BULLETIN (New Series) OF THE
AMERICAN MATHEMATICAL SOCIETY
Volume 45, Number 1, January 2008, Pages 61–75
S 0273-0979(07)01191-3
Article electronically published on October 26, 2007

BARCODES: THE PERSISTENT TOPOLOGY OF DATA

ROBERT GHRIST

Abstract. This article surveys recent work of Carlsson and collaborators on applications of computational algebraic topology to problems of feature detection and shape recognition in high-dimensional data. The primary mathematical tool considered is a homology theory for point-cloud data sets—**persistent homology**—and a novel representation of this algebraic characterization—**barcodes**. We sketch an application of these techniques to the classification of natural images.

1. The shape of data

When a topologist is asked, “How do you visualize a four-dimensional object?” the appropriate response is a Socratic rejoinder: “How do you visualize a three-dimensional object?” We do not see in three spatial dimensions directly, but rather via sequences of planar projections integrated in a manner that is sensed if not comprehended. We spend a significant portion of our first year of life learning how to infer three-dimensional spatial data from paired planar projections. Years of practice have tuned a remarkable ability to extract global structure from representations in a strictly lower dimension.

The inference of global structure occurs on much finer scales as well, with regard to converting discrete data into continuous images. Dot-matrix printers, scrolling LED tickers, televisions, and computer displays all communicate images via arrays of discrete points which are integrated into coherent, global objects. This also is a skill we have practiced from childhood. No adult does a dot-to-dot puzzle with anything approaching anticipation.

1.1. **Topological data analysis.**

Problems of data analysis share many features with these two fundamental integration tasks: (1) how one infers high-dimensional structure from low-dimensional representations; and (2) how one assembles discrete points into global structure.

The principal themes of this survey of the work of Carlsson, de Silva, Edelsbrunner, Harer, Zomorodian, and others are the following:

(1) It is beneficial to replace a set of data points with a family of **simplicial complexes**, indexed by a proximity parameter. This converts the data set into global topological objects.

Received by the editors May 16, 2007, and, in revised form, July 4, 2007.

2000 *Mathematics Subject Classification.* Primary 55N35; Secondary 62H35.

This article is based on the lecture presented at the January 2007 national meeting of the AMS in New Orleans. The author gratefully acknowledges the support of DARPA #HR0011-07-1-0002 and the helpful comments of G. Carlsson, V. de Silva, and A. Zomorodian. The work reviewed in this article is funded by the DARPA program *TDA: Topological Data Analysis.*

(2) It is beneficial to view these topological complexes through the lens of algebraic topology — specifically, via a novel theory of **persistent homology** adapted to parameterized families.
(3) It is beneficial to encode the persistent homology of a data set in the form of a parameterized version of a Betti number: a **barcode**.

This review will introduce these themes and survey an example of these techniques applied to a high-dimensional data set derived from natural images.

1.2. **Clouds of data.** Very often, data is represented as an unordered sequence of points in a Euclidean n-dimensional space $\mathbb{E}^n$. Data coming from an array of sensor readings in an engineering testbed, from questionnaire responses in a psychology experiment, or from population sizes in a complex ecosystem all reside in a space of potentially high dimension. The global 'shape' of the data may often provide important information about the underlying phenomena that the data represent.

One type of data set for which global features are present and significant is the so-called **point cloud data** coming from physical objects in 3-d. Touch probes, point lasers, or line lasers sweep a suspended body and sample the surface, recording coordinates of anchor points on the surface of the body. The cloud of such points can be quickly obtained and used in a computer representation of the object. A temporal version of this situation is to be found in motion-capture data, where geometric points are recorded as time series. In both of these settings, it is important to identify and recognize global features: where is the index finger, the keyhole, the fracture?

Following common usage, we denote by point cloud data any collection of points in $\mathbb{E}^n$, though the connotation is that of a (perhaps noisy) sample of points on a lower-dimensional subset. For point clouds residing in a low-dimensional ambient

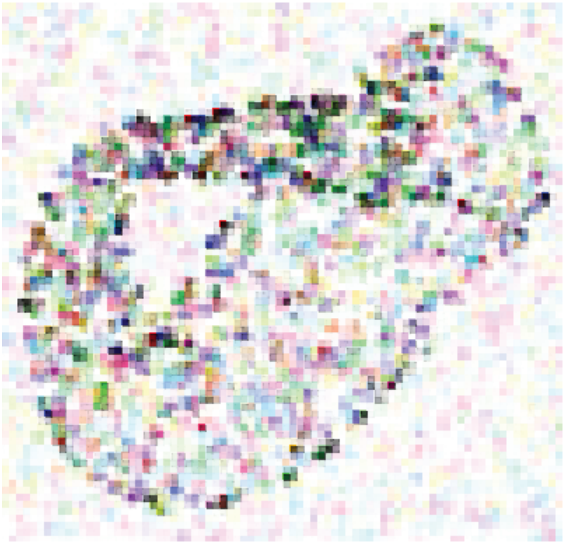

FIGURE 1. Determining the global structure of a noisy point cloud is not difficult when the points are in $\mathbb{E}^2$, but for clouds in higher dimensions, a planar projection is not always easy to decipher.

space, there are numerous approaches for inferring features based on planar projections: reconstruction techniques in the computer graphics and statistics literatures are manifold. From a naive point of view, planar projections would appear to be of limited value in the context of data which is inherently high-dimensional or sufficiently 'twisted' so as to preclude a faithful planar projection (Figure 1).

A more global and intrinsic approach to high-dimensional data clouds has recently appeared in the work of Carlsson and collaborators. This body of ideas applies tools from algebraic topology to extract coarse features from high-dimensional data sets. This survey is a brief overview of some of their work. As a result of our focus on techniques from algebraic topology, we neglect the large body of relevant work in nonlinear statistics (which is rarely topological) and in computer graphics (which is rarely high-dimensional).

