Stevanovich Center
for Financial Mathematics
at the University Of Chicago

5727 South University Avenue
Chicago, IL 60637
773-834-8563

May 20-23, 2019

Geometric Data Analysis

ABSTRACTS

This conference is made possible by the generous philanthropy of
University of Chicago Trustee Steve G. Stevanovich

Yuliy Baryshnikov	Sketches of Manifolds	I will discuss some more and less standard procedures of sketching parametric families of topological spaces, and some of their properties, as well as applications.
Mikhail Belkin	Rethinking the Bias-variance Trade-off	The bias-variance trade-off, often represented as a U-shaped risk curve, is one of the central ideas of machine learning and statistical inference. The key idea it to find a “sweet spot” between under-fitting and over-fitting, where classifiers are sufficiently complex to learn underlying structure but simple enough to avoid learning patterns that are spurious. The textbook corollary is that fitting the data exactly (interpolating) leads to poor performance and is to be avoided, a view still widely accepted. However, it is now a common practice to fit highly complex models such as deep neural networks with (nearly) zero training error. These models would classically be considered overfit to the training data, yet they consistently obtain high accuracy on data previously unseen in training. In this talk I will show how classical bias-variance trade-off and modern machine learning can be reconciled within a new unified "double descent" risk curve that extends the classical U-shaped curve beyond the point of interpolation. I will describe the mechanism underlying the emergence of this curve, its ubiquity for a range of methods and datasets, as well as theoretical evidence that interpolated classifiers can indeed be statistically optimal, even for noisy data. Finally, I will discuss a number of implications understanding and optimizing machine learning models.
Paul Bendich	Self-similarity Matrices for High-dimensional Time Series: Applications to Cross-modal Comparison, Heterogeneous Sensor Fusion, and Phase-aware Data Compression	The self-similarity matrix (SSM) is a two-dimensional image representation of a time-ordered sequence of points in any metric space. This talk surveys several recent (and very recent) applications of SSMs, with a focus on time series data arising from sensor arrays corresponding to multiple sensing modalities. Two powerful tools arising from multi-scale geometric principles, similarity network fusion and the scattering transform, will be described. Payoffs will include novel unsupervised algorithms for heterogeneous data fusion that adapt gracefully to constrained communication situations. *Different aspects of this work involve collaborations with Christopher J. Tralie, John Harer, Lihan Yao, Nathan Borggren, Ken Stewart, Jay Hineman, Abraham Smith, and Peter Zulch*

