Internat. Math. Nachrichten
Nr. 231 (2016), 15–33

# Persistent Homology – State of the art and challenges

**Michael Kerber**

Graz University of Technology

## 1 Motivation for multi-scale topology

A recurring task in mathematics, statistics, and computer science is understanding the connectivity information, or equivalently, the topological properties of a given object. For concreteness, we assume the object in question to be a geometric shape, possibly embedded in a high-dimensional space, although that assumption is not necessary for most of the theory. Algebraic topology offers a toolset for quantifying and comparing topological features of such shapes. The strongest notion of topological equivalence, the existence of an *homeomorphism* between topological spaces, is out of reach in general in computational contexts.[1] An attractive compromise is offered by the theory of homology over a base field $\mathbb{F}$. In informal terms, the $p$-th homology group $H_p(\mathcal{S})$ of a shape $\mathcal{S}$ (with $p \geq 0$) is a $\mathbb{F}$-vector space whose rank counts the number of "$p$-dimensional holes" in $\mathcal{S}$. Concretely, for objects embedded in $\mathbb{R}^3$, $\operatorname{rank} H_{0,1,2}(\mathcal{S})$ count the number of connected components, tunnels, and voids, respectively, induced by the shape $\mathcal{S}$. Homology over fields reveals less topological information then the $\mathbb{Z}$-homology, but this partial information is sufficient for many purposes. The main advantage of restricting to fields is the existence of efficient algorithms. More precisely, if the input is given as a combinatorial cell complex, the homology groups in all dimensions can be computed in cubic time with respect to the number of cells.

**Multi-scale and noise.** We discuss three basic exemplary scenarios in which topological information reveals potentially valuable information. For each scenario, other tools can be employed as well; the goal is rather to underline the general applicability of topology as a tool for data analysis.

[1] The question whether two shapes are homeomorphic is undecidable for shapes of dimension 4 and higher [54].

ISSN 0020-7926 © 2016 Österr. Math. Gesellschaft

- Combustion is a highly complex dynamic process relevant for engineering applications. Consider the goal of analyzing the temperature distribution of a combustion for a fixed moment in time. One approach could be to fix a temperature threshold and decompose the domain into “hot” and “cold” areas. The connectivity of these areas allows an identification of hot or cold pockets which might guide the analyst to areas of importance in the process.

- The task of shape retrieval is to find for a query point cloud (for instance obtained by a 3D-scanner) the closest representation in some database of shapes. A topology-based similarity measure provides a high-level summary which can be used to quickly rule out shapes with very different topology.

- Clustering is one of the most fundamental problems in data analysis. As an example, imagine an internet company collecting data about users in terms of various real-valued parameters. The users form a high-dimensional point cloud, and grouping them into clusters of similar users facilitates decision making (e.g., personalized product offers) and predictions of the user’s behavior in the future. Understanding the topology of that “user space” can be helpful to design a reasonable notion of similarity measures.

The combustion example above contains a scale parameter, identifying what parts are considered hot and cold. A parameter is also intrinsic in the other applications: at first sight, the input is merely a discrete point cloud without interesting topological features. It is required to build a model of the underlying space from which the point cloud was drawn (i.e., the shape that has been scanned). The most frequently employed technique is to replace the points by balls of a fixed radius, and to take the union of these balls as an approximation of the underlying space (cf. Figure 1). In this case, the ball radius constitutes the scale parameter. This raises the question of which radius to choose: a small radius might give a too fine-grained picture while a large radius might blur relevant information contained in the data.

In many applications, there is no natural choice of what is the best scale to look at. In such cases, one might want to consider various scales and to select the best choice afterwards. However, this *multi-scale* approach is affected by the presence of *noise* in the data. For instance, an inaccurate scanning of a shape might lead to a large number of “bubbles” in the approximation, increasing the number of voids in the shape and occluding the real topological features. Such noise can be present at all scales, complicating the task of separating signal and noise in the data.

**Persistent homology.** The main idea of persistence is to connect the homological information gathered across different scales. In this way, we can identify

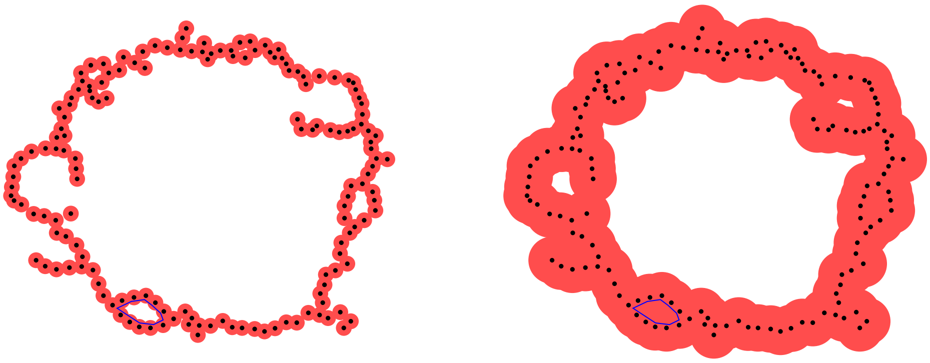


Figure 1: Representation of a point cloud on two different scales as a union of balls. On the smaller scale, we count 5 holes in the shape, or equivalently, $\beta_1 = 5$. On the larger scale, $\beta_1 = 3$. However, the *persistent* Betti number of the inclusion map is 1, because 4 of the 5 small-scale holes disappear after the inclusion. This is illustrated for the bottom hole by the blue cycle generating the corresponding homology class, which becomes trivial in the larger union. Only the larger hole "survives" the inclusion from small to large scale, making it the only persistent feature that spans over this range of scales.

which topological features are present over a large range of scales as opposed to those which are only spuriously present.