1.3. **From clouds to complexes.** The most obvious way to convert a collection of points $\{x_\alpha\}$ in a metric space into a global object is to use the point cloud as the vertices of a combinatorial graph whose edges are determined by proximity (vertices within some specified distance ϵ). Such a graph, while capturing connectivity data, ignores a wealth of higher-order features beyond clustering. These features can be accurately discerned by thinking of the graph as a scaffold for a higher-dimensional object. Specifically, one completes the graph to a **simplicial complex** — a space built from simple pieces (simplices) identified combinatorially along faces. The choice of how to fill in the higher-dimensional simplices of the proximity graph allows for different global representations. Two of the most natural methods for doing so are as follows:

Definition 1.1. Given a collection of points $\{x_\alpha\}$ in Euclidean space $\mathbb{E}^n$, the **Čech complex**,[1] $\mathcal{C}_\epsilon$, is the abstract simplicial complex whose k-simplices are determined by unordered $(k+1)$-tuples of points $\{x_\alpha\}_0^k$ whose closed $\epsilon/2$-ball neighborhoods have a point of common intersection.

Definition 1.2. Given a collection of points $\{x_\alpha\}$ in Euclidean space $\mathbb{E}^n$, the **Rips complex**,[2] $\mathcal{R}_\epsilon$, is the abstract simplicial complex whose k-simplices correspond to unordered $(k+1)$-tuples of points $\{x_\alpha\}_0^k$ that are pairwise within distance ϵ.

The **Čech theorem** (or, equivalently, the "nerve theorem") states that $\mathcal{C}_\epsilon$ has the homotopy type of the union of closed radius $\epsilon/2$ balls about the point set $\{x_\alpha\}$. This means that $\mathcal{C}$, though an abstract simplicial complex of potentially high dimension, behaves exactly like a subset of $\mathbb{E}^n$ (see Figure 2). The Čech complex is a topologically faithful simplicial model for the topology of a point cloud fattened by balls. However, the Čech complex and various topologically equivalent subcomplexes (*e.g.*, the **alpha complex** of [13]) are delicate objects to compute, relying on the precise distances between the nodes in $\mathbb{E}^n$.

From a computational point of view, the Rips complex is less expensive that the corresponding Čech complex, even though the Rips complex has more simplices (in general). The reason is that the Rips complex is a **flag complex**: it is maximal among all simplicial complexes with the given 1-skeleton. Thus, the combinatorics of the 1-skeleton completely determines the complex, and the Rips complex can be

[1] Also known as the **nerve** of the associated cover by balls.

[2] A more appropriate name would be the Vietoris-Rips complex, in recognition of Vietoris' original use of these objects in the early days of homology theory [21]. For brevity we use the term "Rips complex".

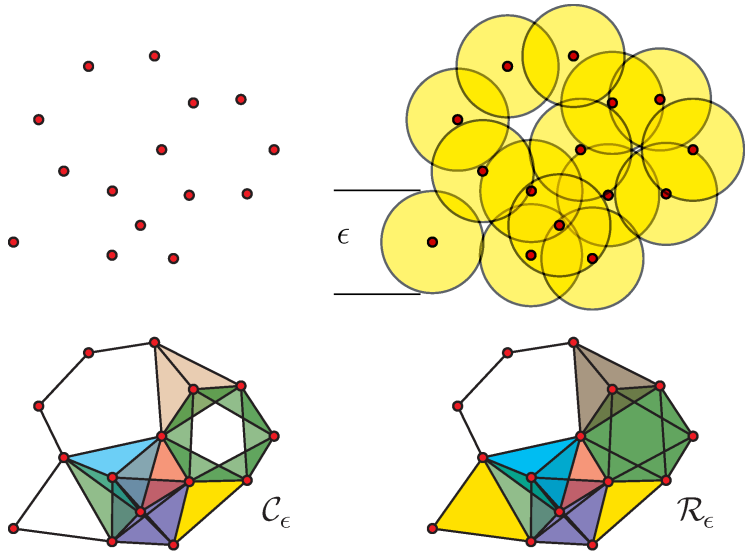

FIGURE 2. A fixed set of points [upper left] can be completed to a Čech complex $\mathcal{C}_\epsilon$ [lower left] or to a Rips complex $\mathcal{R}_\epsilon$ [lower right] based on a proximity parameter ϵ [upper right]. This Čech complex has the homotopy type of the $\epsilon/2$ cover ($S^1 \vee S^1 \vee S^1$), while the Rips complex has a wholly different homotopy type ($S^1 \vee S^2$).

stored as a graph and reconstituted instead of storing the entire boundary operator needed for a Čech complex. This virtue — that coarse proximity data on pairs of nodes determines the Rips complex — is not without cost. The penalty for this simplicity is that it is not immediately clear what is encoded in the homotopy type of $\mathcal{R}$. In general, it is neither a subcomplex of $\mathbb{E}^n$ nor does it necessarily behave like an n-dimensional space at all (Figure 2).

1.4. **Which ϵ?** Converting a point cloud data set into a global complex (whether Rips, Čech, or other) requires a choice of parameter ϵ. For ϵ sufficiently small, the complex is a discrete set; for ϵ sufficiently large, the complex is a single high-dimensional simplex. Is there an optimal choice for ϵ which best captures the topology of the data set? Consider the point cloud data set and a sequence of Rips complexes as illustrated in Figure 3. This point cloud is a sampling of points on a planar annulus. Can this be deduced? From the figure, it certainly appears as though an ideal choice of ϵ, if it exists, is rare: by the time ϵ is increased so as to remove small holes from within the annulus, the large hole distinguishing the annulus from the disk is filled in.

