Topological Data Analysis and Machine Learning Theory

Gunnar Carlsson (Stanford University),
Rick Jardine (University of Western Ontario),
Dmitry Feichtner-Kozlov (University of Bremen),
Dmitriy Morozov (Lawrence Berkeley National Laboratory)

Report contributors: Dominique Attali, Anthony Bak, Mikhail Belkin, Peter Bubenik, Frédéric Chazal, Vin de Silva, Brittany Fasy, Jesse Johnson, Matt Kahle, Gilad Lerman, Facundo Mémoli, Quentin Mérigot, Konstantin Mischaikow, Sayan Mukherjee, Daniel Müllner, Monica Nicolau, Amit Patel, Don Sheehy, Yusu Wang.

October 15–19, 2012

1 Persistent homology

Perhaps the most important idea in applied algebraic topology is persistence. It is a response to the first difficulty that one encounters in attempting to assign topological invariants to statistical data sets: that the topology is not robust and has a sensitive dependence on the length scale at which the data set is being considered. The solution is to calculate the topology (specifically the homology) at all scales simultaneously, and to encode the relationship between the different scales in an algebraic invariant called the persistence diagram. The effective algorithm for doing so was published in 2000 by Edelsbrunner, Letscher and Zomorodian [2]. Topological data analysis would not be possible without this tool.

Since then, persistence has been developed and understood quite extensively. Cohen-Steiner, Edelsbrunner and Harer [3] proved the important (and nontrivial) theorem that the persistence diagram is stable under perturbations of the initial data. Zomorodian and Carlsson [7] studied persistence algebraically, identifying the points of the persistence diagram with indecomposable summands of a module over the polynomial ring $k[t]$, the monomial t representing a change of scale by a fixed increment. Generalising this approach using polynomial rings with two or more variables, they showed that the corresponding situation with two or more independent length scales is in some sense algebraically intractable, with no complete descriptive invariants available. Carlsson and de Silva [1] showed that persistence can be made to work over "non-monotone" parameters, in contrast to "monotone" parameters such as length scale; this is known as zigzag persistence. Bubenik and Scott [4] have described and studied persistence in terms of category theory.

Measured view. In becoming familiar with the literature on persistence, one quickly realises that the existing mathematical foundations are dependent on certain strong finiteness assumptions. For instance, one can do persistence on the sublevelsets of a function on a compact manifold, but it is usually assumed that the function be a Morse function. This ensures that the resulting persistent data is finite: each sublevelset has finite-dimensional homology, and there are only finitely many essentially distinct levels, separated by the critical points. In computational terms, this is a perfectly natural assumption (a computer can only handle finite data), but for some of the theoretical work this assumption is

limiting. This was perhaps first seen in the proof of the stability theorem by Cohen-Steiner et al, which has to navigate around this assumption.

Vin de Silva presented new work [5] with co-authors Chazal, Glisse and Oudot, which addresses these problems through a deep analysis of the structure of so-called "persistence modules": 1-parameter families of vector spaces and maps between them. The basic challenge is to construct the persistence diagram and show that it is stable, without the usual strong finiteness assumptions about the persistence module. They achieve this by establishing an equivalence between persistence diagrams and a certain kind of measure defined on rectangles in the plane. The equivalence is proved under a weaker finiteness condition, called "q-tameness", which is seen to hold quite widely: for any continuous function (not just Morse) on a finite simplicial complex, for the Vietoris–Rips complex on a compact metric space, etc. It turns out that all of the standard results can be proved much more easily, and in greater generality, when one works with these measures. The authors introduce a new notational system, a sort of "quiver calculus", which gives short transparent proofs of the linear algebra lemmas that show up in the persistence literature, by interpreting those results as statements about the indecomposable summands of certain quiver representations.

Statistics of diagrams. An important workshop theme emerged during John Harer's talk: the need to introduce statistical techniques to topological data analysis. The overarching idea is that the input to most analysis techniques is generated stochastically. Therefore, it is interesting to talk not only about a single persistence diagram, but about an entire collection of them, so that we can try to calculate means, variances and apply statistical inference techniques.

Harer described his work [8] with Mileyko and Mukherjee showing that the space of persistence diagrams allows for the definition of probability measures, which support expectations and variances, among other properties. The authors showed that the space of persistence diagrams with the Wasserstein metric is complete and separable, and described when it is compact. Harer also sketched an algorithm [9] to compute Fréchet means in this space.

Inference using landscapes. Peter Bubenik gave an alternative approach [10] to this problem. He mapped persistence diagrams (also known as barcodes) to certain functions $\lambda : \mathbb{N} \times \mathbb{R} \to \mathbb{R}$, called persistence landscapes. Let B be a barcode. Call half the length of an interval its radius. Then $\lambda_k(t)$ is the largest radius such that the interval of radius r centered at t is a subinterval of k intervals in B. This space inherits nice structure from $L^p(\mathbb{N} \times \mathbb{R})$. In L^2, the Fréchet means and variances are the pointwise means and integral of the pointwise variances, respectively. One has a strong law of large numbers and a central limit theorem. One can apply statistical inference using the permutation. Bubenik applied these ideas to synthetic data drawn from a torus and a sphere and to brain MRI data. In the latter he showed that the persistence landscapes of the triangulation of the outer surface filtered by angle defect seemed to discriminate between high-functioning autistic subjects and controls.