To describe the idea mathematically, consider two spaces $X \subseteq Y$, corresponding to representations of data on different scales (think about two sublevel sets of a function, or two unions of balls with different radius). The inclusion map $X \hookrightarrow Y$ induces, for any $p \geq 0$, a linear map between the vector spaces

$$\phi : H_p(X) \to H_p(Y),$$

as a consequence of the functorial properties of homology [56]. We define the *persistent Betti number* with respect to $(X,Y)$ as

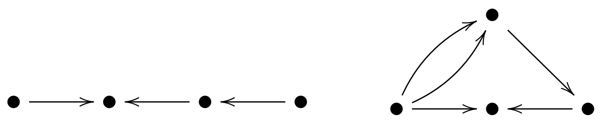
$$\operatorname{rank}(\operatorname{Im}\phi),$$

which counts the number of homological features in $Y$ which have already been present in $X$ (see Figure 1 for an example). Having a multi-scale representation of a given data set, we obtain a persistent Betti number for each pair of scales. They constitute a topological multi-scale summary of the data, which provides more information than only the ranks of the individual homology groups. A catchy one-liner for this idea is that "the homology of a sequence is worth more than a sequence of homologies" [41].

# 2 Quivers and Barcodes

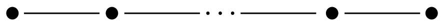
Under some mild assumptions, there are effective ways to visualize the persistent homology of a sequence. They are called *persistence diagrams* or *barcodes*. We describe the latter using notions from representation theory. The content of this section is a shortened version of the recent exposition by Oudot [58].

**Quivers and representations.** A *quiver* is a directed multigraph with *nodes* and *arrows*. A quiver is called *finite* if both the number of nodes and arrows is finite. Here are two examples of quivers



(1)

A finite quiver is called $A_n$*-type* if after removing all its arrowheads, it takes the form:



(2)

For a fixed quiver $Q$ with node set $V$ and arrow set $A$ and a base field $\mathbb{F}$, a *representation* $V = ((V_i)_{i \in V}, (f_{ij})_{ij \in A})$ is an assignment of a $\mathbb{F}$-vector space $V_i$ to each node $i$ of $Q$ and a linear map $f_{ij} : V_i \to V_j$ to each arrow from $i$ to $j$. There are no further conditions on the resulting diagram of vector spaces and linear maps, in particular, the maps do not have to commute. A representation is called *finite-dimensional*, if $\dim V_i < \infty$ for all nodes $i$. The simplest example of a representation is the trivial one, assigning the trivial vector space to every node.

Our motivating example originates from a sequence

$$S_1 \hookrightarrow S_2 \hookrightarrow \ldots \hookrightarrow S_{n-1} \hookrightarrow S_n$$

of growing shapes, for example representing a given data set for scales $\alpha_1 < \alpha_2 < \ldots < \alpha_n$. Applying $\mathbb{F}$-homology for fixed dimension $p$ yields a sequence of vector spaces and linear maps

$$H_p(S_1) \xrightarrow{h_1} H_p(S_2) \xrightarrow{h_2} \ldots \xrightarrow{h_{n-1}} H_p(S_{n-1}) \xrightarrow{h_n} H_p(S_n) \qquad (3)$$

which is a representation of an $A_n$-type quiver with all arrows directed to the right. While we focus on finite quivers in this article, the theory can be extended to the infinite case as explained in [58].

Having two representations $V$ and $W$ of the same quiver, we can form another representation $V \oplus W$ naturally by taking the direct sums of vector spaces and linear maps over every node and arrow. Vice versa, we call a representation $V$ *indecomposable*, if $V = W_1 \oplus W_2$ implies that $W_1$ or $W_2$ is the trivial representation.

**Decompositions.** Let us consider the simplest quiver $\bullet$, consisting of one node and no arrow. A finite-dimensional representation is simply a finite-dimensional vector space, and thus isomorphic to $\mathbb{F}^k = \mathbb{F} \oplus \ldots \oplus \mathbb{F}$ for some $k$. Thus, every representation decomposes into a unique direct sum of indecomposable elements up to isomorphism, and the only indecomposable representation is $\mathbb{F}$. For more general quivers, it turns out that the former statement remains valid, while the classification of indecomposable elements is more involved.

Before we can state the result, we have to define isomorphisms of representations in general. A *morphism* $\phi$ between two representations $V = (V_i, f_{ij})$ and $W = (W_i, g_{ij})$ of the same quiver $Q$ is a collection of linear maps $\phi_i : V_i \to W_i$ such that for any arrow from $i$ to $j$ in $Q$, the diagram

$$\begin{tikzcd} V_i \arrow[r, "f_{ij}"] \arrow[d, "\phi_i"] & V_j \arrow[d, "\phi_j"] \\ W_i \arrow[r, "g_{ij}"] & W_j \end{tikzcd} \tag{4}$$

commutes. A morphism is called *isomorphism* if each $\phi_i$ is an isomorphism of vector spaces. The following theorem, attributed to Krull, Remak, and Schmidt, settles the existence and uniqueness of a decomposition of finite representations.

**Theorem 1.** *Let $V$ be a non-trivial, finite-dimensional representation of a finite quiver. Then, $V = V_1 \oplus \ldots \oplus V_k$, where each $V_i$ is non-trivial and indecomposable. This decomposition is unique up to permutations and isomorphism.*

What are the indecomposable representations of a quiver? It turns out that for $A_n$-type quivers, the situation is well-behaved. This result is due to Gabriel [39].