2. Algebraic topology for data

Algebraic topology offers a mature set of tools for counting and collating holes and other topological features in spaces and maps between them. In the context of high-dimensional data, algebraic topology works like a telescope, revealing objects and features not visible to the naked eye. In what follows, we concentrate on homology for its balance between ease of computation and topological resolution. We

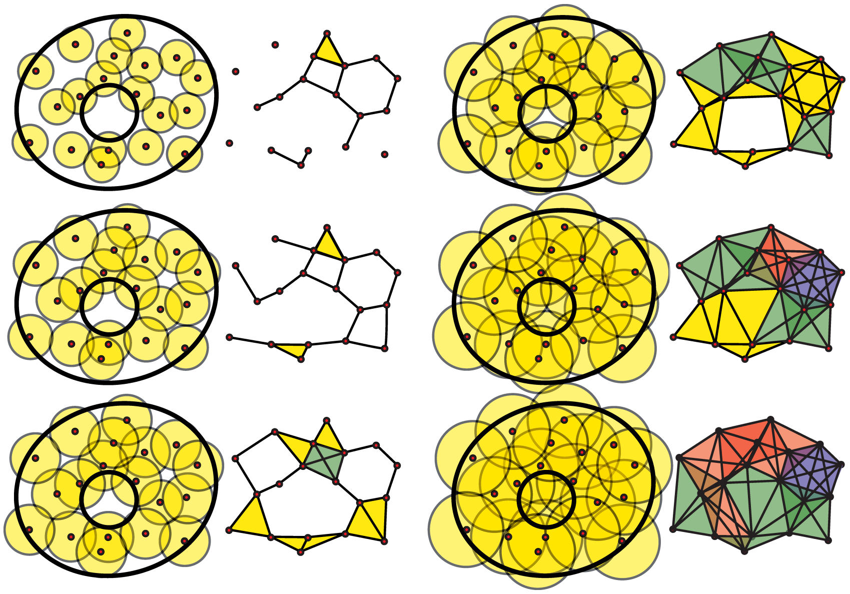

FIGURE 3. A sequence of Rips complexes for a point cloud data set representing an annulus. Upon increasing ϵ, holes appear and disappear. Which holes are real and which are noise?

assume a rudimentary knowledge of homology, as is to be found in, say, Chapter 2 of [15].

Despite being both computable and insightful, the homology of a complex associated to a point cloud at a particular ϵ is insufficient: it is a mistake to ask which value of ϵ is optimal. Nor does it suffice to know a simple 'count' of the number and types of holes appearing at each parameter value ϵ. Betti numbers are not enough. One requires a means of declaring which holes are essential and which can be safely ignored. The standard topological constructs of homology and homotopy offer no such slack in their strident rigidity: a hole is a hole no matter how fragile or fine.

2.1. **Persistence.** Persistence, as introduced by Edelsbrunner, Letscher, and Zomorodian [12] and refined by Carlsson and Zomorodian [22], is a rigorous response to this problem. Given a parameterized family of spaces, those topological features which persist over a significant parameter range are to be considered as signal with short-lived features as noise. For a concrete example, assume that $\mathsf{R} = (\mathcal{R}_i)_1^N$ is a sequence of Rips complexes associated to a fixed point cloud for an increasing sequence of parameter values $(\epsilon_i)_1^N$. There are natural inclusion maps

$$\mathcal{R}_1 \overset{\iota}{\hookrightarrow} \mathcal{R}_2 \overset{\iota}{\hookrightarrow} \cdots \overset{\iota}{\hookrightarrow} \mathcal{R}_{N-1} \overset{\iota}{\hookrightarrow} \mathcal{R}_N. \tag{2.1}$$

Instead of examining the homology of the individual terms $\mathcal{R}_i$, one examines the homology of the iterated inclusions $\iota : H_*\mathcal{R}_i \to H_*\mathcal{R}_j$ for all $i < j$. These maps reveal which features persist.

As a simple example, persistence explains why Rips complexes are an acceptable approximation to Čech complexes. Although no single Rips complex is an especially faithful approximation to a single Čech complex, pairs of Rips complexes 'squeeze' the appropriate Čech complex into a manageable hole.

Lemma 2.1. *For any $\epsilon > 0$, there is a chain of inclusion maps*

$$\mathcal{R}_\epsilon \hookrightarrow \mathcal{C}_{\epsilon\sqrt{2}} \hookrightarrow \mathcal{R}_{\epsilon\sqrt{2}}. \tag{2.2}$$

(See [10] for the tight dimension-dependent expansion bound smaller than $\sqrt{2}$.) This implies that any topological feature which persists under the inclusion $\mathcal{R}_\epsilon \hookrightarrow \mathcal{R}_{\epsilon'}$ is in fact a topological feature of the Čech complex $\mathcal{C}_{\epsilon'}$ when $\epsilon'/\epsilon \geq \sqrt{2}$. *Moral:* The homology of the inclusion $\iota_* : H_*\mathcal{R}_\epsilon \to H_*\mathcal{R}_{\epsilon'}$ reveals information that is not visible from $H_*\mathcal{R}_\epsilon$ and $H_*\mathcal{R}_{\epsilon'}$ unadorned. This is a foreshadowing of the broader idea of persistence arising in an arbitrary sequence of chain complexes.

2.2. **Persistent homology.** One begins with a **persistence complex**: a sequence of chain complexes $\mathsf{C} = (C_*^i)$ together with chain maps $x : C_*^i \longrightarrow C_*^{i+1}$. (For notational simplicity, we do not index the chain maps x.) This is motivated by having a sequence of Rips or Čech complexes of increasing ϵ sampled at an increasing sequence of parameters $\{\epsilon_i\}$. Since Rips or Čech complexes grow with ϵ, the chain maps x are naturally identified with inclusions.

Definition 2.2. For $i < j$, the (i,j)-persistent homology of C, denoted $H_*^{i\to j}(\mathsf{C})$, is defined to be the image of the induced homomorphism $x_* : H_*(C_*^i) \to H_*(C_*^j)$.

As an example, consider the filtration $\mathsf{R} = (\mathcal{R}_i)$ of Rips complexes parameterized by proximities ϵ_i. Lemma 2.1 implies that if $\epsilon_j/\epsilon_i \geq \sqrt{2}$, then $H_k^{i\to j}(\mathsf{R}) \neq 0$ implies $H_k(\mathcal{C}_{\epsilon_j}) \neq 0$. Holes in the Čech complex are inferred by the persistent homology of the Rips filtration.

There is a good deal more algebraic structure in the interleaving of persistent homology groups, as explained in the work of Carlsson and Zomorodian. Fix a PID of coefficients R and place a graded $R[x]$-module structure on C with x acting as a shift map. That is, a unit monomial $x^n \in R[x]$ sends C_*^i to C_*^{i+n} via n applications of x. One assumes a finite-type condition that each C_*^i is finitely generated as an $R[x]$-module and that the sequence stabilizes in i (in the case of an infinite sequence of chain complexes).