Computational advances. The rapid growth in the range of applications of algebraic topology suggests the need for efficient algorithms for computing homology groups, persistent homology, and induced maps on homology. There are a variety of strategies that have been adopted. The most fundamental is to view Smith diagonalization as a purely algebraic problem and then to seek an optimal algorithm. The worst case analysis of such algorithms suggests a supercubical complexity with respect to the size of the complex, which is prohibitive for large datasets. An alternative strategy is to develop efficient algorithms for restricted problems, for example by restricting the dimension of the complex or restricting the computation to that of Betti numbers. Konstantin Mischaikow described an approach [11] that uses discrete Morse theory to pre-process the data, thereby producing a smaller complex on which an algebraic Smith normal form computation can be performed.

Mischaikow's talk centered around three topics:

1. A review of the notion of complex as introduced by Tucker and Lefschetz where the focus is on incidence numbers that relate neighboring cells. In the contexts of data analysis and computational dynamics which motivate this work, this local information is often natural to the input whereas the associated boundary operator must be derived from this information. The explicit construction of this boundary operator, which can be costly for large complexes, is unnecessary in the approach.

2. Algorithms for efficiently computing Morse complexes which are based on the co-reduction algorithm.

3. The use of Morse complexes to efficiently compute the induced maps on homology.

Generalization: Étalage. Persistent homology quantifies topological information in data. Very importantly, this quantification is stable to measurable perturbations of the data. With this in mind, Amit Patel presented abstractions of various notions in the theory of persistent homology to multi-parameter families of spaces.

A fundamental notion is the persistent homology group. Let $f : \mathbb{X} \to \mathbb{R}$ be a continuous map from a topological space to the real line. The map f defines a 1-parameter family of spaces $\mathbb{X}_t = f^{-1}(t)$. The *persistent homology group* over an interval $(a, b) \subseteq \mathbb{R}$ is the intersection of the images of the two homomorphisms $i : \mathsf{H}_*(f^{-1}(a)) \to \mathsf{H}_*(f^{-1}[a, b])$ and $j : \mathsf{H}_*(f^{-1}(b)) \to \mathsf{H}_*(f^{-1}[a, b])$ induced by inclusion of spaces. Roughly speaking, the persistent homology group is the homology that is common to all fibers $\mathbb{X}_t$ over the interval (a, b). Now let $g : \mathbb{X} \to \mathbb{M}$ be a map to an oriented m-manifold. This map defines an m-parameter family of spaces $\mathbb{X}_p = g^{-1}(p)$. One would like an abstraction of the notion of the persistent homology group to path-connected open sets $U \subseteq \mathbb{M}$. Patel introduced the well group, which serves this purpose. The *well group* is the image of a homomorphism $\Phi : \mathsf{H}^{\mathrm{c}}_{*+m}(f^{-1}(U)) \to \mathsf{H}_*(f^{-1}(U))$ from the homology of $f^{-1}(U)$ with closed support to the homology of $f^{-1}(U)$ with compact support shifted down by m dimensions. This homomorphism is the cap product with the pullback of the orientation on $\mathbb{M}$. The well group is the homology that is common to all fibers above U, and, furthermore, it is stable to homotopic perturbations of the map g.

The theory of persistent homology assembles the local information of the persistent homology groups into a global structure called the *persistence diagram*. Patel introduced the *étalage* of $\mathbb{M}$ which abstracts the notion of a persistence diagram to the higher dimensional setting. An étalage of $\mathbb{M}$ is a Hausdorff topological space $\mathbb{E}$ along with a continuous map $\pi : \mathbb{E} \to \mathbb{M}$ that is locally a homeomorphism. In addition, assigned to each connected component of $\mathbb{E}$ is an integer. As with persistence diagrams, one can read off the rank of the stable homology of each fiber $\mathbb{X}_p$ by summing the integers assigned to each point in $\pi^{-1}(p)$.

Persistence stability for geometric complexes. The classical theory of persistent homology, restricted to tame functions and filtrations of finite simplicial complexes, does not directly address certain questions in topological data analysis: e.g., multiscale homology inference for metric spaces or scalar fields analysis on discrete data. To overcome this issue, Frédéric Chazal and collaborators extended and generalized persistent homology and its stability results. As an application they have proven the robustness of the persistent homology of various families of geometric filtered complexes built on top of compact metric spaces or spaces endowed with a similarity measure.

Theorem 1 ([6]). *Let X be a pre-compact metric space. Then the Vietoris–Rips and Čech filtrations built on top of X induce, at the homology level, persistence modules that are q-tame. In particular, they have well-defined persistence diagrams. Moreover, if X and Y are two compact metric spaces then the bottleneck distance between the persistence diagrams of the Vietoris–Rips (resp. Čech) filtrations built on top of X and Y is upper bounded by the Gromov–Hausdorff distance between X and Y.*

This result is an ingredient for mathematically well-founded statistical developments for topological data analysis using persistence theory.

Persistent homology and metric geometry. Facundo Mémoli presented ongoing work on importing constructions from metric geometry into persistent topology. His aim is to define a notion of distance d_F between filtered finite spaces compatible with the standard stability result for persistence diagrams of filtrations. More precisely, $d_F(X, Y)$ is defined as the infimal $\varepsilon > 0$ for which one can find a finite set Z and surjective maps $\varphi_X : Z \to X$ and $\varphi_Y : Z \to Y$ such that the L^∞ norm of the difference of the pullback filtrations $\varphi_X^* F_X, \varphi_Y^* F_Y : \mathrm{pow}(Z) \to \mathbb{R}$ is bounded by ε.