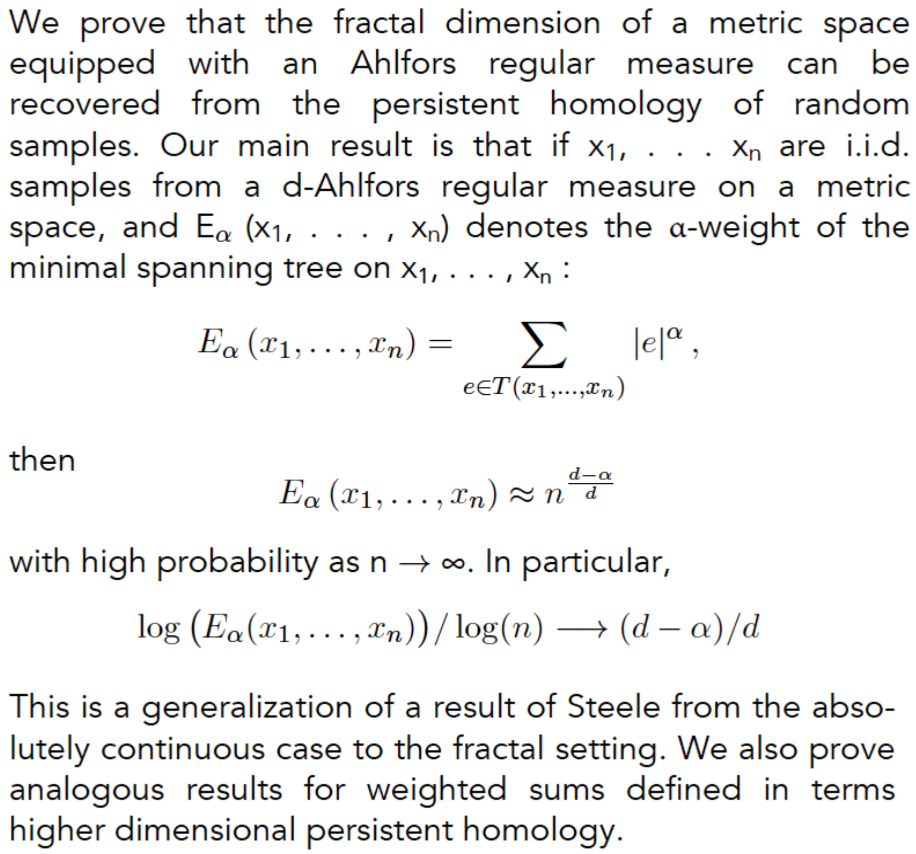
Omer Bobrowski	Homological Percolation: the Formation of Giant Cycles	In probability theory and statistical physics, the field of percolation studies the formation of "giant" (possibly infinite) connected components in various random graphs. In this talk, we will discuss a higher dimensional analogue of this phenomenon. Suppose that we have a random point cloud generated over a manifold M. Constructing a filtration (e.g. Cech, Rips), and computing its persistent homology, we can divide the resulting cycles into two classes. By "giant" cycles we refer to those cycles in the random complex that are mapped to nontrivial cycles in M. The rest will be considered "small" cycles (aka "noise"). Similarly, to percolation theory, our goal is to study the phase transitions describing the emerging (birth-time) of these giant cycles. We will discuss the use of these results for differentiating between signal and noise in persistence diagrams. Finally, we will present some unexpected connection to the Euler characteristic curve. Joint work with Primoz Skraba
Jeff Brock	A Case for Model-Driven Discovery and a Geometric Lens on Topological Data	In this two-part talk I will first seek to engage in a conversation about the role of data science in science, and the need for explainable models in data analysis. I will then turn to some open questions in geometric topology that lend themselves to analysis with the new tools of topological data analysis and suggest the value of a geometric perspective.
Tamal Krishna Dey	Generalized Persistence Algorithm for Multi-parameter Persistence Modules	There is no known generalization of the classical matrix reduction-based persistence algorithm for simplicial filtration in a multi-parameter setting. We present for the first time such a generalization. It improves over the Meataxe algorithm commonly used for the purpose by several orders. *Joint work with Cheng Xin*
Herbert Edelsbrunner	Tri-partition of a Polytopal Complex	We prove that for every polytopal complex, K, and every dimension, p, there is a partition of the p-cells into a maximal p-tree, a maximal p-cotree, and the remaining p-cells defining the p-th homology of K. Given a monotonic order of the cells, this tri-partition is unique and can be computed by matrix reduction. Collecting the sets over all monotonic orders, we get matroids over the set of p-cells. As an application, we consider the manipulation of the hole structure in geometric shapes, using the tri-partition to facilitate the opening and closing of holes in subcomplexes of K. In a concrete application, we let K be the Delaunay triangulation of a finite set, and we extract a partial order on the filtration induced by the radius function, whose cuts define the subcomplexes that can be constructed with this method. *Joint work with Katharina Oelsboeck*