**Theorem 2.** *Let $V$ be an indecomposable, finite-dimensional representation of an $A_n$-quiver. Then, $V$ is isomorphic to the representation $I_{b,d}$, with $1 \leq b \leq d \leq n$, which is*

$$\underbrace{0 \overset{0}{\text{——}} \cdots \overset{0}{\text{——}} 0}_{b-1} \overset{0}{\text{——}} \underbrace{\mathbb{F} \overset{id}{\text{——}} \cdots \overset{id}{\text{——}} \mathbb{F}}_{d-b+1} \overset{0}{\text{——}} \underbrace{0 \overset{0}{\text{——}} \cdots \overset{0}{\text{——}} 0}_{n-d}$$

In particular, every representation satisfying the requirements of the theorem can be characterized as a finite collection of intervals. We call this collection of intervals the *barcode* of the representation.

**Persistent barcodes.** What do these results imply for the homology sequence in (3)? A simple observation is that the barcode reveals the Betti number of $H_p(S_i)$ for all $i$, just by counting the number of intervals that span over $i$. But equally, the persistent Betti numbers are also encoded in the barcode: for $i < j$, let $\beta_{ij} = \operatorname{rank} \operatorname{Im} f$, where $f : H_p(S_i) \to H_p(S_j)$ is induced by the inclusion map $S_i \hookrightarrow S_j$. By functoriality, $f = h_{j-1} \circ \ldots \circ h_i$, and consequently, $\beta_{ij}$ equals the number of intervals in the barcode that span over the whole range $[i, j]$. Vice versa, the persistent Betti numbers also uniquely determine the barcode: the number of indecomposables of the form $I_{b,d}$ is given by

$$\beta_{b,d} - \beta_{b-1,d} - \beta_{b,d+1} + \beta_{b-1,d+1} \tag{5}$$

by the inclusion-exclusion principle.

The intervals in the barcode can also be interpreted in intuitive geometric terms: it is instructive to imagine the sequence $S_1 \hookrightarrow \ldots \hookrightarrow S_n$ as a sequence of growing balls with a fixed set of centers. Setting $p = 2$, the barcode captures the formation of voids in this sequence of balls. An interval $[b, d]$ means that a new void comes into existence when the balls have reached the scale $\alpha_b$. This void persists until scale $\alpha_d$ where it is completely filled up, and disappears. Similar considerations are true for tunnels ($p = 1$), and connected components ($p = 0$). Figure 2 illustrates this idea for an example in the plane.

While barcodes can be defined without the use of the rather heavy machinery of quivers (for instance, using (5)), this abstract point of view has several advantages: First of all, it underlines that the concept of persistence is rather independent of homology and applies to sequences of vector spaces in general (with $\mathbb{F}$-homology being only one instance of it). More importantly, we obtain a non-trivial generalization for free. Consider the following example of a *zigzag sequence* of spaces

$$S_1 \hookrightarrow S_2 \hookrightarrow S_3 \hookleftarrow S_4 \hookrightarrow S_5 \hookleftarrow S_6.$$

We can interpret this sequence again in the context of data analysis, allowing cases where the approximation is allowed to expand or shrink when the scale parameter increases. Functoriality of homology now yields a sequence of homology groups and linear maps

$$H_p(S_1) \to H_p(S_2) \to H_p(S_3) \leftarrow H_p(S_4) \to H_p(S_5) \leftarrow H_p(S_6)$$

in the same way as before. Because the arrows point in different directions, the concept of persistent Betti numbers does not carry over to this context. However, the homology groups still form a representation of a $A_n$-type quiver. Therefore, Theorem 2 applies also to this case and ensures the existence of a barcode!

Finally, the representation-theoretic point of view sheds some light on the theory of *multidimensional persistence*, where one considers more than one scale parameter to analyze the data set. The complete version of Gabriel's theorem [39] shows that finding a compact description of persistent homology in more than one dimension becomes a delicate issue. We will discuss this in some more detail in Section 4.

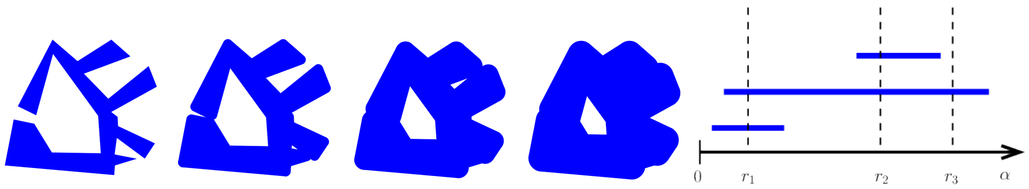


Figure 2: The 4 images on the left show snapshots of a nested sequence of shapes $S_1 \hookrightarrow S_2 \hookrightarrow \ldots \hookrightarrow S_{n-1} \hookrightarrow S_n$. Observe the formation and vanishing of holes in this process. The barcode on the right summarizes this process. Each bar (i.e., indecomposable) corresponds to a hole in the process and spans over the range of scales for which the hole is present in the data. The vertical alignment of the bars is not important. This illustration already appeared in [45].

## 3 About the history of persistence

Although persistent homology only exists for about 15 years in the literature, the substantial amount of work makes a comprehensive survey a difficult task. Moreover, any such attempt is doomed to be deprecated within short time due to the rapid evolvement of the research field. We therefore do not even aim for completeness, but rather focus on a few highlights in the theory, applications and algorithmic aspects of persistent homology. The interested reader can find more details in one of the numerous surveys on the topic [40, 10, 32, 67, 35, 66, 27]. There are also various textbooks available covering persistent homology [33, 31, 68, 41, 58].