Now, one has a generalization of the combinatorial stability theorem of [21] that does not assume that the two filtrations are defined on the same space. Below, d_B is the bottleneck distance between persistence diagrams.

Theorem 2 ([23]). *For all (X, F_X) and (Y, F_Y), $d_B\big(D_*(F_X), D_*(F_Y)\big) \leq d_F(X, Y)$.*

Mémoli also described different ways in which metric spaces and metric measure spaces induce filtrations, and how these induced filtrations are stable with respect to suitable notions of distance. Each of these "ways" gives rise to a *filtration functor*. In this manner, a measure-dependent notion of Vietoris-Rips filtration arises, which is quantitatively stable in a precise sense.

To be clear, given a finite mm-space (X, d_X, μ_X), the *weighted* Vietoris-Rips filtration $F_X^{\omega R} : \mathrm{pow}(X) \to \mathbb{R}$ is

given for $p \geq 1$ by

$$\sigma \mapsto \left(\sum_{x,x' \in \sigma} \big(d_X(x,x')\big)^p \mu_X(x)\mu_X(x') \right)^{1/p}.$$

The hope is that with such constructions one would be able to capture topological features of a given dataset in a manner which is robust to noise or outliers. In order to express stability, in the case of metric spaces the natural notion of distance is the Gromov-Hausdorff distance, whereas in the case of mm-spaces, the notion is the Gromov-Wasserstein distance $d_{GW,\infty}$ [20]. From Theorem 2 above one obtains:

Corollary 1. *[23] For all finite mm-spaces X and Y one has $d_B\big(D_*(F_X^{\omega R}), D_*(F_Y^{\omega R})\big) \leq d_{GW,\infty}(X,Y)$.*

Mémoli also revisited a theme, which was discussed in other talks in the workshop, namely, the issue of pinning down a suitable notion of what it might mean to do statistics on persistence diagrams. He described a measured construction that can be regarded as a step that takes place *before* the computation or definition of the Fréchet mean of a collection of persistence diagrams.

2 Topological connections

Topological statistical mechanics. Configuration spaces of points are well-studied in several branches of mathematics, including algebraic topology, geometric group theory, and combinatorics. Give the particles thickness, and you have what physicists might describe as phase space for a hard spheres gas. When the points are points, the topology of the configuration space is well understood. But hardly anything is known when points have thickness. The changes in the topology as the thickness varies could be thought of as topological phase transitions.

In joint work [12] with Baryshnikov and Bubenik, Matthew Kahle has started to develop a Morse theory for these configuration spaces. In particular, they have proved a theorem that "critical points", where the topology changes, correspond to mechanically balanced configurations of spheres.

With a similar point of view, Carlsson, Gorham, Mason, and Kahle [13] implemented a computational approach to study these spaces. They find complicated behavior, even for a small number of particles. With only five disks in a square, it seems that the topology of the configuration space changes a few dozen times as the radius of the particles varies.

Finally, MacPherson and Kahle have a work in progress where they consider the asymptotic behavior of Betti numbers for n disks in an infinite strip as $n \to \infty$. They find that that there is a regime where the Betti numbers grow polynomially, and a regime where they grow exponentially. Understanding these kinds of asymptotics for any bounded region seems to be an attractive and wide-open problem.

Topological dimensionality reduction. Data in high dimensional spaces is ubiquitous across a variety of domains. A major problem in data analysis is how to create effective schemes to reduce the dimensionality while maintaining or improving the ability to do geometric and statistical inference. Anthony Bak showed how ideas from topological data analysis can guide intelligent dimension reduction choices.

Many data analysis problems consist of a collection of objects, each of which has associated features. For the case that the features are real-valued we organize this information into a matrix $S = (s_{ij})$, where s_{ij} is the value of the j^{th} feature on the i^{th} object. The goal is to understand the objects, or row space of this matrix. As a preprocessing step, we analyze the geometry of the column space, or "dual space", with the goal of reducing the number of columns.

Bak applied these ideas in two very different situations:

1. A simulation in the spirit of "evasion" and sensor network problems. We have a series of sensors in the plane observing a different set of moving particles. Each sensor is "dumb" in that it only records aggregate information on all of the particles within a certain radius. In simulations Bak laid the sensors out in a series of circles with varying density. The goal was to reduce the number of sensors while maintaining the topological coverage so that one could, for example, detect when an object passes in or out of a circle.

2. Real world microarray data consisting of gene activation levels for E. coli from a collection of 600 experiments. E. coli genes are physically laid out in a circle (unlike the human genome which is laid out in a line on each chromosome) and genes near each other have a tendency to activate together in what is called an "operon" group. The goal is to reconstruct the circular ordering of the genome from the microarray data.

The central idea in both cases was to build a metric space using correlation distance between sensors and to remove sensors until the topology, as measured with zigzag persistence, changed. In both examples he reported a dramatic dimensionality reduction, going from hundreds of features to 50 in the first case and only five in the case of E coli. For E. coli the ability to reconstruct the ordering improved over both using all the experiments and using PCA as the dimension reduction step.

Novel techniques for clustering. The clustering problem in data analysis is to decide how to split a large data set into a number of smaller sets based on the geometry of the point cloud. The goal is for points in each subset to be closer to each other than to points in the other subsets. Jesse Johnson discussed a new approach [14] to this problem based on ideas from three-dimensional geometric topology called thin position. In the pure topology setting, these techniques are very effective at finding two dimensional surfaces that efficiently cut three-dimensional spaces into smaller pieces. Because of the close link between topology and geometry in dimension three, thin position also tends to find geometrically efficient partitions of these spaces, so it is very natural to apply them to finding efficient geometric partitions of data sets.