Daniela Egas Santander	Topology and Neuroscience	I will present some of the applications of topology and topological data analysis to neuroscience through an exploration of the collaboration between the applied topology group at EPFL and the Blue Brain Project. In particular, I will describe how we are using topology to try to understand the relationship between structural data and data obtained from voltage-sensitive dye imaging techniques.
Tingran Gao	Multi-Frequency Angular Synchronization	We propose a novel formulation for phase synchronization -- the statistical problem of jointly estimating alignment angles from noisy pairwise comparisons -- as a nonconvex optimization problem that enforces consistency of the alignment angles across multiple irreducible representations of the unitary group (in this case, frequency channels). This problem is prototypical for the class averaging algorithm in cryo-EM image analysis. Inspired by harmonic retrieval in signal processing, we develop a simple yet efficient two-stage algorithm that leverages the multi-frequency information. We demonstrate in theory and practice that the proposed algorithm significantly outperforms state-of-the-art phase synchronization algorithms, at a mild computational cost incurred by using the extra frequency channels. We also extend our algorithmic framework to general synchronization problems over compact Lie groups.
Clément Levrard	Estimation and Approximations of Distance Functions (for geometric inference)	Estimating a shape embedded in an Euclidean space boils down to estimating the distance to this shape. This basic fact allows to cast many geometric inference problems into the theoretical framework of distance estimation from a point cloud, and is at the core of robust geometric inference procedures based on relaxations of the classical distance function (for instance the distance to measure). This talk will summarize some recent theoretical results on what precision can be expected for the estimation of these distances functions from a possibly corrupted point cloud. I will also expose some procedures that allows to reduce the complexity of the usual distance estimators, in several noise settings.

Mauro Maggioni	Statistical Learning & Dynamical Systems: Exploiting Hidden Low-dimensional Structures	Inferring the laws of motion of physical systems from observations is a fundamental challenge. Different tools have been brought to bear in different scenarios, with statistical and machine learning techniques becoming more prominent and useful with the abundance of data. Many challenges still remain. In this talk we discuss two examples of geometry-based statistical learning techniques for learning approximations to certain classes of high-dimensional dynamical systems. In the 1st scenario, we consider systems that are well-approximated by a stochastic process of diffusion type on a low-dimensional manifold. Neither the process nor the manifold are known, but we have access to a way of sampling initial conditions and a (typically expensive) simulator that returns short paths of the stochastic system with those initial conditions. We introduce ATLAS, an estimator that -given the above- outputs a stochastic system near the manifold with good large time accuracy guarantees. *Continued on next page* In the 2nd scenario we consider a system of interacting agents: given only observed trajectories of the system, we are interested in estimating the interaction laws between the agents. We consider both the mean-field limit (i.e. the number of agents going to infinity) and the case of a finite number of agents, with an increasing number of observations. We show that at least in particular cases, where the interaction is governed by an (unknown) function of pairwise distances, the high-dimensionality of the state space of the system does not affect the learning rates. In these cases, we achieve an optimal learning rate for the interaction kernel, equal to that of a one-dimensional regression problem. We exhibit efficient algorithms for constructing our estimator for the interaction kernels, with statistical guarantees, and demonstrate them on various simple examples.

Facundo Mémoli	Stable Persistent Homology for Dynamic Metric Spaces	Characterizing the dynamics of time-evolving data within the framework of topological data analysis (TDA) has been attracting increasingly more attention. Popular instances of time-evolving data include flocking/swarming behaviors in animals and social networks. A natural mathematical model for such collective behaviors is a dynamic point cloud, or more generally a dynamic metric space (DMS). In this paper we extend the Rips filtration stability result for (static) metric spaces to the setting of DMSs. We do this by devising a certain three-parameter "spatiotemporal" filtration of a DMS. Applying the homology functor to this filtration gives rise to multi-dimensional persistence module derived from the DMS. We show that this multidimensional module enjoys stability under a suitable generalization of the Gromov-Hausdorff distance which permits metrizing the collection of all DMSs. On the other hand, it is recognized that, in general, comparing two multidimensional persistence modules leads to intractable computational problems. For the purpose of practical comparison of DMSs, we focus on both the rank invariant or the dimension function of the multidimensional persistence module that is derived from a DMS. We specifically propose to utilize a certain metric for comparing these invariants: In our work this metric is either (1) a certain generalization of the erosion distance by Patel, or (2) a specialized version of the well-known interleaving distance.
Bertrand Michel	Statistical Analysis and Parameter Selection for Mapper	The Mapper algorithm is a method for topological data analysis introduced by Singh, Mémoli and Carlsson. In this work, we study the statistical convergence of the 1-dimensional Mapper to its continuous analogue, the Reeb graph. We show that the Mapper is an optimal estimator of the Reeb graph, which gives, as a byproduct, a method to automatically tune its parameters and compute confidence regions on its topological features, such as its loops and flares. This allows to circumvent the issue of testing a large grid of parameters and keeping the most stable ones in the brute-force setting, which is widely used in visualization, clustering and feature selection with the Mapper. *Joint work with Mathieu Carriere and Steve Oudot*