**Theory** The term "persistent homology" was coined by Edelsbrunner, Letscher, and Zomorodian [34], who introduced persistent Betti numbers, persistence diagrams (a different, but equivalent representation of barcodes) and an efficient algorithm for filtrations of alpha shapes in the case $\mathbb{F} = \mathbb{Z}_2$. Zomorodian and Carlsson [69] extended this algorithm to arbitrary fields; moreover, they provided an algebraic description of persistence as a graded $\mathbb{F}[t]$-module, and argued that all persistent Betti numbers are determined uniquely by the module decomposition. The connection of persistence to quiver theory, as described earlier, was introduced by Carlsson and de Silva [11] to develop the concept of *zigzag persistence*.

A cornerstone for the importance of persistence is its *stability*: it means that a small perturbation of data leads only to a small change in its barcode summary; to make the statement precise, a distance measure on barcodes has to be defined, which we omit in this article. Cohen-Steiner, Edelsbrunner, and Harer [24] provided the first such stability result for the so-called *bottleneck distance*, and this result was extended by Cohen-Steiner et al. [25] to a wider family of distance measures. Once again abstracting from the geometric context, stability has been rephrased in algebraic terms by the concept of *interleavings* by Chazal et al. [16].

The survey by Carlsson [10] discusses many of these aspects and also popularized the idea of using the theory of persistent homology as a general technique for data analysis tasks. This has led to the shapening of *topological data analysis (tda)* as a new research discipline in which persistent homology is a key concept. We point out that tda is a wider area, covering aspects that are not discussed in this article, including size theory [38], Morse-Smale complexes [44], sheaf theory [26], and Reeb graphs [5]. We remark that an extension of Reeb graphs, the Mapper algorithm [63] forms the basis of the startup company Ayasdi[2], underlining the relevance of topological tools in industrial applications of data analysis.

**Applications** There is a large bandwidth of application scenarios on which persistent homology has been proved to be useful. A comprehensive list goes beyond the scope of this article, but we mention applications in coverage problems in sensor networks [28], measuring the dimension of fractal shapes [53], robust length measuring of tube-like shapes [36], the analysis of growth of rice plant roots [4], the effect of mixture of genome material in evolution [15], the effects of drug influence on brain networks [59], and the visualization of cyclical behavior of memory assignments in the execution of machine programs [23]. The recent book by Oudot [58, p.8] contains a longer (and mostly disjoint) list. We point out that the last three mentioned applications deal with data of non-geometric nature, but the data still has "shape" for which topology reveals meaningful information.

Among the numerous applications, we illustrate two major templates of how topological information is used by describing two applications in slightly more detail: Chazal et al. [20] consider the problem of clustering point clouds. Among the many approaches for this problem, *mode-seeking* methods [50] construct a density function $f$ based on the point cloud, create one cluster center per local minimum, and cluster the point set using the *basins of attraction* for each minimum (with respect to the gradient flow). A problem with this method is the instability of the clustering under small perturbation of $f$, and the authors use persistent homology to tackle this problem: using the persistent barcode defined by the function $f$, they classify the clusters into important ones and noisy ones, based on the range of scales in which a cluster is active. Then, they employ a robust variant of mode-seeking clustering where the basins of noisy clusters are charged to important ones; see [20] for more details. This is an example of a *denoising*: the topological internals of a particular data set are analyzed, allowing a simplified and more robust outcome for the given task (this was also the original motivation of introducing persistent homology from [34]).

The second template of applications uses topological information as a proxy in order to compare and classify data sets. The majority of contemporary applications falls in this category. An instructive example is given by Adcock, Rubin, and

[2] `www.ayasdi.com`

Carlsson [1], who study the task of classifying images of liver lesions into predefined categories, for the purpose of computer-assisted diagnosis. For that, they compute a barcode on an image, and compute the pairwise distances of that barcode to the barcode of a set of reference images. This defines a high-dimensional feature vector, where each coordinate is based on a topological distance. Having represented an image as a high-dimensional point, the authors use standard techniques from machine learning, such as support vector machines, for the classification task, and report on satisfying results. While this result approaches the classification task solely on topological descriptors, topology can also be used to complement other (e.g., geometric) descriptors [42, 64].

**Algorithms** A major reason for the success of persistent homology as a discipline is the existence of fast algorithms to compute the topological summary. For computations, the multi-scale representation of the data is usually written as an inclusion of combinatorial cell complexes, and is represented by the *ordered boundary matrix* of that cell complex. Persistence is computed by a simple reduction procedure that resembles Gaussian elimination. While its theoretical worst-case complexity is cubic in the size of the matrix, the algorithm shows a significantly better behavior in practice, thanks to the initial sparseness of the boundary matrix.

Because of the demand for practically efficient implementations, there is a substantial body of literature describing speed-ups of the original matrix reduction. One line of research attempts to identify shortcuts in the reduction process exploiting the special structure of boundary matrices, and achieves remarkable speed-ups with rather simple heuristics [2, 21]. These techniques have also lead to the first practical distributed algorithm to compute persistent homology [3]. Also successful has been the approach of computing persistent *cohomology* instead, relying on a duality result for persistent homology and cohomology by de Silva et al. [29]. Boissonat et al. [6] provided several optimizations of the original algorithm under the name of *annotations* [30]. Yet another way of improving is the combination of Discrete Morse Theory and persistence [43, 55]: the idea is to reduce the size of the initial simplicial complex through collapses guided by a Morse matching, and to invoke the matrix reduction algorithm solely on a matrix representation of the collapsed complex, which is often of significantly smaller size. All the aforementioned techniques have been implemented in publicly available software packages – we refer to [57] for a recent comparative survey.

The standard problem of comparing two barcodes can be reduced to a maximum-cardinality matching problem in complete bipartite graphs [33, §VIII.4]. It has been observed recently that the special (geometric) structure of barcodes can be used to significantly speed-up these computations in practice [47].

