Persistent Homology

A Topic in Topological Data Analysis

Kindness Chen [1], Yejin Lee [1,2], Parker Sherry [1,3]

[1] Department of Mathematics and Statistics

[2] Department of Physiology

[3] Department of Physics

Directed Reading Program
McGill University
July 1, 2024

Contents

Introduction

The explosive growth of data collection has necessitated the development of effective tools for data analysis. Topological Data Analysis (TDA) is a relatively new branch of statistics that emerged in the late 2000s. Mathematics being the language of sciences, TDA borrows concepts from algebraic topology and functions as a powerful tool for extracting meaningful geometric insights from complex real-life datasets. Of particular interest is the study of persistent homology, which quantifies the intuitive notions of the number and sizes of “holes” in a scatterplot of points representing data value tuples.

Section 1 outlines the theoretical background on which persistent homology is based, and introduces the reader to relevant definitions of topological spaces, simplicial complexes, and homology. Sections 1.4 and 1.5 cover the application of the previous concepts to arbitrary datasets in $\mathbb{R}^n$ via the Čech and Vietoris-Rips complexes, the persistence diagram, and the barcode, along with a formalization of “nested” complexes (i.e. filtration).

Section 2 makes explicit the mathematical concepts which underpin one of the greatest practical advantages of persistent homology, namely that of being stable against noise. The culmination of this section is the stability theorem for finite sets in $\mathbb{R}^n$.

Section 3 gives examples of how persistent homology is applied to real datasets; in particular, the barcodes of a distance matrix and the molecular conformations of cyclooctane are computed, and the method of applying persistent homology to biological aggregations is discussed.

Keywords: Simplicial complex, filtration value, Vietoris-Rips complex, persistent barcode, Betti number

1 Theoretical Background

1.1 Topology

1.1.1 Definition (Topology)

Let X be a nonempty set. For $\tau \subseteq \mathcal{P}(X)$, we say that τ is a **topology** on X if

1. X, $\emptyset \in \tau$
2. τ is closed under taking unions: $\tau_i \in \tau \implies \bigcup_{i \in I} \tau_i \in \tau$ for I a (potentially infinite) index set.
3. τ is closed under *finite* intersection: $\tau_i \in \tau \implies \bigcap_{i \in I} \tau_i \in \tau$ for I a finite index set.

The pair (X,τ) is called a **topological space**. A set S is **open** in τ if it is a member of τ; if the complement of S, $X \setminus S$, is open in τ, we say that S is **closed**.

Though not necessarily apparent from the definition given above, topology is concerned with the properties of geometric objects that are preserved through deformations. As a quantitative tool, topology lends itself well to analyzing sets which consistently exhibit certain kinds of geometric behaviour in spite of small perturbations. In the context of TDA, what is meant by "small perturbations" will be clarified in Section 2.

Since data values are often plotted in a space where the notion of distance has some geometric meaning, it will also be worth it to introduce the concept of a metric space and the topology it generates; later, we will make use of these concepts to make inferences about the persistence of geometric features captured using TDA.

1.1.2 Definition (Metric Space)

A **metric space** is a set X equipped with a **distance function** d, denoted (X, d). For any two elements $x, y \in X$, we require $d(x, y) \in \mathbb{R}^+$; i.e., the distance between any two elements is a non-negative real number. The following properties must also hold for d, for all $x, y, z \in X$:

1. $d(x, y) = 0 \iff x = y$
2. $d(x, y) = d(y, x)$
3. (Triangle Inequality) $d(x, y) + d(y, z) \geq d(x, z)$

Example 1: The *Euclidean* metric $d : \mathbb{R}^n \to \mathbb{R}$ assigns to two vectors x and y in Euclidean n-space the nonnegative real number

$$d(x, y) = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2}$$

If $n = 2$, one obtains the familiar Pythagorean Theorem.

Example 2: For two points P and Q on the surface of a sphere, both the straight line distance *and* the spherical distance between them represent viable distance functions.