Ezra Miller	Primary Distance for Multipersistence	When persistent homology is used to summarize data objects, distances between the resulting persistence modules serve as proxies for distances between the data objects themselves. In the presence of more than one parameter, module distances are complicated by the rich algebraic structure of multipersistence. In particular, unboundedness of the set of parameters presents problems with integration, interleaving, and other measures. Primary decomposition and algebraic operations related to it provide canonical (functorial) ways to extract bounded parameter sets, yielding convergence for existing measures that are based on integration. In addition, primary distances isolate from mixtures of multipersistence types pure contributions that would, in many existing measures, otherwise introduce bias when truncation or enforced decay are used without taking into account the algebraic structure.
Jonathan Taylor	Proximal change of measure in selective inference	We are interested in the problem of (conditional) selective inference after solving a (randomized) convex statistical learning program in the form of a penalized o\r constrained loss function. We first describe a change-of-measure formula related to the proximal mapping of the penalty in the convex program, yielding a representation of many conditional sampling problems of interest. The result is model-agnostic in the sense that users may provide their own statistical model for inference, we simply provide the modification of each distribution in the model after the selection. We describe some of the geometric structure in the Jacobian appearing in the change of measure, drawing connections to curvature measures appearing in Weyl-Steiner volume-of-tubes formulae. This Jacobian is necessary for problems in which the convex penalty is not polyhedral, with the prototypical example being the group LASSO.
Sayan Mukherjee	Fiber Bundles in Probabilistic Models	We will present how fiber bundles can be used in probabilistic modeling. Applications in geometric morphometrics will be used as examples, this means we will model shapes and surfaces for evolutionary biology and biomedical applications. We will consider three problems: (1) regression using shapes as covariates, (2) sub-shape or variable selection, and (3) Gaussian process models indexed via fiber bundles. For the first two problems we will use ideas based in integral geometry and for the third we will extend approaches based on diffusions on fiber bundles to a hierarchical Bayesian Gaussian process model.

Takashi Owada	Weak convergence results for topological crackle	The main objective of this work is to study the topological crackle from the viewpoints of Topological Data Analysis. Topological crackle frequently appears in the context of manifold learning, and refers to the layered structure of homological cycles generated by ``noisy" samples, where the support is unbounded. We aim to establish weak convergence results for topological objects, including Betti numbers -- a basic quantifier of cycles, and persistence diagrams -- a point process representation for persistent homology, where each homological cycle is represented by its (birth, death) coordinates. If time allows, I will also discuss the case in which the sample possesses non-trivial dependency structure.
Gennady Samorodnistsky	The Betti Numbers in the Multiparamater Model	We establish central limit theorems for the Betti numbers in the multiparamater Costa-Farber model.
Benjamin Schweinhart	Persistent Homology of Random Geometric Complexes on Fractals	 We prove that the fractal dimension of a metric space equipped with an Ahlfors regular measure can be recovered from the persistent homology of random samples. Our main result is that if $x_1, . . . x_n$ are i.i.d. samples from a d-Ahlfors regular measure on a metric space, and E_α (x_1, . . . , x_n) denotes the α-weight of the minimal spanning tree on x_1, . . . , x_n : $E_\alpha (x_1, \ldots, x_n) = \sum_{e \in T(x_1,\ldots,x_n)} \|e\|^\alpha ,$ then $E_\alpha (x_1, \ldots, x_n) \approx n^{\frac{d-\alpha}{d}}$ with high probability as $n \to \infty$. In particular, $\log \left(E_\alpha(x_1, \ldots, x_n)\right) / \log(n) \longrightarrow (d-\alpha)/d$ This is a generalization of a result of Steele from the absolutely continuous case to the fractal setting. We also prove analogous results for weighted sums defined in terms higher dimensional persistent homology.