# 4 Current developments

Persistent homology has shown to be a useful tool to analyze data sets under a topological lens. Nevertheless, many questions remain unanswered both in terms of generalization and scalability. We end this article by highlighting three areas of active research which have the potential to significantly extend the range of applications of the theory.

**Multidimensional persistence** A limitation of standard persistent homology is the restriction to a single scale parameter. In many applications, one would like to filter the data along two or more axes: for instance, in the combustion example from before, we would probably prefer to consider a time-varying sequence of functions measuring temperature, and to track topological changes for progress in time as well as for changes in the threshold.

The simplest formalization of this process is a diagram of spaces and maps

$$\begin{tikzcd}
S_{m1} \arrow[r, hook] & S_{m2} \arrow[r, hook] & \dots \arrow[r, hook] & S_{mn} \\
\vdots \arrow[u, hook] & \vdots \arrow[u, hook] & & \vdots \arrow[u, hook] \\
S_{21} \arrow[u, hook] \arrow[r, hook] & S_{22} \arrow[u, hook] \arrow[r, hook] & \dots \arrow[r, hook] & S_{2n} \arrow[u, hook] \\
S_{11} \arrow[u, hook] \arrow[r, hook] & S_{12} \arrow[u] \arrow[r] & \dots \arrow[r] & S_{1n} \arrow[u]
\end{tikzcd} \tag{6}$$

where all little squares commute (the time-varying example above would better be modeled by a zigzag diagram, but we try to keep the exposition simple). Applying homology yields a representation of the quiver whose shape is the integer grid.

How much of the theory for one dimension carries over? Theorem 1 from Section 2 applies to the quiver, stating that the representation decomposes into finitely many indecomposables. However, Theorem 2 only holds for $A_n$-type quivers (and slight generalizations of it). The structure of indecomposables is way more complicated in general: there is an infinite number of isomorphisms classes, already for the case of a square-shaped quiver, which prevents a direct generalization of barcodes to higher dimensions. These difficulties with the multidimensional case have been observed first by Carlsson and Zomorodian [12] (without using quiver theory).

Despite these negative results, multidimensional persistence has received growing attention in the last years. While a complete topological invariant like the barcode in one dimension is out of reach, the primary question is which incomplete invariants can be useful for the data analysis applications. The first proposal was the *rank invariant* [12] which generalizes the persistent Betti numbers: in two dimensions, it is defined as

$$\operatorname{rank}\big(H_p(S_{ij}) \to H_p(S_{k\ell})\big)$$

for any $i \leq j, k \leq \ell$. Cerri et al. [13] have considered one-dimensional sections of the multi-dimensional filtration. In the setting of (6), any monotone path from $S_{11}$ to $S_{mn}$ defines a one-dimensional barcode, and the collection of all these barcodes is equivalent to the rank invariant. Very recently, Lesnick and Wright [52] developed a software to visualize this collection of barcodes, along with improved algorithms to compute the rank invariant.

Another research front is the efficient comparison of multidimensional representations. Lesnick [51] extended the *interleaving distance* to the multidimensional case. Chacholski et al. [14] proposed an formal algebraic definition of noise and define the distance between two representation as the minimal noise in which they differ. While both approaches are mathematically sound, no efficient algorithms to compute or at least approximate these distances are known, and no hardness results have been settled.

Because of the demand for analyzing data in multi-dimensional scale spaces, we expect further research to define, compute, visualize, and compare meaningful invariants for the case of multidimensional persistence.

**Statistical tda** A recent line of research is the combination of persistent homology and statistical methods. A central question in this context is the definition of an average of a collection of diagrams. Difficulties arise from the fact that the space of persistent barcodes has a complicated structure; while so-called *Fréchet means* of barcodes can be defined in this space, they are not unique and difficult to compute [65]. An alternative idea is to embed the space of barcodes into a larger and better behaved space, in which means are well-defined and simple to compute.

We have already discussed an example of such a strategy for the diagnosis of liver lesions [1] in Section 3. Recall that the barcode of an image was converted into a point in $\mathbb{R}^d$, constituting a transition into standard Euclidean space for which a large toolset of statistical methods applies. Another concept is that of *persistent landscapes* by Bubenik [8]. A persistence barcode is converted into a sequence of functions $\ell_i : \mathbb{R} \to \mathbb{R}$. Having two or more landscapes, averaging is easily achieved through a pointwise average of the $i$-th level functions. However, the average landscape in general cannot be translated back to a persistence barcode. Landscapes satisfy basic statistical properties such as a law of large numbers and a central limit theorem, and standard statistical methods like bootstrapping [18, 19] and subsampling [17] have been brought into the field of topological data analysis.

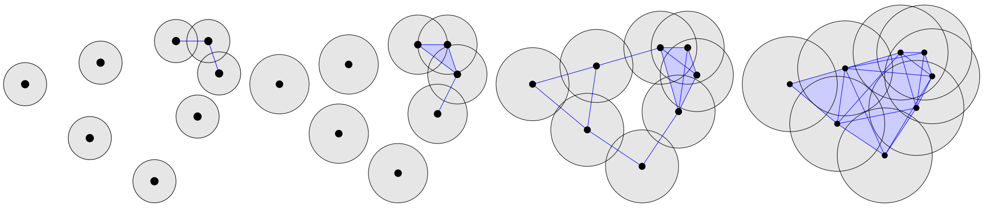


Figure 3: Illustration of the Čech filtration as the intersection complex of a union balls at various scales.