This topic was also a natural complement to the workshop's theme of homological techniques because in the pure topology setting, thin position tends to find structure that homology misses, while missing the structure that homology measures. To get a complete understanding of a three-dimensional manifold, one must understand both its homology groups and its thin position structure. It appears that the same is true for data analysis: To best understand a large, high dimensional data set, one should consider information that comes both from persistence homology and from the algorithms based on thin position.

Biological applications. The past decade has witnessed developments in the field of biology that have brought about profound changes in understanding the dynamic of disease and of biological systems in general. New technology has given biologists an unprecedented wealth of information, but it has generated data that is hard to analyze mathematically, thereby making its biological interpretation difficult. These challenges, stemming in part from the very high dimensionality of the data, have given rise to a myriad novel exciting mathematical problems and have provided an impetus to modify and adapt traditional mathematics tools, as well as develop novel techniques to tackle the data analysis problems raised in biology.

Monica Nicolau discussed data transformations and topological methods for solving biology-driven problems. The general approach of her work was to address some of the computational challenges of these large data types by combining data transformations and topological methods. Through the definition of high-dimensional mathematical models for different biological states, for example healthy vs. disease, or various developmental stages of cells, data can be transformed to mod out all characteristics but those relevant or statistically significant to the problem under study. Once data has been transformed, analysis of significance involves analysis of the shape of the data. Nicolau described how adaptation of topological methods in the context of discrete point clouds has proved to be a powerful tool for identifying statistical significance as well as providing methods for visualizing data.

Scale selection. One of the keywords in topological data analysis, besides "shape", is "scale". This is not an obvious fact, given that a great part of algebraic topology deals with utmost flexibility in the sense of homotopy invariance and could thus be considered scale-free. On the other hand, it exemplifies what we must give up as soon as ideal shapes are approximated by point clouds, and how topology, at the interface with applied statistics, can use its inherent flexibility and global point of view to find out at which scale features are present in a data set and to also detect and describe multi-scale phenomena.

Daniel Müllner focused on a situation where several scale choices are made on overlapping parts of a data set. This happens at the core of the "Mapper" algorithm, a tool for visualization, exploration and data analysis, which has been successful both academically and commercially. The original Mapper algorithm divides a data set into overlapping slices and chooses a scale independently for each, in order to cluster the data within each slice. Müllner demonstrated how this can lead to inconsistencies and false positives of topological features and in general makes it hard to determine which representations of a data set among many different possible outputs are appropriate. However, one can leverage spatial coherence and link scale choices for neighboring regions together in a "scale graph". By choosing optimal paths through the scale graph, those shortcomings can be resolved successfully. Müllner and co-authors developed heuristics for the scale graph method with the goal of making it work well for a broad range of data sets without optimizing too much for a particular context.

One problem in this process, which is still a big challenge in all of applied topology, is that noise in sampled data quickly smudges or destroys features. As an independent idea, which works very well together with the scale graph, Müllner proposed a new way of looking at dendrograms from hierarchical clustering. The basic idea is that stability intervals for the number of clusters in a dendrogram are not necessarily disjoint, but it is beneficial to consider more than one choice as feasible for a given scale if two clusterings differ for example only by inclusion or exclusion of noise points. This leads to a new rating of stability intervals which is conservative in the sense that it changes very little for low noise but improves the clustering considerably for higher noise levels. Since this new approach is not restricted to the Mapper context of multiple scale choices, it also has potential outside the application it was invented for, and it seems worthwhile to explore where it can improve clustering choices in other situations.

3 Geometric connections

Shape reconstruction. Dominique Attali reported on reconstructing shape of data points distributed in low-dimensional subspaces of a high-dimensional ambient space. This problem arises in many fields, including computer graphics and machine learning. Typical shape reconstruction methods start by building the Delaunay complex and then select a set of simplices such as the α-complex. Such approaches work well for point clouds in two- and three-dimensional spaces, which have Delaunay triangulations of affordable size. But, as the dimension of the ambient space increases, the size of the Delaunay triangulation explodes and other strategies must be found. If the data points lie on a low-dimensional submanifold, it seems reasonable to ask that the result of the reconstruction depends only upon the intrinsic dimension of the data. This motivated de Silva [19] to introduce *weak Delaunay complexes* and Boissonnat and Ghosh [18] to define *tangential Delaunay complexes*. For medium dimensions, Boissonnat et al. [17] have modified the data structure representing the Delaunay complex and are able to manage complexes of reasonable size up to dimension 6 in practice. In particular, they avoid the explicit representation of all Delaunay simplices by storing only edges in what they call the *Delaunay graph*, an idea close to that of using Rips complexes that Attali described.

Given a point set P and a scale parameter α, the *Vietoris-Rips complex* is the simplicial complex whose simplices are subsets of points in P with diameter at most 2α. Rips complexes are examples of *flag complexes*, and as such enjoy the property that a subset of P belongs to the complex if and only if all its edges belong to the complex. In other words, Rips complexes are completely determined by the graph of their edges. This compressed form of storage makes Rips complexes very appealing for computations, at least in high dimensions.

