matrix norm — in particular one can study each curvature set via PH. In this project students will explore the characterization of curvature sets of spheres and tori using persistent homology constructions on random samples taken from these manifolds.

**Künneth formula for persistent homology.** In this project students will explore the following situation: Given two finite metric spaces $\mathcal{X} = (X, d_X)$ and $\mathcal{Y} = (Y, d_Y)$ and an assignment $\mathrm{prod}(\mathcal{X}, \mathcal{Y})$ of a metric on $X \times Y$, express the persistence diagrams of the Vietoris-Rips filtration $\mathcal{R}(\mathrm{prod}(\mathcal{X}, \mathcal{Y}))$ based on those of $\mathcal{R}(\mathcal{X})$ and $\mathcal{R}(\mathcal{Y})$.

**Classification of Music Data Streams: Music Information Retrieval.** Many musical streams may be represented by a similarity matrix [10, 11]. As such, these datasets are amenable to analysis via TDA related ideas. In this projects students will apply Persistent Homology techniques in order to automatically classify databases of song data into different genres.

**Analysis of hippocampal networks.** It is believed that ensembles of cells in the hippocampus of animals have the ability of storing spatial memories [4, 3, 7]. These cell ensembles produce time series data in the form of spike trains. In the course of a certain time interval, experimentalists obtain spike trains from about a hundred different hippocampal cells. The concerted spiking patterns of these ensembles of cells encode for the location of an animal in-

side its habitual environment, and it is believed that the by studying these patterns one can recover structural information about the environment itself: how many obstacles are there in the environment? how many different tunnels, etc? One approach for answering these questions relies on representing these cell ensembles as *networks*: directed weighted graphs with vertex set coinciding with the set of cells. The weights associated to a directed edge are often defined to be the time-delayed correlation of the spike trains corresponding to the two intervening cells. Methods based on computing the persistent homology of these networks have been demonstrated to successfully recover some topological features of the environment.

In this project students will (1) learn about this neuroscience problem, will (2) explore different ways of encoding the co-spiking activity of cells as networks, and (3) will analyze the resulting networks using PH techniques.

# 4 Professional Development

Students will learn how to write reports in LaTeX. In order to help them develop their communication skills, each team will be asked to give a 10-15 min. presentation every Friday. Other students will be encouraged to give feedback. We will hold panels with Brown graduate students on (1) graduate school decisions/applications, and (2) graduate fellowships (NSF GRFP, DoD NDSEG, DOE CSGF).

# References


[1] Aaron Adcock, Erik Carlsson, and Gunnar Carlsson. *The ring of algebraic functions on persistence bar codes.* `http://comptop.stanford.edu/u/preprints/multitwo.`

[2] Gunnar Carlsson. Topology and data. *Bulletin of the American Mathematical Society*, 46(2):255–308, 2009.

[3] Yuri Dabaghian, Facundo Mémoli, Loren Frank, Gunnar Carlsson. A topological paradigm for hippocampal spatial map formation using persistent homology. PLoS Comput Biol, 2012.

[4] Carina Curto, Vladimir Itskov. Cell groups reveal structure of stimulus space PLoS Comput Biol 4 (10), e1000205

[5] Gunnar Carlsson and Vin De Silva. Zigzag persistence. *Foundations of computational mathematics*, 10(4):367–405, 2010.

[6] Frédéric Chazal, David Cohen-Steiner, Leonidas J Guibas, Facundo Mémoli, and Steve Y Oudot. *Gromov-hausdorff stable signatures for shapes using persistence.* In *Computer Graphics Forum*, volume 28, pages 1393–1403. Wiley Online Library, 2009.

[7] Chad Giusti, Eva Pastalkova, Carina Curto and Vladimir Itskov, "Clique topology reveals intrinsic geometric structure in neural correlations" in Proceedings of the National Academy of the Sciences USA, 2015.

[8] Herbert Edelsbrunner and John Harer. *Computational topology: an introduction.* American Mathematical Soc., 2010. [12] F.Mémoli. Some Properties of Gromov–Hausdorff Distances. Discrete and Computational Geometry, 2012.

[9] Sara Kališnik. *Tropical Coordinates on the space of Persistence Diagrams*, `http://arxiv.org/pdf/1604.00113v1.pdf`.

[10] Katherine M. Kinnaird. *Aligned Hierarchies for Sequential Data.* PhD thesis, Dartmouth College, 2014.

[11] Katherine M. Kinnaird. *Aligned Hierarchies: A Multi-Scale Structure-Based Representation for Music-Based Data Streams*, ISMIR 2016.

[12] Facundo Mémoli. *Some Properties of Gromov–Hausdorff Distances." Discrete & Computational Geometry* 48.2 (2012): 416-440.

[13] Jose Perea and John Harer. "Sliding windows and persistence: An application of topological methods to signal analysis." Foundations of Computational Mathematics 15.3 (2015): 799-838.

[14] Jose Perea et al. "Sw1pers: Sliding windows and 1-persistence scoring; discovering periodicity in gene expression time series data." BMC Bioinformatics 16.1 (2015): 257.

[15] `http://tosca.cs.technion.ac.il/book/resources_data.html`