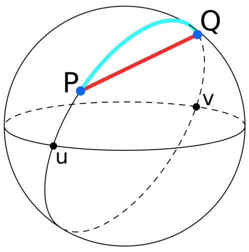


Figure 1: Straight line versus spherical distance between two points on the surface of a sphere. [18]

1.1.3 Definition (Metric Topology)

For a metric space (X, d), the **metric topology** on X induced by d is the topology τ whose open sets are the balls $B(x, \epsilon) := \{y \in X : d(x, y) < \epsilon\}$ for any $x \in X$ and for any $\epsilon > 0$, and unions of these balls. We say that the balls $B(X, \epsilon)$ form the **basis** of the metric space topology τ.

We will also make use of the topology induced by a subset of X; for convenience, we define it here.

1.1.4 Definition (Subspace Topology)

Given a topological space X with topology τ and a subset $Y \subseteq X$, Y is also a topological space with the **subspace topology**, where a set $U \subseteq Y$ is open in Y if and only if $U = Y \cap O$ for some open set $O \subseteq X$. In particular, if X is a metric space with distance function d and its generated metric space topology d, then the metric space topology generated by d on Y and the subspace topology on Y are the same.

Throughout this paper, we adopt the convention of referring to the metric space topology on $\mathbb{R}^n$ or any of its subsets with the Euclidean metric as just "the topology" of the set. Also, throughout all subsequent sections, we assume that all sets are subsets of $\mathbb{R}^n$.

1.2 Simplex, Simplicial Complex, Geometric Realization

We now want to introduce the geometric structure which we will use to assemble our data points into a topological object. We begin by defining a purely symbolic collection of "points" along with the set-theoretic rules which inform how they connect to each other, before going on to explain how they generate a corresponding topological space.

1.2.1 Definition (Abstract Simplicial Complex)

Given a set $V = \{v_0, v_1, ..., v_{n-1}, v_n\}$, a *finite* **abstract simplicial complex** (ASC) or just *simplicial complex*, denoted K, is a subset of $\mathcal{P}(V)$ (the power set of V) such that any subset of an element of K is also in K. Formally, we require

1. $\sigma \in K \wedge \tau \subseteq \sigma \implies \tau \in K$
2. $\forall v_i \in V, \{v_i\} \in K$

where the latter condition requires all the elements in V to appear as singletons in K.

Example: Let $V = \{a, b, c, d, e\}$. One valid ASC of V would be the set

$$K = \{\{a\}, \{b\}, \{c\}, \{d\}, \{e\}, \{a, b\}, \{b, c\}, \{a, b, c\}\}$$

1.2.2 Definition (Subcomplex)

A **subcomplex** of a simplicial complex K refers to a subset L of K which also is a simplicial complex. These subcomplexes allow us to represent a small portion of a given complex.

Example: In the above example, a valid subcomplex of K would be the ASC

$$L = \{\{a\}, \{b\}, \{c\}, \{d\}, \{e\}, \{a, b\}\}$$

1.2.3 Definition (Simplex)

Let K be an ASC over a finite set $V = \{v_0, v_1, ..., v_{n-1}, v_n\}$, and let $\Delta^k \subset V$ be the power set of $k+1$ elements $v_0, v_1, \ldots, v_k$ such that $\{v_0, v_1 ..., v_k\} \in K$. We say that Δ^k describes a **k-simplex**, and we denote $\Delta^k = [v_0, v_1, \ldots, v_k]$.