As the filtering of C is via chain maps x (*cf.* the setting of Rips complexes — simplices are added but never removed as ϵ increases), C is free as an $R[x]$-module. The resulting homology $H_*(\mathsf{C})$ retains the structure of an $R[x]$-module, but, unlike the chain module, is not necessarily free. Nor is it easily classified: the Artin-Rees theory from commutative algebra implies that the problem of classifying (finite-type) persistence modules such as $H_*(\mathsf{C})$ is equivalent to classifying finitely generated nonnegatively graded $R[x]$-modules. This is known to be very difficult in, say, $\mathbb{Z}[x]$.

However, for coefficients in a field F, the classification of $F[x]$-modules follows from the Structure Theorem for PID's, since the only graded ideals of $F[x]$ are of the form $x^n \cdot F[x]$. This implies the following:

Theorem 2.3 ([22]). *For a finite persistence module C with field F coefficients,*

$$H_*(\mathsf{C}; F) \cong \bigoplus_i x^{t_i} \cdot F[x] \;\oplus\; \left(\bigoplus_j x^{r_j} \cdot (F[x]/(x^{s_j} \cdot F[x]))\right). \tag{2.3}$$

This classification theorem has a natural interpretation. The free portions of Equation (2.3) are in bijective correspondence with those homology generators which come into existence at parameter t_i and which persist for all future parameter values. The torsional elements correspond to those homology generators which appear at parameter r_j and disappear at parameter $r_j + s_j$. At the chain level, the Structure Theorem provides a birth-death pairing of generators of C (excepting those that persist to infinity).

2.3. **Barcodes.** The parameter intervals arising from the basis for $H_*(\mathsf{C}; F)$ in Equation (2.3) inspire a visual snapshot of $H_k(\mathsf{C}; F)$ in the form of a **barcode**. A barcode is a graphical representation of $H_k(\mathsf{C}; F)$ as a collection of horizontal line segments in a plane whose horizontal axis corresponds to the parameter and whose vertical axis represents an (arbitrary) ordering of homology generators. Figure 4 gives an example of barcode representations of the homology of the sampling of points in an annulus from Figure 3 (illustrated in the case of a large number of parameter values ϵ_i).

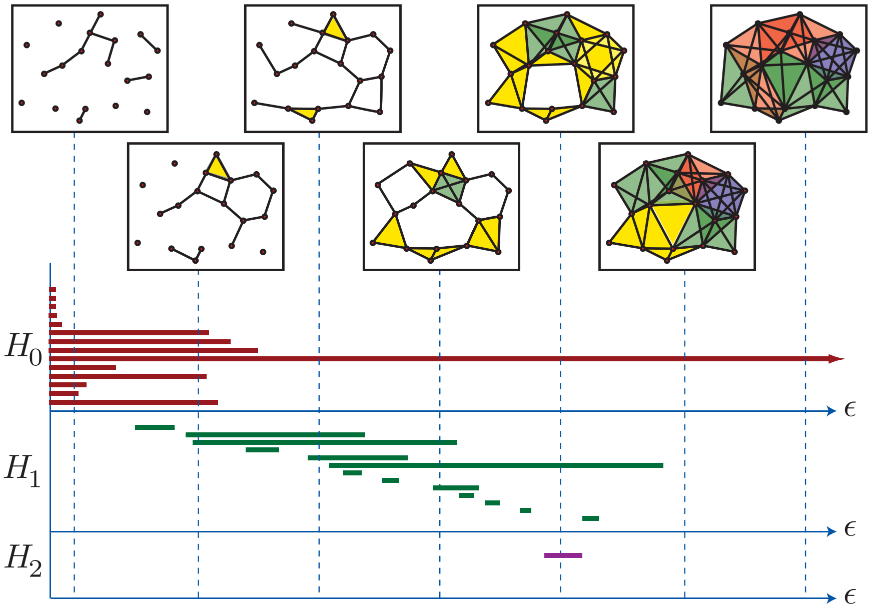

FIGURE 4. [bottom] An example of the barcodes for $H_*(\mathsf{R})$ in the example of Figure 3. [top] The rank of $H_k(\mathcal{R}_{\epsilon_i})$ equals the number of intervals in the barcode for $H_k(\mathsf{R})$ intersecting the (dashed) line $\epsilon = \epsilon_i$.

Theorem 2.3 yields the fundamental characterization of barcodes.

Theorem 2.4 ([22])**.** *The rank of the persistent homology group* $H_k^{i\to j}(\mathsf{C}; F)$ *is equal to the number of intervals in the barcode of* $H_k(\mathsf{C}; F)$ *spanning the parameter interval* $[i, j]$*. In particular,* $H_*(C_*^i; F)$ *is equal to the number of intervals which contain* i*.*

A barcode is best thought of as the persistence analogue of a Betti number. Recall that the k^{th} Betti number of a complex, $\beta_k := \text{rank}(H_k)$, acts as a coarse numerical measure of H_k. As with β_k, the barcode for H_k does not give any information about the finer structure of the homology, but merely a continuously parameterized rank. The genius of a barcode representation is the ability to qualitatively filter out topological noise and capture significant features. Indeed, as shown in [7], barcodes are stable in the presence of noise added to a [Morse] filtration. For example, in Figure 4, one sees (from a *very* coarse sampling) that the point cloud likely represents a connected object with one or two significant 'holes' as measured by H_1 and no significant higher homology.

2.4. **Computation.** Most invariants in modern algebraic topology are not known for their ease of computation. Homology (in its simplest manifestations) appears exceptional in that the invariants arise as quotients of finite-dimensional vector spaces. In the context of applications, 'finite' may exceed reasonable bounds. There is no recourse to chanting *"Homology is just linear algebra"* when faced with millions of simplices: one needs good algorithms. Fortunately, such exist with increasing scope and speed. The text [16] gives a comprehensive introduction to issues of and algorithms available for computing homology for realistic problems in application domains.