Attali (with André Lieutier and David Salinas) obtained two results. First, they established that Rips complexes can capture the homotopy type when distances are measured using the ℓ_∞ norm [15]. Unfortunately, the sampling condition was not as weak as they were hoping for, especially as the dimension of the ambiant space increases. Encouraged by experiments that were indicating this result should also hold when measuring distances using the Euclidean norm instead of the ℓ_∞ norm, they revisited the question and found conditions under which Rips complexes reflect the homotopy type of shapes when measuring distances using the Euclidean norm [16].

Mesh generation. A common approach to Topological Data Analysis (TDA) starts with a point cloud, proceeds to a function induced by the points, builds a simplicial complex, and then analyzes the persistent homology of the sublevel sets of that function as a filtration on the complex. The most common case of this is to explore the distance function to a set of points in Euclidean space. If we generalize to consider Lipschitz smooth functions on low-dimensional Euclidean space, there are many other choices of functions that apply. From this perspective, it is natural to consider the well-established field of mesh generation; it aims to efficiently build small simplicial complexes that give good approximations to Lipschitz functions. Don Sheehy showed how to apply results from mesh generation, some standard and some new, to give guaranteed approximate persistence diagrams for a wide class of functions.

His main results [24, 25, 26, 27] state that given a point set and any t-Lipschitz function f bounded from below by c times the second-nearest neighbor distance function to the point set, there exists a filtered mesh that has approximately the same persistence diagram, i.e. the diagrams drawn on the log-scale are close in bottleneck distance. Moreover, this mesh is independent of the function. It depends only on the point set, the constants c, t, and the desired approximation guarantee. Specifically, for a so-called ε-refined mesh, the resulting persistence diagram will be a $(1 + ct\varepsilon/(1 - \varepsilon))$-approximation to the persistence diagram of f itself. Moreover, for most reasonable point sets, one can guarantee that the mesh will have only linear size, though the constant factors depend exponentially on the ambient dimension.

The number of vertices will depend on the input and the amount of extra refinement. The dependence is a simple exponential in $1/\varepsilon$. The number of simplices incident to any vertex will be $2^{O(d^2)}$. It is not clear if it is possible to fill

space with significantly fewer simplices per vertex. Volume arguments immediately imply a $2^{\Omega(d \log d)}$ lower bound for quality meshes, ruling out the possibility of a simple exponential dependence on the dimension.

Distance to measure. The notion of distance to a measure was introduced [28] in order to extend existing geometric and topological inference results from the usual Hausdorff sampling condition to a more probabilistic model of noise. This function can be rewritten as a minimum of a finite number of quadratic functions, one per isobarycenter of k distinct points of the point set, thus allowing to compute the topology of its sublevel sets using weighted alpha-complexes, weighted Rips complexes, etc.

However, as the number of isobarycenters grows exponentially with the number of points, it is necessary in practice to approximate this function by another function that can be written as a minimum of quadratic functions. A natural problem is then the following: given a target error ε, determine the minimum number of quadratic functions needed to approximate the distance to the measure with error ε. Quentin Mérigot presented recent probabilistic lower bounds on this number.

4 Machine Learning.

Learning mixtures of Gaussians. In recent years there has been an increase of interest in using algebraic-geometric methods to analyze data by recovering structure in probability distributions, in the field of algebraic statistics. This development has been somewhat parallel to using algebraic-topological methods to understand the shape of the data through recovering the homology groups or other topological and geometric invariants of the data.

Mikhail Belkin discussed recent work [29] on using real algebraic geometry to recover the parameters of mixtures of high-dimensional Gaussian distributions as well as other parametric families. Unlike most of the existing work in algebraic statistics, these parametric families are typically not exponential families. Specifically, his main result is that a mixture of Gaussians with a fixed number of components can be learned using the number of samples and operations polynomial in the dimension of the space and other relevant parameters. Moreover, a version of this statement for fixed dimension holds for a much broader class of distributions, called "polynomial families", i.e. families whose moments are polynomial functions of the parameters.

The overarching point of Belkin's talk was that there may be interesting connections between the fields of algebraic statistics, where the object of study is the space of parameters, and topological data analysis, where the geometry of a space is analyzed directly.

Modes in the mixtures of Gaussians. Brittany Fasy presented a recent result [30] on Gaussian mixtures. The mixture analyzed was the sum of $n+1$ identical isotropic Gaussians, where each Gaussian is centered at the vertex of a regular n-simplex. All critical points of this mixture are located on one-dimensional lines (axes) connecting barycenters of complementary faces of the simplex. Fixing the width of the Gaussians and varying the diameter of the simplex from zero to infinity by increasing a parameter called the scale factor, gives the window of scale factors for which the Gaussian mixture has more modes, or local maxima, than components. Using the one-dimensional axes containing the critical points, the interval of scale factors can be computed. Furthermore, the extra mode created is subtle, but becomes more pronounced as the dimension increases.

A natural open question in the area is: Restricting our attention to some class of kernels (for example, unimodal continuous kernels), which kernel observes the most (or the fewest) ghost modes?

Robus PCA. Consider a dataset of vector-valued observations that consists of a modest number of noisy inliers, which are explained well by a low-dimensional subspace, along with a large number of outliers, which have no linear structure. Lerman, McCoy, Tropp and Zhang [32] have suggested a convex optimization problem that can reliably fit a low-dimensional model to this type of data. They first minimize the function $F(\mathbf{Q}) := \sum_{i=1}^{N} \|\mathbf{Q}\mathbf{x}_i\|$ over $\{\mathbf{Q} \in \mathbb{R}^{D\times D} : \mathbf{Q} = \mathbf{Q}^T, \mathrm{tr}(\mathbf{Q}) = D - d \text{ and } \mathbf{Q} \preccurlyeq \mathbf{I}\}$. The subspace is then defined as the span of the bottom d eigenvectors of this minimizer. They referred to this subspace recovery optimization as the REAPER optimization problem.