Amit Singer	Multi-target Detection and Cryo-EM imaging by Autocorrelation Analysis	We consider the multi-target detection problem of recovering a set of signals that appear multiple times at unknown locations in a noisy measurement. In the low noise regime, one can estimate the signals by first detecting occurrences, then clustering and averaging them. In the high noise regime however, neither detection nor clustering can be performed reliably, so that strategies along these lines are destined to fail. Notwithstanding, using autocorrelation analysis, we show that the impossibility to detect and cluster signal occurrences in the presence of high noise does not necessarily preclude signal estimation. Specifically, to estimate the signals, we derive simple relations between the autocorrelations of the observation and those of the signals. These autocorrelations can be estimated accurately at any noise level given a sufficiently long measurement. To recover the signals from the observed autocorrelations, we solve a system of polynomial equations. We explain how these principles can be applied to image 3-D structures of biological macromolecules using cryo-electron microscopy in extreme noise levels.
Katharine Turner	Injectivity Results Relating to the Persistent Homology Transform and the Euler Characteristic Transform	The persistent homology transform (PHT) and the Euler characteristic transform (ECT) are two topological transforms that can be viewed as topological analogues of the Radon transform. They provide a way of summarising shapes in a topological, yet quantitative, way. They take a shape, viewed as a tame subset M of $\mathbb{R}^d$, and associates to each direction $v \in S^{d-1}$ a shape summary obtained by scanning M in the direction v. These shape summaries are either persistence diagrams or piecewise constant integer valued functions called Euler curves. This talk will explore some recent injectivity results relating to these transforms. An inversion theorem of Schapira implies they are injective on the space of shapes—each shape has a unique transform. By making use of a stratified space structure on the sphere, induced by hyperplane divisions, we can also prove additional uniqueness results in terms of distributions on the space of Euler curves. Finally, we can construct a finite bound of the number of directions required to specify any shape in certain uncountable families of shapes. *Joint work with Justin Curry and Sayan Mukherjee*

Rebecca Willett	Algebraic Variety Models for High-Rank Matrix Completion	The past decade of research on matrix completion has shown it is possible to leverage linear dependencies to impute missing values in a low-rank matrix. However, the corresponding assumption that the data lies in or near a low-dimensional linear subspace is not always met in practice. Extending matrix completion theory and algorithms to exploit low-dimensional nonlinear structure in data will allow missing data imputation in a far richer class of problems. In the low rank matrix completion (LRMC) problem, the low rank assumption means that the columns (or rows) of the matrix to be completed are points on a low-dimensional linear subspace. This work extends this thinking to cases where the columns are points on a low-dimensional nonlinear algebraic variety, a problem we call Low Algebraic Dimension Matrix Completion (LADMC). Matrices whose columns belong to a union of subspaces (UoS) are an important special case. We propose a LADMC algorithm that leverages existing LRMC methods on a tensorized representation of the data. For example, a second-order tensorization representation is formed by taking the outer product of each column with itself, and we consider higher order tensorizations as well. This approach will succeed in many cases where traditional LRMC is guaranteed to fail because the data are low-rank in the tensorized representation but not in the original representation. We also provide a formal mathematical justification for the success of our method. In particular, we show bounds of the rank of these data in the tensorized representation, and we prove sampling requirements to guarantee uniqueness of the solution. Interestingly, the sampling requirements of our LADMC algorithm nearly match the information theoretic lower bounds for matrix completion under a UoS model. We also provide experimental results showing that the new approach significantly outperforms existing state-of-the-art methods for matrix completion in many situations. *Joint work with Greg Ongie, Laura Balzano, Daniel Pimentel-Alarcón, and Robert Nowak*