Yet another approach by Reininghaus et al. [60] defines a kernel for persistence barcodes which induces a Hilbert space structure on barcodes and permits topological classifiers in machine learning applications, such as Support Vector Machines and Principle Component Analysis. Two recent software libraries provide methods to apply statistical methods on persistence diagrams [9, 37].

We foresee further applications of statistical methods in the analysis of realistic data sets. Besides a comparison of existing techniques to embed the barcode space, plenty of algorithmic challenges need to be resolved: how can we efficiently compute and represent such an embedding? What are meaningful statistical tests, and how can they be performed efficiently in the context of persistence?

**Efficient creation of cell complexes** The first step in the computational pipeline of persistent homology is the generation of a sequence of shapes, representing the input data on different scales. We remind the reader of the popular example of point clouds, and their approximation by a union of balls, whose radius increases throughout the sequence. For computational purposes, it is common to dualize the construction, and to consider the *nerve* of the balls, which is a simplicial complex that captures the intersection patterns of the balls, called the Čech complex (Figure 3). The major drawback is the sheer size of this complex: for $n$ points in $\mathbb{R}^d$, it grows to a size of $O(n^{d+1})$ simplices, too much for realistic applications already when $d$ is small.

For low dimensions, especially $d = 2$ and $d = 3$, the complex size can be reduced by the use of *alpha complexes* [33], forming a subset of the Delaunay triangulation of the point set. But this improvement does only slightly improve the asymptotic bound for high dimensions (to $O(n^{\lceil d/2 \rceil})$) and raises computational questions since computing Delaunay triangulations in high dimensions is a non-trivial task.

A promising direction is to use geometric approximation techniques to approximate cell complexes: instead of computing a homotopically equivalent representation of desired shapes, the goal is to find approximate complexes which are significantly smaller in size, but with a provable guarantee of closeness of the exact and approximate persistent barcode. Sheehy [61] gave the first construction for the related *Vietoris-Rips* complexes with a size of $O(n \cdot 2^{d^2})$ (the precise bound is more fine-grained, but we restrict to the worst-case estimate for brevity) for an arbitrary fixed constant approximation quality $\varepsilon$. Similar results for Rips and Čech complexes with the same asymptotics have been derived subsequently [7, 30, 49]. Because of the decoupling of $n$ and $d$ in the bound, these techniques have the potential to broaden the range of data sets for which persistence can be applied. The practical evaluation of these techniques is one of the major challenges of algorithmic topology within the next years.

There is also a line of research dealing with very high-dimensional input (i.e., if $d$ is in the same order as $n$). In this case, the aforementioned approaches do not improve the naive construction. Instead, dimension reduction techniques have been considered. The celebrated Johnson-Lindenstrauss lemma [46] states that a point cloud in high-dimensionsal Euclidean space can be embedded into $O(\log n)$ dimensions with arbitrary small distortion. As shown by Sheehy [62] and by Kerber and Raghvendra [48], this property extends in the following way: the Čech complex of a point set in high dimensions yields a persistent barcode that is close to the barcode of the same point set projected to $O(\log n)$ dimensions.

Very recently, Choudhary, Kerber, and Raghvendra [22] developed a new approximation technique that yields an approximation complex with only $O(n \cdot 2^{d \log d})$ simplices, at the price of a weaker approximation guarantee. Combined with dimension reduction techniques, their results yield an approximation complex whose size is $n^{O(1)}$, independent of the dimensionality $d$ of the point set.

"Big data" is one of the buzzwords of our time – how can we design algorithms that are able to cope with the increasing volume of acquired data? Approximation techniques appear to be the most promising paradigm to process the immense amounts of data in a reasonable time. The aforementioned efforts can be interpreted as an attempt of transferring these technique into the context of tda. The question of how far this transfer will go has to be carried out by research in the upcoming years.

**Acknowledgments** I thank Aruni Choudhary and Hannah Schreiber for their helpful suggestions.

# References

[1] A. Adcock, D. Rubin, and G. Carlsson. Classification of hepatic lesions using the matching metric. *Computer Vision and Image Understanding*, 121:36–42, 2014.

[2] U. Bauer, M. Kerber, and J. Reininghaus. Clear and compress: Computing persistent homology in chunks. In *Topological Methods in Data Anal-*

*ysis and Visualization III*, Mathematics and Visualization, pages 103–117. Springer, 2014.

[3] U. Bauer, M. Kerber, and J. Reininghaus. Distributed computation of persistent homology. In *Workshop on Algorithm Engineering and Experiments (ALENEX)*, pages 31–38, 2014.

[4] P. Bendich, H. Edelsbrunner, and M. Kerber. Computing robustness and persistence for images. *IEEE Transactions on Visualization and Computer Graphics*, 16(6):1251–1260, 2010.

[5] S. Biasotti, D. Giorgi, M. Spagnuolo, and B. Falcidieno. Reeb graphs for shape analysis and applications. *Theoretical Computer Science*, 392(13):5 – 22, 2008. Computational Algebraic Geometry and Applications.

[6] J. Boissonnat, T. Dey, and C. Maria. The compressed annotation matrix: An efficient data structure for computing persistent cohomology. In *European Symposium on Algorithms (ESA)*, pages 695–706, 2013.

[7] M. Botnan and G. Spreemann. Approximating persistent homology in Euclidean space through collapses. *Applied Algebra in Engineering, Communication and Computing*, 26(1-2):73–101, 2015.

[8] P. Bubenik. Statistical topological data analysis using persistence landscapes. *The Journal of Machine Learning Research*, 16:77–102, 2015.

[9] P. Bubenik and P. Dlotko. A persistence landscapes toolbox for topological statistics. *arXiv*, abs/1501.00179, 2015.