When the inliers are contained in a low-dimensional subspace they provided a rigorous theory describing when this optimization can recover the subspace exactly. The theory (based on an earlier work of Zhang and Lerman [33]) establishes exact recovery under some combinatorial conditions (which ask for sufficient spread of inliers throughout the underlying subspace and non-concentration of outliers per directions as well as some control on the magnitude

of outliers). It also shows that under some probabilistic settings (e.g., Gaussian distributions of inliers and outliers) these combinatorial conditions, and thus the subspace recovery, are guaranteed. An example of such a probabilistic guarantee is formulated as follows, where $\mathbf{P}_L$ denotes the orthogonal projector onto the subspace L and $\mathcal{N}(\mathbf{0}, \mathbf{V})$ denotes a normal distribution with mean $\mathbf{0}$ and covariance matrix $\mathbf{V}$.

Theorem 3. *Fix a number $\beta > 0$, and assume that $1 \leq d \leq (D-1)/2$. Let L_* be an arbitrary d-dimensional subspace of $\mathbb{R}^D$, and draw N_{in} inliers i.i.d. $\mathcal{N}(\mathbf{0},\ (\sigma_{\text{in}}^2/d)\,\mathbf{P}_{L_*})$ and N_{out} outliers i.i.d. $\mathcal{N}(\mathbf{0},\ (\sigma_{\text{out}}^2/D)\,\mathbf{I}_D)$. Let C_1, C_2 and C_3 be positive universal constants. If the sampling sizes and the variances satisfy the relation*

$$\frac{N_{\text{in}}}{d} \geq \mathrm{C}_1 + \mathrm{C}_2\,\beta + \mathrm{C}_3 \cdot \frac{\sigma_{\text{out}}}{\sigma_{\text{in}}} \cdot \left(\frac{N_{\text{out}}}{D} + 1 + 4\beta \right), \tag{1}$$

then L_* is the unique solution to the REAPER optimization problem except with probability $4\,\mathrm{e}^{-\beta d}$.

Furthermore, Lerman, McCoy, Tropp and Zhang [32] presented an efficient iterative algorithm for solving this optimization problem, which converges linearly and its computational cost is comparable to that of the non-truncated SVD. Coudron and Lerman [31] established probabilistic convergence rates for the REAPER optimization problem (of an i.i.d. sampled data from a continuous setting); it is of the same order as that of full PCA. These imply some nontrivial robustness to noise (similar to the one of PCA) as well as sample complexity estimates for robust PCA (like those of PCA).

Towards understanding the Gaussian-weighted graph Laplacian. The Gaussian-weighted graph Laplacian, as a special form of graph Laplacians with general weights, has been a popular empirical operator for data analysis applications, including semi-supervised learning, clustering, and denoising. There have been various studies of the properties and behaviors of this empirical operator; most notably, its convergence behavior as the number of points sampled from a hidden manifold goes to infinity. Yusu Wang presented two new results on the theoretical properties of the Gaussian-weighted graph Laplacian. The first one is about its behavior as the input points are sampled from a *singular manifold*; while previous theoretical study of the Gaussian-weighted graph Laplacian typically assumes that the hidden domain is a compact smooth manifold. A singular manifold can consist of a collection of potentially intersecting manifolds with boundaries; it represents one step towards modeling more complex hidden domains. The second result is about the stability of the Gaussian-weighted Graph Laplacian if the hidden manifold has certain small perturbation. The goal is to understand how the spectrum of Gaussian-weighted Graph Laplacian changes with respect to perturbations of the domain.

Both of these problems are connected to topological data analysis. In particular, the work on singular manifolds is closely related to learning stratified spaces from point samples, although approached from a different direction from current topological methods for stratification inference. It would be interesting to investigate how these two somewhat complimentary lines of work can be combined to produce efficient algorithms for stratification learning. It is likely that one should focus on constraint families of stratified spaces so that theoretical guarantees can be obtained when they are inferred from point samples.

For the work on the stability of Gaussian-weighted graph Laplacian, a key question is a general perturbation model that allows small topological changes of the underlying domain. The presented perturbation model allows certain topological changes, but is still rather special. It would be interesting study under what generalized perturbation models one can still obtain bounded perturbations in the spectrum of graph Laplacian.

Manifold learning via Lie groups. The Cheeger inequality [37, 36] is a classic result that relates the isoperimetric constant of a manifold (with or without boundary) to the spectral gap of the Laplace-Beltrami operator. An analog of the manifold result was also found to hold on graphs [35, 34, 47] and is a prominent result in spectral graph theory. Given a graph G with vertex set V, the Cheeger number is the following isoperimetric constant

$$h := \min_{\emptyset \subsetneq S \subsetneq V} \frac{|\delta S|}{\min\{|S|, |\overline{S}|\}}$$

where δS is the set of edges connecting a vertex in S with a vertex in $\overline{S} = V \setminus S$. The Cheeger inequality on the graph relates the Cheeger number h to the algebraic connectivity λ [42] which is the the second eigenvalue of the graph Laplacian. It states that

$$2h \geq \lambda \geq \frac{h^2}{2 \max_{v \in V} d_v}$$

where d_v is the number of edges connected to vertex v (also called the degree of the vertex). For more background on the Cheeger inequality see [39].