More fortunate still, there is an excellent algorithm available for the computation of persistent homology groups and barcodes. The algorithm takes as its argument the filtered simplicial complex consisting of pairs (σ_i, τ_i), where σ_i is a simplex and τ_i is the time at which that simplex appears in the filtration. This algorithm first appears in the paper of Edelsbruner, Letscher, and Zomorodian [12] for simplicial subcomplexes of $\mathbb{E}^3$ with $\mathbb{Z}_2$ coefficients and in that of Carlsson and Zomorodian [22] for general persistence complexes with field coefficients. The Matlab-based front end `Plex` by de Silva and Perry [11] incorporates the C++ persistent homology library of Kettner and Zomorodian with tools for inputting and manipulating simplicial complexes.

It is worth noting that for chain filtrations arising from realistic data sets, the Rips complexes are of an unmanageable size. This necessitates efficient sampling or reduction of the complex with accurate topology. The **witness complex** of Carlsson and de Silva [8, 9, 14] is one solution to this problem.

2.5. **Other directions.** We note that the above is the briefest of treatments of what quickly becomes a fascinating and very active sub-topic of computational topology. For those interested in the algebraic-topological aspects of the theory, we note the following recent developments:

- There are other filtrations besides those associated to Čech or Rips complexes which are natural settings in which to contemplate persistence. The **Morse filtration** of a space X outfitted with $f : X \to \mathbb{R}$ is a filtration of X by excursion sets $X_t = \left\{f^{-1}\left((-\infty, t]\right)\right\}$. This (or a discretized version thereof) is one commonly investigated setting [1, 7], as is filtration by means of curvature data [5].
- Our discussion of persistence is couched in the setting of chain complexes indexed by a single parameter. There are strong motivations for wanting to treat multi-parameter families of complexes. However, there are fundamental algebraic difficulties in constructing an analogous theory of persistence modules in this setting [24].
- The computation of persistent relative homology is more subtle, since the ensuing parameterized chain complex C is no longer free as an $F[x]$-module. Bendich and Harer [in progress] have developed an algebraic construction

for defining and computing persistent homology which has a particularly clean form in the setting of a Morse filtration. The analogue of Theorem 2.3 provides a perfect pairing of Morse critical points.

- The computation of persistent cohomology is not straightforward. As shown by de Silva [in progress], if you take the graded free $F[x]$-module chain complex C for the Morse filtration X_t of a space X and dualize it as a graded free $F[x]$-module, *i.e.*, if you construct $\mathrm{Hom}_{F[x]}(\mathsf{C}, F[x])$, then the homology of the resulting object as a graded $F[x]$-module is *not* the persistent cohomology of $H^*(X_t)$, but rather that of the relative cohomology $H^*(X, X_t)$. Computing absolute persistent cohomology necessitates a recourse to duality and the theory of Bendich-Harer above.

3. Example: natural images

One recent example of discovering topological structure in a high-dimensional data set comes from **natural images**. A collection of 4,167 digital photographs of random outdoor scenes was assembled in the late 1990s by van Hateren and van der Schaaf [20]. Mumford and others have posed several fascinating questions about the structure and potential universality of the statistics of this and similar sets of images in the context of visual perception [17].

3.1. **"Round about the cauldron go".** Mumford, Lee, and Pederson [18] construct a data set by choosing at random 5,000 three-pixel by three-pixel squares within each digital image and retaining the top 20% of these with respect to contrast. Each such square is a matrix of grey-scale intensities. The full data set consists of roughly 8,000,000 points in $\mathbb{E}^9$. By normalizing with respect to mean intensity and restricting attention to high-contrast images (those away from the origin), the data set is projected to a set of points $\mathcal{M}$ on a topological seven-sphere $S^7 \subset \mathbb{E}^8$. The details of this data set construction require a choice of natural basis with respect to a particular norm for values of contrast patches. We refer the interested reader to [18] for details.

3.2. **"Hover through the fog".** So coarse a reduction of natural images (into three-by-three squares of greyscale intensities) still leads to a point cloud of too high a dimension to visualize. Worse still, what structure there is is blurred and foggy: points appear at first to be distributed over the entire S^7. A resort to density considerations is thus in order. The subject of density filtration is a well-trod area of statistics: see, *e.g.*, [19].

A **codensity** function is used in [3] as follows. Fix a positive integer $k > 0$. For any point x_α in the data set, define $\delta_k(x_\alpha)$ as the distance in $\mathbb{E}^n$ from x_α to the k^{th} nearest neighbor of x_α in the data set. For a fixed value of k, δ_k is a positive distribution over the point cloud which measures the radius of the ball needed to enclose k neighbors. Values of δ_k are thus inversely related to the point cloud density. The larger a value of k used, the more averaging occurs among neighbors, blurring finer variations.

The codensity is used to filter the data as follows. Denote by $\mathcal{M}[k, T]$ the subset of $\mathcal{M}$ in the upper T-percent of density as measured by δ_k. This is a two-parameter subset of the point cloud which, for reasonable values of k and T, represents an appropriate core.

3.3. **"When shall we three meet again?"** The first interesting persistent homology computation on this data set occurs at the level of H_1: to what extent are there 'loops' in the data set along which the cloud is concentrated?

Taking a density threshold of $T = 25$ at neighbor parameter $k = 300$, with 5,000 points sampled at random from $\mathcal{M}[k, T]$, computing the barcode for the first homology H_1 reveals a unique persistent generator [3]. See Figure 5. This indicates that the data set is diffused about a primary circle in the 7-sphere. The structure of the barcode is robust with respect to the random sampling of the points in $\mathcal{M}[k, T]$.

The goal of the homology computation is to discover a 'hidden' feature of a data set that is not discernable by clustering and connectivity alone. The simplest such feature would be, as indicated by the computation above, a primary circle about which the data is scattered. To what might this correspond? A close examination of the data point corresponding to the primary circle reveals a pattern of 3-by-3 patches with one light region and one dark region separated by a linear transition. This **nodal curve** between light and dark is linear and appears in a circular family parameterized by the angle of the nodal line, as shown in Figure 5.