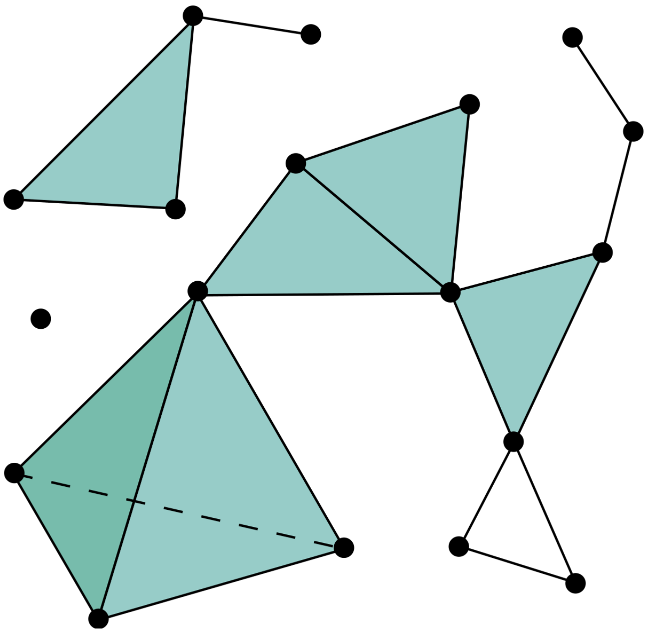


Figure 2: The geometric realization of an abstract simplicial complex. [20]

In essence, a simplex is the n-dimensional generalization of a polygon (e.g. polyhedron), and for this reason the set upon which K is formed is often referred to as the set of **vertices** of K. In this light, a simplicial complex can be thought of as a geometric structure consisting of many different kinds of simplices (lines, tetrahedrons, etc.). Often, simplicial complexes are embedded in $\mathbb{R}^n$, although analogues in hyperbolic space ($\mathbb{H}^n$) and the n-dimensional sphere ($\mathbb{S}^n$) exist. The focus of this paper will be restricted to Euclidean space, and the underlying topology will be the topology on $\mathbb{R}^n$.

Also of importance here is the fact that the elements of V are ordered. For now, this is arbitrary, but when we begin to introduce other concepts from algebraic topology it will be important to keep this ordering in mind.

1.2.4 Definition (Boundary of a Simplex)

Let Δ^n be an n-simplex, which we know is described by $n+1$ vertices. Since $\mathcal{P}(\Delta^n) \subseteq K$, by removing some vertex v_i we obtain a set which describes an $n-1$-simplex. This simplex, Δ_i^{n-1}, corresponds to one of the **faces** of Δ^n, and the union of all such simplices corresponds to the **boundary** of Δ^n, denoted $\delta\Delta^n$.

1.2.5 Definition (Geometric Simplicial Complex, Geometric Realization)

Let K be an ASC. The **geometric simplicial complex** (GSC) is the topological space generated by K, denoted $|K|$. Specifically, a GSC is a pair $(|K|, F)$, where $|K| \subset \mathbb{R}^n$ is the underlying space and F is a set of functions which map the simplices in K to $|K|$. For a function $f \in F$, we require that

1. The restriction of f to the interior[1] of Δ^n is injective
2. The restriction of f to one of the faces of Δ^n can be identically expressed as $f : \Delta_i^{n-1} \to |K|$
3. A set $U \subseteq |K|$ is open if and only if $f^{-1}(U)$ is open for all f

$|K|$ is the union of the images of all such f, and we call it the **geometric realization** of K.

Note that while the underlying topology has not been made explicit, given $|K| \subset \mathbb{R}^n$ and τ the standard topology on $\mathbb{R}^n$, we have that τ induces a subspace topology on $|K|$. It is also worth it to mention that every two geometric realizations of the same ASC, even in Euclidean spaces of different dimensions, are homeomorphic. Therefore, given an ASC K, one can speak of the *unique* geometric realization, $|K|$. [9]

For a deeper characterization of simplices and simplicial complexes, see [7].

1.3 Simplicial Homology

The classic joke about topologists is that they are unable to distinguish between a donut and a coffee cup. While there are many reasons one can give to explain why an ill-advised mathematician would make this mistake, one reason which is quite relevant to the interests of this paper is that the two objects have *homology groups* which coincide.

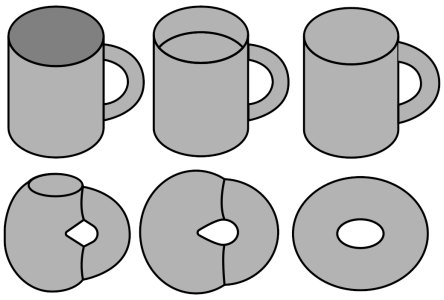


Figure 3: Half a dozen donuts. [19]

Loosely, homology is a technique used to enumerate and classify the holes in a topological object. This is done by constructing a correspondence between an algebraic object and the set of k-simplices in our simplicial complex. In analogy with the connection between donuts and coffee cups, homology allows us to draw parallels between different sets of data points which are nonetheless *topologically similar*. The construction of a set's homology group proceeds as follows.