A key motivation for studying the Cheeger inequality has been understanding expander graphs [44] — sparse graphs with strong connectivity properties. The edge expansion of a graph is the Cheeger number in these studies and expanders are families of regular graphs $\mathcal{G}$ of increasing size with the property $h(G) > \varepsilon$ for some fixed $\varepsilon > 0$ and all $G \in \mathcal{G}$. A generalization of the Cheeger number to higher dimensions on simplicial complexes, based on ideas in [45, 46], was defined and expansion properties studied in [40] via cochain complexes. In addition, it has long been known [41] that the graph Laplacian generalizes to higher dimensions on simplicial complexes. In particular, one can generalize the notion of algebraic connectivity to higher dimensions using the cochain complex and relate an eigenvalue of the k-dimensional Laplacian to the k-dimensional Cheeger number. This raises the question of whether the Cheeger inequality has a higher-dimensional analog.

Sayan Mukherjee examined the combinatorial Laplacian which is derived from a chain complex and a cochain complex. First, the negative result: for the cochain complex a natural Cheeger inequality does not hold. For an m-dimensional simplicial complex we denote λ^{m-1} as the analog of the spectral gap for dimension $m-1$ on the cochain complex, and we denote h^{m-1} as the $(m-1)$-dimensional coboundary Cheeger number. In addition, let S_k be the set of k-dimensional simplexes, and for any $s \in S_k$ let d_s be the number of $(k+1)$-simplices incident to s. The following result implies that there exists no Cheeger inequality of the following form for the cochain complex. Specifically, there are no constants p_1, p_2, C such that either of the inequalities

$$C(h^{m-1})^{p_1} \geq \lambda^{m-1} \qquad \text{or} \qquad \lambda^{m-1} \geq \frac{C(h^{m-1})^{p_2}}{\max_{s \in S_{m-1}} d_s}$$

hold in general for an m-dimensional simplicial complex X with $m > 1$. The case of h^0 and λ^0 with $p_1 = 1$ and $p_2 = 2$ reduces to the Cheeger inequality on the graph and the Cheeger inequality holds.

For the chain complex Mukherjee and co-authors obtain a positive result: there is a direct analogue for the Cheeger inequality in certain well-behaved cases. Whereas the cochain complex is defined using the coboundary map, the chain complex is defined using the boundary map. Denote γ_m as the analog of the spectral gap for dimension m on the chain complex and h_m as the m-dimensional Cheeger number defined using the boundary map. If the m-dimensional simplicial complex X is an orientable pseudomanifold or satisfies certain more general conditions, then

$$h_m \geq \gamma_m \geq \frac{h_m^2}{2(m+1)}.$$

This inequality can be considered a discrete analog of the Cheeger inequality for manifolds with Dirichlet boundary condition [37, 36].

References

[1] G. Carlsson and V. de Silva. Zigzag Persistence. *Found. Comput. Math.* (2010).

[2] H. Edelsbrunner, D. Letscher and A. Zomorodian. Topological persistence and simplification. *Discrete Comput. Geom.* **28** (2002), 511-533.

[3] David Cohen-Steiner, Herbert Edelsbrunner, and John Harer. Stability of Persistence Diagrams. *Discrete and Computational Geometry,* **37**:103–120, 2007.

[4] P. Bubenik and J. A. Scott. Categorification of persistent homology. Preprint, arXiv:1205.3669, 2012.

[5] V. de Silva, F. Chazal, M. Glisse, and S. Oudot. The structure and stability of persistence modules. Manuscript, 2012.

[6] V. de Silva, F. Chazal, and S. Oudot. Persistence stability for geometric complexes. Manuscript, 2012.

[7] A. Zomorodian and G. Carlsson. Computing Persistent Homology. *Discrete and Computational Geometry,* **33**(2):249–274, 2005.

[8] Yuriy Mileyko, Sayan Mukherjee, and John Harer. Probability measures on the space of persistence diagrams. Inverse Problems, **27**, 2011.

[9] Katharine Turner, Yuriy Mileyko, Sayan Mukherjee, and John Harer. Fréchet Means for Distributions of Persistence Diagrams. Preprint, 2012.

[10] Peter Bubenik. Statistical topology using persistence landscapes. Manuscript, arXiv:1207.6437, 2012.

[11] S. Harker, K. Mischaikow, M. Mrozek, and V. Nanda. Discrete Morse Theoretic Algorithms for Computing Homology of Complexes and Maps. Preprint, 2012.

[12] Yuliy Baryshnikov, Peter Bubenik, and Matthew Kahle. Min-type Morse theory for configuration spaces of hard spheres. Preprint, 2011.

[13] Gunnar Carlsson, Jackson Gorham, Jeremy Mason, and Matthew Kahle. Computational topology for configuration spaces of hard disks. Physical Review E, 2012.

[14] Jesse Johnson. Topological graph clustering with thin position. Manuscript, arXiv:1206.0771, 2012.

[15] D. Attali and A. Lieutier. Reconstructing shapes with guarantees by unions of convex sets. In *Proc. 26th Ann. Sympos. Comput. Geom.*, pages 344–353, Snowbird, Utah, June 13-16 2010. [download], [hal-00468610].

[16] D. Attali, A. Lieutier, and D. Salinas. Vietoris-Rips complexes also provide topologically correct reconstructions of sampled shapes. *Computational Geometry: Theory and Applications (CGTA)*, 2012. In Press. [download].