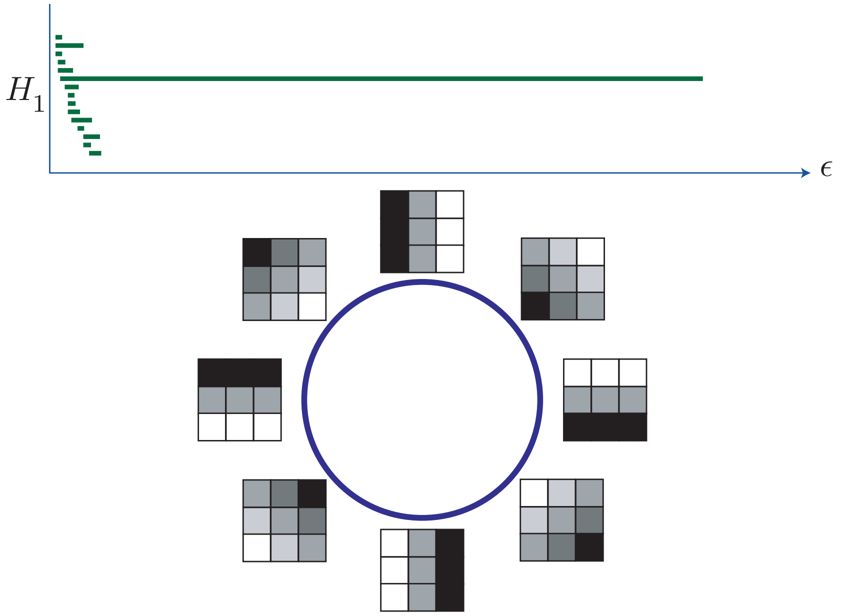

FIGURE 5. The H_1 barcode for a random sampling of 5,000 points of $\mathcal{M}[300, 25]$ yields a single generator. This generator indicates the nodal line between a single light and single dark patch as being the dominant feature of the primary circle in $\mathcal{M}$.

As seen from the barcode, this generator is dominant at the threshold and codensity parameters chosen. An examination of the barcodes for the first homology group H_1 of the data set filtered by codensity parameter $k = 15$ and threshold $T = 25$ reveals a different persistent first homology. The reduction in k leads to less averaging and more localized density sensitivity. The barcode of Figure 6 reveals that the persistent H_1 of samples from $\mathcal{M}[k, T]$ has Betti number five. This does not connote the presence of five disjoint circles in the data set. Rather, by focusing on the generators and computing the barcode for H_0, it is observed [3]

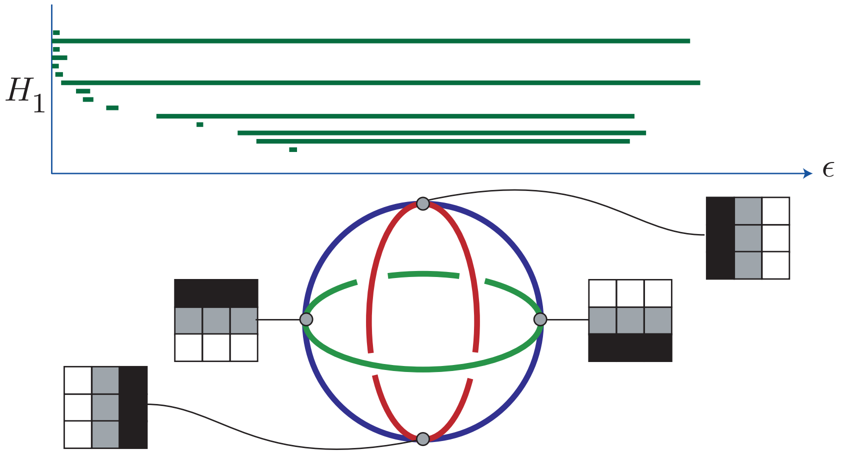

FIGURE 6. The H_1 barcode for $\mathcal{M}[15, 25]$ reveals five persistent generators. This implies the existence of two secondary circles, each of which intersects the third, large-k, primary circle twice.

that, besides the primary circle from the high-k H_1 computation, there are two secondary circles which come into view at the lower density parameter.

A close examination of these three circles reveals that each intersects the high-k primary twice, yet the two secondary circles are disjoint. To which features in the data might these secondary circles correspond? As noted in [3], each secondary circle regulates images with three contrasting regions and interpolates between these states and the primary circle. The difference between the two secondary circles lies in their bias for horizontal and vertical stratification respectively. Figure 7 gives an interpretation of the meanings of the secondary circles.

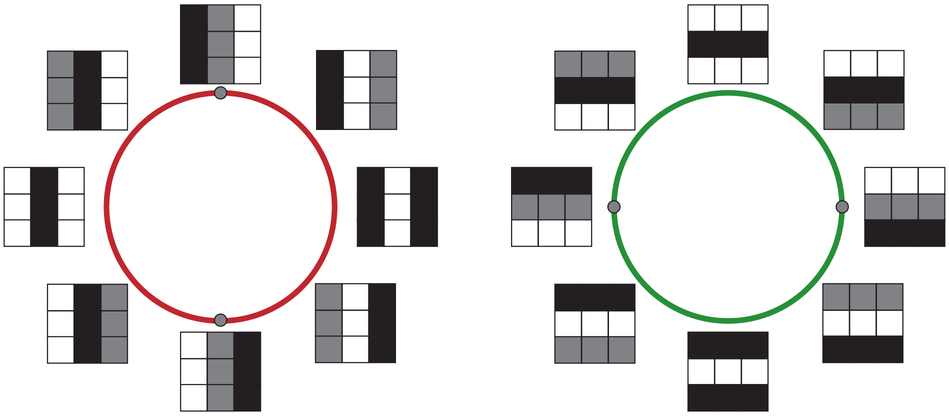

FIGURE 7. The secondary generators of H_1 for $\mathcal{M}[15, 25]$ have an interpretation as regulating changes from dual-patch to triple-patch high contrast regions in horizontal and vertical biases respectively.