1.3.1 Formal Sum

Let B be a finite set with n elements. A **finite formal sum** over B is a linear combination of the elements of B of the form

$$\sum_{i=1}^{n} c_i b_i, \qquad c_i \in \mathbb{Z},\ b_i \in B$$

where by $c_i b_i$ we mean:

1. if $c_i > 0$ then $c_i b_i = \underbrace{b_i + \cdots + b_i}_{c \text{ times}}$
2. if $c_i < 0$ then $c_i b_i = \underbrace{b_i^{-1} + \cdots + b_i^{-1}}_{c \text{ times}}$
3. if $c_i = 0$ then $c_i b_i = 0$, the empty sum

1.3.2 Definition: Simplicial n-Chains and C_n

Let K be an ASC, let $|K|$ be its geometric realization, and let there be k n-simplices in K. A **simplicial n-chain** is a finite formal sum of the form

$$\sum_{i=1}^{k} c_i (\Delta^n)_i, \qquad \Delta^n \text{ an } n\text{-simplex}^{2}$$

---

[1] This criterion will not be very important to the ideas we put forth in the subsequent sections of this paper, so the notion of the "interior" of a simplex as it has been described is somewhat glossed over. Hatcher defines a simplex using a weighted sum of coordinates $V = \{v_1, ..., v_n\} \in \mathbb{R}^n$, for which the interior can be thought of as the convex hull of V. [7]

We denote the group of such n-chains as C_n, or

$$C_n := \left\{ \sum_{i=1}^{k} c_i(\Delta^n)_i : c_i \in \mathbb{Z},\ \Delta^n \text{ an } n\text{-simplex} \right\}$$

It can be readily seen that C_n is the free abelian group over basis K. By convention, we have that the interchanging of two vertices flips the sign of the simplex in the chain, i.e. $[v_0, ..., v_i, ..., v_j, ..., v_n] = -[v_0, ..., v_j, ..., v_i, ..., v_n]$.

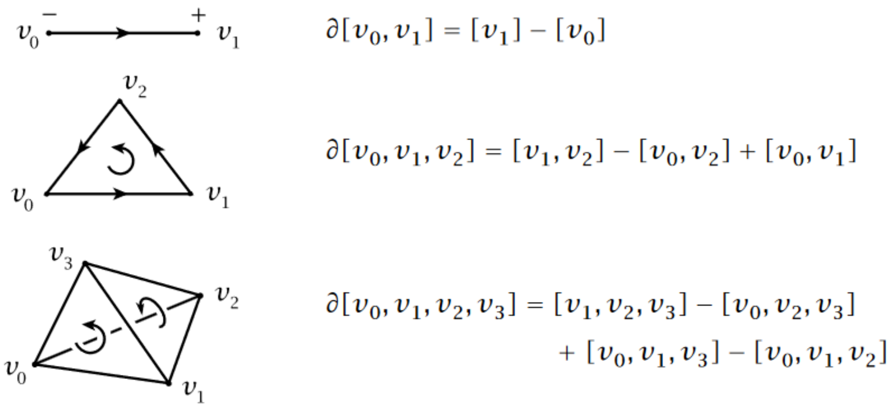
The basis of C_n is in one-to-one correspondence with the set of n-simplices in K, which is itself in correspondence with its image in $|K|$. By defining n-chains in this way, we effectively set up an analogy between arithmetic combinations of simplices in C_n and "traversing" an ordered path of simplices in $|K|$. This is illustrated in Fig.4, though a full appreciation of the image requires the addition of one more concept, that of the boundary map.