[10] G. Carlsson. Topology and data. *Bulletin of the AMS*, 46:255–308, 2009.

[11] G. Carlsson and V. de Silva. Zigzag persistence. *Foundations of Computational Mathematics*, 10(4):367–405, 2010.

[12] G. Carlsson and A. Zomorodian. The theory of multidimensional persistence. *Discrete & Computational Geometry*, 42(1):71–93, 2009.

[13] A. Cerri, B. Di Fabio, M. Ferri, P. Frosini, and C. Landi. Betti numbers in multidimensional persistent homology are stable functions. *Mathematical Methods in the Applied Sciences*, 36:1543–1557, 2013.

[14] W. Chacholski, A. Lundman, R. Ramanujam, M. Scolamiero, and S. Öberg. Multidimensional persistence and noise. *arXiv*, abs/1505.06929, 2015.

[15] J. Chan, G. Carlsson, and R. Rabadan. Topology of viral evolution. *Proceedings of the National Academy of Sciences*, 110(46):18566–18571, 2013.

[16] F. Chazal, D. Cohen-Steiner, M. Glisse, L. Guibas, and S. Oudot. Proximity of persistence modules and their diagrams. In *ACM Symposium on Computational Geometry (SoCG)*, pages 237–246, 2009.

[17] F. Chazal, B. Fasy, F. Lecci, B. Michel, A. Rinaldo, and L. Wasserman. Subsampling methods for persistent homology. In *International Conference on Machine Learning (ICML)*, pages 2143–2151, 2015.

[18] F. Chazal, B. Fasy, F. Lecci, A. Rinaldo, A. Singh, and L. Wasserman. On the bootstrap for persistence diagrams and landscapes. *Modeling and Analysis of Information Systems*, 20:96–105, 2014.

[19] F. Chazal, B. Fasy, F. Lecci, A. Rinaldo, and L. Wasserman. Stochastic convergence of persistence landscapes and silhouettes. In *ACM Symposium on Computational Geometry (SoCG)*, pages 474–483, 2014.

[20] F. Chazal, L. J. Guibas, S. Y. Oudot, and P. Skraba. Persistence-based clustering in Riemannian manifolds. *Journal of the ACM*, 60(6):41:1–41:38, Nov. 2013.

[21] C. Chen and M. Kerber. Persistent homology computation with a twist. In *European Workshop on Computational Geometry (EuroCG)*, pages 197–200, 2011.

[22] A. Choudhary, M. Kerber, and S. Raghvendra. Polynomial-sized topological approximations using the permutahedron. In *Accepted for the 32nd International Symposium on Computational Geometry (SoCG)*, 2016.

[23] A. Choudhury, B. Wang, P. Rosen, and V. Pascucci. Topological analysis and visualization of cyclical behavior in memory reference traces. In *IEEE Pacific Visualization Symposium (PacificVis)*, pages 9–16, 2012.

[24] D. Cohen-Steiner, H. Edelsbrunner, and J. Harer. Stability of persistence diagrams. *Discrete & Computational Geometry*, 37:103–120, 2007.

[25] D. Cohen-Steiner, H. Edelsbrunner, J. Harer, and Y. Mileyko. Lipschitz functions have $L_p$-stable persistence. *Foundations of Computational Mathematics*, 10(2):127–139, 2010.

[26] J. Curry, R. Ghrist, and V. Nanda. Discrete morse theory for computing cellular sheaf cohomology. *Foundations of Computational Mathematics*, pages 1–23, 2015.

[27] J. M. Curry. Topological data analysis and cosheaves. *Japan Journal of Industrial and Applied Mathematics*, 32(2):333–371, 2015.

[28] V. De Silva and R. Ghrist. Coordinate-free coverage in sensor networks with controlled boundaries via homology. *International Journal on Robotics Research*, 25(12):1205–1222, 2006.

[29] V. de Silva, D. Morozov, and M. Vejdemo-Johansson. Dualities in persistent (co)homology. *Inverse Problems*, 27(12):124003+, 2011.

[30] T. Dey, F. Fan, and Y. Wang. Computing topological persistence for simplicial maps. In *ACM Symposium on Computational Geometry (SOCG)*, page 345, 2014.

[31] H. Edelsbrunner. *A Short Course in Computational Geometry and Topology*. Springer, 2015.

[32] H. Edelsbrunner and J. Harer. Persistent homology a survey. In *Surveys on Discrete and Computational Geometry: Twenty Years Later*, Contemporary Mathematics, pages 257–282. American Mathematical Society, 2008.

[33] H. Edelsbrunner and J. Harer. *Computational Topology. An Introduction.* American Mathematical Society, 2010.

[34] H. Edelsbrunner, D. Letscher, and A. Zomorodian. Topological persistence and simplification. *Discrete & Computational Geometry*, 28(4):511–533, 2002.

[35] H. Edelsbrunner and D. Morozov. Persistent homology: theory and practice. In *Proceedings of the European Congress of Mathematics*, pages 31–50, 2012.

[36] H. Edelsbrunner and F. Pausinger. Stable length estimates of tube-like shapes. *Journal of Mathematical Imaging and Vision*, 50(1-2):164–177, 2014.

[37] B. Fasy, J. Kim, F. Lecci, and C. Maria. Introduction to the R package TDA. *arXiv*, abs/1411.1830, 2014.

[38] P. Frosini and C. Landi. Size theory as a topological tool for computer vision. *Pattern Recognition and Image Analysis*, 9(4):596–603, 1999.

[39] P. Gabriel. Unzerlegbare Darstellungen I. *manuscripta mathematica*, 6(1):71–103, 1972.

[40] R. Ghrist. Barcodes: The persistent topology of data. *Bulletin of the AMS*, 45:61–75, 2008.