3.4. **"Come like shadows, so depart!"** What is the good of temporary topological features which emerge and dissolve as a function of the parameter ϵ? Does this lead to anything more than a heuristic for high-dimensional data sets that are hard to visualize? While the work of Carlsson et al. is very recent, there are several applications of the topological approach to data analysis which argue in favor of the proposition that homological structures in high-dimensional data sets are of scientific significance. Besides the Mumford data set reviewed here, persistent homology computations are being applied in several disparate contexts, including geometric features of curves (*e.g.*, optical character recognition) [5] and spike train data from implanted electrode arrays in the primary visual cortex of Macaque monkeys [4]. The latter project has as its goal the understanding of how the topology of a parameterized space of images is represented in neural data: just as the lens manipulates and projects an image onto the retina, an image parameter space is transformed and projected directly into the visual cortex.

Regarding the natural image data, it is instructive to think of the persistent homology of $\mathcal{M}$ as something akin to a series expansion of the true space. The reduction of the full data set to an S^7 via projection is really a normalization to eliminate the zero-order (or "single patch") terms in the data set. Following this analogy, the H_1 primary generator fills the role of a next term in the expansion of the homotopy type of the data set, collating the nodal curve between two contrasting patches. The secondary circles, interpolating between single and dual nodal curves, act as higher-order terms in the expansion, in which horizontal and vertical biases arise.

It is here that one gets deeper insight into the data set. Inspired by the meaning of the H_1 barcodes of $\mathcal{M}$, further investigation reveals what appears to be an intrinsic bias toward horizontal and vertical directions in the natural image data, as opposed to an artifact of the (right) angle at which the camera was held: [3] reports that a repetition of the experiment with a camera held at a constant angle $\pi/4$ yields a data set whose secondary persistent H_1 generators exhibit a bias towards true vertical and true horizontal: the axis of pixellation appears less relevant than the axis of gravity in natural image data.

Is there any predictive power in the barcodes of the data set? Recent progress [3] demonstrates the insight that a persistent topology approach can yield. The

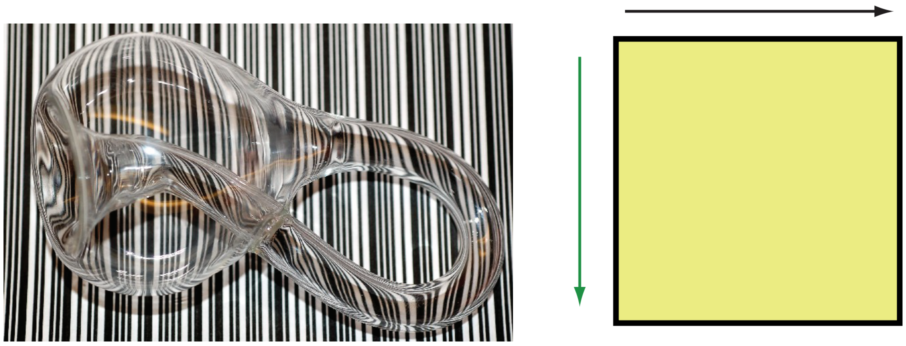

Figure 8. A Klein bottle (pictured against a computer-generated UPC barcode of the string "Klein bottle") [left] is the non-orientable surface obtained by identifying opposite sides of a square as shown [right].

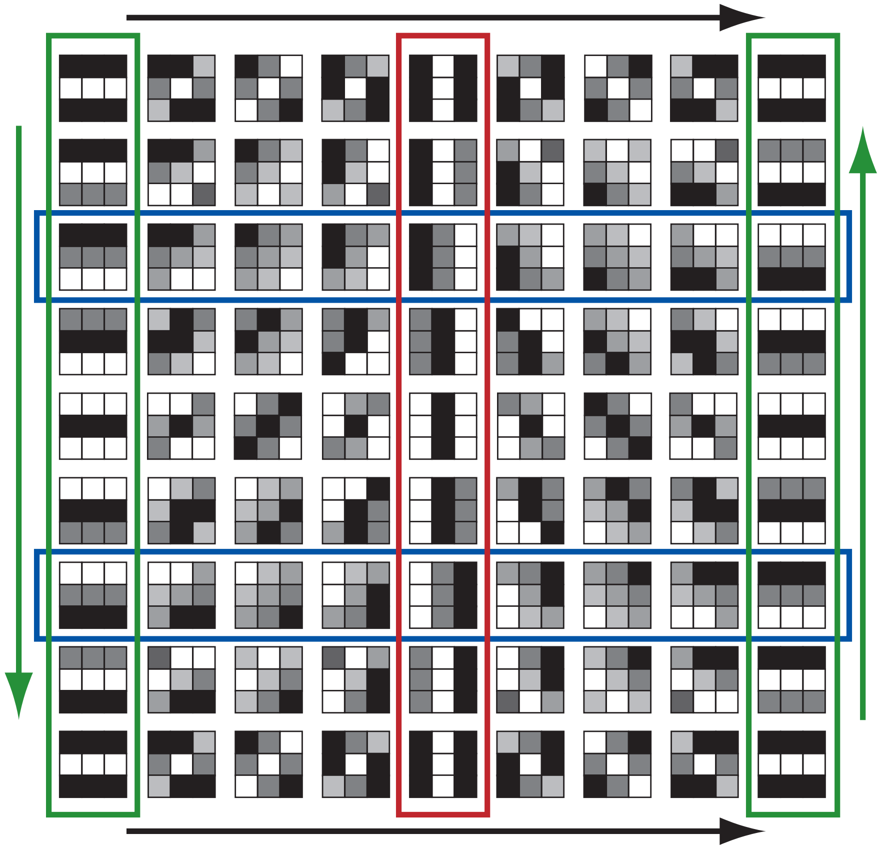

FIGURE 9. A Klein bottle embeds naturally in the parameter space as a completion of the 3-circle model. In the unfolded identification space shown, the primary circle wraps around the horizontal axis twice. The two secondary circles each wrap around the vertical axis once (note: the circle on the extreme left and right are glued together with opposite orientation). Each secondary circle intersects the primary circle twice.