[17] J. Boissonnat, O. Devillers, and S. Hornus. Incremental construction of the Delaunay triangulation and the Delaunay graph in medium dimension. In *Proceedings of the 25th annual symposium on Computational geometry*, pages 208–216. ACM, 2009.

[18] J. Boissonnat and A. Ghosh. Manifold reconstruction using tangential Delaunay complexes. In *Proceedings of the 2010 annual symposium on Computational geometry*, pages 324–333. ACM, 2010.

[19] V. De Silva. A weak characterisation of the Delaunay triangulation. *Geometriae Dedicata*, 135(1):39–64, 2008.

[20] Facundo Mémoli. Gromov-Wasserstein distances and the metric approach to object matching. *Foundations of Computational Mathematics*, pages 1–71, 2011. 10.1007/s10208-011-9093-5.

[21] David Cohen-Steiner, Herbert Edelsbrunner, and Dmitriy Morozov. Vines and vineyards by updating persistence in linear time. In *SCG '06: Proceedings of the twenty-second annual symposium on Computational geometry*, pages 119–126, New York, NY, USA, 2006. ACM.

[22] Facundo Mémoli. Some properties of Gromov-Hausdorff distances. *Discrete & Computational Geometry*, pages 1–25, 2012. 10.1007/s00454-012-9406-8.

[23] Facundo Mémoli. Persistent Homology and Metric Geometry. In preparation. May 2012.

[24] B. Hudson, G. L. Miller, S. Y. Oudot, and D. R. Sheehy. Topological Inference via Meshing. Proceeding of Annual ACM Symposium on Computational Geometry, pages 277–286, 2010.

[25] D. R. Sheehy. Mesh Generation and Geometric Persistent Homology. PhD Thesis, 2011.

[26] D. R. Sheehy. New Bounds on the Size of Optimal Meshes. Computer Graphics Forum, 2012.

[27] G. L. Miller, T. Phillips, and D. R. Sheehy. Beating the Spread: Time-Optimal Point Meshing. Proceeding of Annual ACM Symposium on Computational Geometry, 2012.

[28] F. Chazal, D. Cohen-Steiner, and Q. Mérigot. Geometric inference for probability measures. *Foundations of Computational Mathematics*, 11:733–751, 2011.

[29] M. Belkin and K. Sinha. Polynomial Learning of Distribution Families. Proceeding of Annual IEEE Symposium on Foundations of Computer Science, 2010.

[30] Herbert Edelsbrunner, Brittany Terese Fasy, and Günter Rote. Add Isotropic Gaussian Mixtures at Own Risk: More and More Resilient Modes in Higher Dimensions. Proceedings of Annual ACM Symposium on Computational Geometry, 2012.

[31] M. Coudron and G. Lerman. On the sample complexity of robust pca. NIPS, 2012.

[32] G. Lerman, M. McCoy, J. A. Tropp, and T. Zhang. Robust computation of linear models, or How to find a needle in a haystack. *ArXiv e-prints 1202.4044*, Feb. 2012.

[33] T. Zhang and G. Lerman. A novel M-estimator for robust pca. Submitted, available at arXiv:1112.4863.

[34] N. Alon. Eigenvalues and expanders. *Combinatorica*, 6(2):83–96, 1986.

[35] N. Alon and V.D. Milman. λ_1, Isoperimetric Inequalities for Graphs, and Superconcentrators. *Journal of Combinatorial Theory, Series B*, 38(1):73–88, 1985.

[36] P. Buser. On Cheeger's Inequality $\lambda_1 \geq h^2/4$. In *Proc. Sympos. Pure Math*, volume 36, pages 29–77, 1980.

[37] J. Cheeger. A lower bound for the smallest eigenvalue of the Laplacian. *Problems in analysis*, pages 195–199, 1970.

[38] F. Chung. Random walks and local cuts in graphs. *Linear Algebra and its applications*, 423(1):22–32, 2007.

[39] F.R.K. Chung. *Spectral graph theory*. Amer. Mathematical Society, 1997.

[40] D. Dotterrer and M. Kahle. Coboundary expanders. *Arxiv preprint arXiv:1012.5316*, 2010.

[41] B. Eckmann. Harmonische Funktionen und Randwertaufgaben in einem Komplex. *Comm. Math. Helv.*, 17(1):240–255, 1944.

[42] M. Fiedler. Algebraic connectivity of graphs. *Czechoslovak Mathematical Journal*, 23(2):298–305, 1973.

[43] A. Gundert and U. Wagner. On Laplacians of random complexes. In *Proceedings of the 2012 Symposium on Computational Geometry*, pages 151–160. ACM, 2012.

[44] S. Hoory, N. Linial, and A. Wigderson. Expander graphs and their applications. *Bulletin of the American Mathematical Society*, 43(4):439, 2006.

[45] N. Linial and R. Meshulam. Homological connectivity of random 2-complexes. *Combinatorica*, 26(4):475–487, 2006.

[46] R. Meshulam and N. Wallach. Homological connectivity of random k-dimensional complexes. *Random Structures & Algorithms*, 34(3):408–417, 2009.

[47] B. Mohar. Isoperimetric numbers of graphs. *Journal of Combinatorial Theory, Series B*, 47(3):274–291, 1989.

[48] O. Parzanchevski, R. Rosenthal, and R.J. Tessler. Isoperimetric inequalities in simplicial complexes. *arXiv preprint arXiv:1207.0638*, 2012.