[41] R. Ghrist. *Elementary Applied Topology*. CreateSpace Independent Publishing Platform, 2014.

[42] C. Gu, L. Guibas, and M. Kerber. Topology-driven trajectory synthesis with an example on retinal cell motions. In *International Workshop on Algorithms in Bioinformatics (WABI)*, pages 326–339, 2014.

[43] D. Günther, J. Reininghaus, H. Wagner, and I. Hotz. Efficient computation of 3D MorseSmale complexes and persistent homology using discrete Morse theory. *The Visual Computer*, 28(10):959–969, 2012.

[44] A. Gyulassy, V. Natarajan, V. Pascucci, and B. Hamann. Efficient computation of Morse-Smale complexes for three-dimensional scalar functions. *IEEE Transactions on Visualization and Computer Graphics*, 13(6):1440–1447, Nov. 2007.

[45] D. Halperin, M. Kerber, and D. Shaharabani. The offset filtration of convex objects. In *Proceedings of the 23rd Annual European Symposium on Algorithms (ESA)*, pages 705–716, 2015.

[46] W. Johnson and J. Lindenstrauss. Extensions of Lipschitz mappings into a Hilbert space. *Contemporary Mathematics*, 26:189–206, 1982.

[47] M. Kerber, D. Morozov, and A. Nigmetov. Geometry helps to compare persistence diagrams. In *Workshop on Algorithm Engineering and Experiments, ALENEX*, pages 103–112, 2016.

[48] M. Kerber and S. Raghvendra. Approximation and streaming algorithms for projective clustering via random projections. In *Canadian Conference on Computational Geometry (CCCG)*, pages 179–185, 2015.

[49] M. Kerber and R. Sharathkumar. Approximate Čech complex in low and high dimensions. In *International Symposium on Algortihms and Computation (ISAAC)*, pages 666–676, 2013.

[50] W. Koontz, P. Narendra, and K. Fukunaga. A graph-theoretic approach to nonparametric cluster analysis. *IEEE Transactions on Computing*, 24:936–944, 1976.

[51] M. Lesnick. The theory of the interleaving distance on multidimensional persistence modules. *Foundations of Computational Mathematics*, 15(3):613–650, 2015.

[52] M. Lesnick and M. Wright. Interactive visualization of 2D persistence modules. *arXiv*, abs/1512.00180, 2015.

[53] R. MacPherson and B. Schweinhart. Measuring shape with topology. *Journal of Mathematical Physics*, 53, 2012.

[54] A. Markov. The insolvibility of the problem of homeomoprhy. *Dokl. Akad. Nauk SSSR*, 121:218–220, 1958. (Russian).

[55] K. Mischaikow and V. Nanda. Morse theory for filtrations and efficient computation of persistent homology. *Discrete & Computational Geometry*, 50(2):330–353, 2013.

[56] J. Munkres. *Elements of algebraic topology*. Westview Press, 1984.

[57] N. Otter, M. Porter, U. Tillmann, P. Grindrod, and H. Harrington. A roadmap for the computation of persistent homology. *arXiv*, abs/1506.08903, 2015.

[58] S. Oudot. *Persistence theory: From Quiver Representation to Data Analysis*, volume 209 of *Mathematical Surveys and Monographs*. American Mathematical Society, 2015.

[59] G. Petri, P. Expert, F. Turkheimer, R. Carhart-Harris, D. Nutt, P. J. Hellyer, and F. Vaccarino. Homological scaffolds of brain functional networks. *Journal of the Royal Society Interface*, 11(101), 2014.

[60] J. Reininghaus, S. Huber, U. Bauer, and R. Kwitt. A stable multi-scale kernel for topological machine learning. In *IEEE Conference on Computer Vision & Pattern Recognition (CVPR)*, 2015.

[61] D. Sheehy. Linear-size approximation to the Vietoris-Rips filtration. *Discrete & Computational Geometry*, 49:778–796, 2013.

[62] D. Sheehy. The persistent homology of distance functions under random projection. In *ACM Symposium on Computational Geometry (SoCG)*, 2014.

[63] G. Singh, F. Memoli, and G. Carlsson. Topological Methods for the Analysis of High Dimensional Data Sets and 3D Object Recognition. In *Eurographics Symposium on Point-Based Graphics*. The Eurographics Association, 2007.

[64] C. N. Topp, A. S. Iyer-Pascuzzi, J. T. Anderson, C.-R. Lee, P. R. Zurek, O. Symonova, Y. Zheng, A. Bucksch, Y. Mileyko, T. Galkovskyi, et al. 3D phenotyping and quantitative trait locus mapping identify core regions of the rice genome controlling root architecture. *Proceedings of the National Academy of Sciences*, 110(18):E1695–E1704, 2013.

[65] K. Turner, Y. Mileyko, S. Mukherjee, and J. Harer. Fréchet means for distributions of persistence diagrams. *Discrete & Computational Geometry*, 52(1):44–70, 2014.

[66] M. Vedjemo-Johansson. Sketches of a platypus: a survey of persistent homology and its algebraic foundations. *Contemporary Mathematics*, 620, 2014.

[67] S. Weinberger. What is ... persistent homology. *Notices of the AMS*, 58:36–39, 2011.

[68] A. Zomorodian. *Topology for Computing*. Cambridge University Press, 2009.

[69] A. Zomorodian and G. Carlsson. Computing persistent homology. *Discrete & Computational Geometry*, 33(2):249–274, 2005.

*Author's address: Michael Kerber, Institute of Geometry, Technische Universität Graz. Kopernikusgasse 24, A-8010 Graz. email kerber@tugraz.at.*