barcodes for the second persistent homology H_2 are more volatile with respect to changes in density and thresholding. This is not surprising: the lowest order terms in any series expansion are always most easily perceived. However, there is indication of a persistent H_2 generator (in $\mathbb{Z}_2$ coefficients) at certain settings of k and T. Combined with the basis of H_1 generators, one obtains predictive insight to the structure of the space of high-contrast patches. At certain density thresholds, the H_2 barcode, suitably trimmed with Occam's razor, suggests a two-dimensional completion of the low-k persistent H_1 basis into a **Klein bottle** (see Figure 8). Recall that this nonorientable surface can be realized as an identification space of a square, as in the figure. Figure 9 illustrates an embedding of this surface in the space of pixellated images. One notes that this is a natural completion of the low-density persistent H_1 readings: the primary and secondary circles appear with the appropriate intersection properties. Fortunately, a pair of homology computations in $\mathbb{Z}_2$ and $\mathbb{Z}_3$ coefficients — finite fields being most natural for computer experiments — is efficacious in verifying that the persistent surface found is a Klein bottle.

We emphasize that the point cloud data set $\mathcal{M}$ is vast, high-dimensional, and not at all concentrated sharply along distinct features. A cursory viewing of the data seems to indicate that the 7-sphere is filled densely with data points and that there is seemingly no coherent structure to be found. It is through the lens of persistent homology — suitably tuned and aimed — that cogent features emerge and fade with changing parameters. These persistent generators, upon close examination, do correspond to meaningful structures in the data, inspiring a sensible parametrization of the global structure of the data set. This is the type of explanatory power that any exemplar of good applied mathematics provides to a scientific challenge.

About the author

Robert Ghrist is a professor in the Department of Mathematics, with affiliation in the Coordinated Science Laboratory at the University of Illinois, Urbana-Champaign.

References

[1] P. Bubenik and P. Kim, "A statistical approach to persistent homology", preprint (2006), math.AT/0607634.

[2] E. Carlsson, G. Carlsson, and V. de Silva, "An algebraic topological method for feature identification", *Intl. J. Computational Geometry and Applications*, 16:4 (2006), 291-314. MR2250511 (2007c:52015)

[3] G. Carlsson, T. Ishkhanov, V. de Silva, and A. Zomorodian, "On the local behavior of spaces of natural images", *Intl. J. Computer Vision*, in press.

[4] G. Carlsson, T. Ishkhanov, F. Mémoli, D. Ringach, and G. Sapiro, "Topological analysis of the responses of neurons in V1", in preparation (2007).

[5] G. Carlsson, A. Zomorodian, A. Collins, and L. Guibas, "Persistence barcodes for shapes", *Intl. J. Shape Modeling*, 11 (2005), 149-187.

[6] F. Chazal and A. Lieutier, "Weak feature size and persistent homology: computing homology of solids in $\mathbb{R}^n$ from noisy data samples", in *Proc. 21st Sympos. Comput. Geom.* (2005).

[7] D. Cohen-Steiner, H. Edelsbrunner, and J. Harer, "Stability of persistence diagrams", in *Proc. 21st Sympos. Comput. Geom.* (2005), 263–271.

[8] V. de Silva, "A weak definition of Delaunay triangulation", preprint (2003).

[9] V. de Silva and G. Carlsson, "Topological estimation using witness complexes", in *SPBG'04 Symposium on Point-Based Graphics* (2004), 157-166.

[10] V. de Silva and R. Ghrist, "Coverage in sensor networks via persistent homology", *Alg. & Geom. Topology*, 7 (2007), 339–358.

[11] V. de Silva and P. Perry, PLEX home page, http://math.stanford.edu/comptop/programs/plex/.

[12] H. Edelsbrunner, D. Letscher, and A. Zomorodian, "Topological persistence and simplification", *Discrete Comput. Geom.*, 28:4 (2002), 511-533. MR1949898 (2003m:52019)

[13] H. Edelsbrunner and E.P. Mücke, "Three-dimensional alpha shapes", *ACM Transactions on Graphics*, 13:1 (1994), 43-72.

[14] L. Guibas and S. Oudot, "Reconstruction using witness complexes", in *Proc. 18th ACM-SIAM Sympos. on Discrete Algorithms* (2007).

[15] A. Hatcher, *Algebraic Topology*, Cambridge University Press (2002). MR1867354 (2002k:55001)

[16] T. Kaczynski, K. Mischaikow, and M. Mrozek, *Computational Homology,* Applied Mathematical Sciences, 157, Springer-Verlag (2004). MR2028588 (2005g:55001)

[17] D. Mumford, "Pattern Theory: The Mathematics of Perception", Proc. Intl. Congress of Mathematicians, Vol. I (2002), 401–422. MR1989195 (2004k:91168)

[18] D. Mumford, A. Lee, and K. Pedersen, "The nonlinear statistics of high-contrast patches in natural images", *Intl. J. Computer Vision*, 54 (2003), 83–103.

[19] B. Silverman, *Density Estimation for Statistics and Data Analysis*, Chapman & Hall/CRC (1986). MR848134 (87k:62074)

[20] J. van Hateren and A. van der Schaff, "Independent Component Filters of Natural Images Compared with Simple Cells in Primary Visual Cortex", *Proc. R. Soc. London*, B 265 (1998), 359–366.

[21] L. Vietoris, "Über den höheren Zusammenhang kompakter Räume und eine Klasse von zusammenhangstreuen Abbildungen", *Math. Ann.*, 97 (1927), 454–472. MR1512371

[22] A. Zomorodian and G. Carlsson, "Computing persistent homology", *Discrete Comput. Geom.*, 33 (2005), 249–274. MR2121296 (2005j:55004)

[23] A. Zomorodian and G. Carlsson, "Localized homology", *Proc. Shape Modeling International* (2007), 189–198.

[24] A. Zomorodian and G. Carlsson, "The theory of multidimensional persistence", *Proc. Symposium on Computational Geometry* (2007), 184–193.

Department of Mathematics and Coordinated Science Laboratory, University of Illinois, Urbana, Illinois 61801

E-mail address: ghrist@math.uiuc.edu