1.3.3 Definition (Boundary Map)

The n**th boundary map** $\partial_n : \mathcal{C}_n \to \mathcal{C}_{n-1}$ is a homomorphism defined as

$$\partial_n(\Delta^n) := \sum_i (-1)^i \Delta_i^{n-1}$$

i.e., the map sending an n-simplex to the alternating formal sum of its faces. For $n = 0$, we define $\partial_0(\Delta^n) := 0$, i.e. the zero map.



$$\partial[v_0, v_1] = [v_1] - [v_0]$$

$$\partial[v_0, v_1, v_2] = [v_1, v_2] - [v_0, v_2] + [v_0, v_1]$$

$$\partial[v_0, v_1, v_2, v_3] = [v_1, v_2, v_3] - [v_0, v_2, v_3] + [v_0, v_1, v_3] - [v_0, v_1, v_2]$$

Figure 4: Geometric interpretation of the boundary map. Formal sums correspond to "paths" defined by the orientation of the the vertices of the underlying simplex. [7].

1.3.4 Lemma (Composition of Boundary Maps)

Claim: The composition $\partial_n \circ \partial_{n+1}$ gives the zero map.

Proof. The following derivation is given by Hatcher [7]. We have

$$\begin{aligned}
(\partial_n \circ \partial_{n+1})(\Delta^{n+1}) &= \partial_n(\sum_i (-1)^i \Delta_i^n) \\
&= \sum_{i,j} (-1)^{i+j} \Delta_{ij}^{n-1}, \qquad i \neq j \\
&= \sum_{i<j} (-1)^{i+j} \Delta_{ij}^{n-1} + \sum_{j<i} (-1)^{i+j} \Delta_{ji}^{n-1} \\
&= \sum_{i<j} (-1)^{i+j} \Delta_{ij}^{n-1} + \sum_{j<i} (-1)^{i+j+1} \Delta_{ij}^{n-1} \\
&= 0
\end{aligned}$$ □

In the jump from the second-to-last line to the last line we have used the fact that ∂_n is a homomorphism, i.e.,

$$\begin{aligned}
\partial_n([v_0, ..., v_i, v_j, ..., v_n]) + \partial_n([v_0, ..., v_j, v_i, ..., v_n]) &= \partial_n([v_0, ..., v_i, v_j, ..., v_n] + [v_0, ..., v_j, v_i, ..., v_n]) \\
&= \partial_n([v_0, ..., v_i, v_j, ..., v_n] - [v_0, ..., v_i, v_j, ..., v_n]) \\
&= 0
\end{aligned}$$

This result tells us something interesting: if we apply the boundary map to *the boundary* of a simplex, it is sent to zero. This implies that the boundary of any $n+1$-simplex is contained in the kernel of the n^{th} boundary map, i.e.

$$\delta\Delta^{n+1} \subset ker(\partial_n)$$

We refer to the elements of this subgroup as **cycles**. Geometrically, they correspond to closed paths one can take along n-dimensional simplices, starting and ending at the same simplex.

---

[2] Do not be confused by notation: whereas in the previous section we defined Δ_i^{n-1} to be the face of Δ^n generated by removing its i^{th} vertex, $(\Delta^n)_i$ is defined here to be the ith n-simplex in K, for some arbitrary ordering of simplices.

1.3.5 Definition (Homology Group, Betti Number, Homology)

Let K be an ASC with a geometric realization, and let $N \in \mathbb{N}$ be the highest dimension of any simplex in K. Let C_n be the group of n-chains generated by the n-simplices in K for $n = 0, 1, ..., N$. Let ∂_n be the boundary map which sends C_n to C_{n-1} for any such n. The n**th homology group** of K is defined as

$$H_n(K) := \frac{ker(\partial_n)}{im(\partial_{n+1})}$$

The reasoning behind this construction is as follows: an n-dimensional "hole" in a simplicial complex should be given by a closed cycle of n-simplices which do not enclose any n+1-simplices: in other words, we are looking for **cycles which are not boundaries**.
If we define the n**th Betti number** as

$$Betti_n(K) := \text{rank}(H_n(K)),$$

then $Betti_n$ denotes the number of n-dimensional holes in K. Note that, for $n = 0$, we have $ker(\partial_0) = C_0$, while $im(\partial_1)$ is the set of 0-chains which correspond to connected vertices. This implies that $Betti_0$ denotes the number of *connected components* in K.

The **homology** of K is defined as the set of homology groups

$$\mathcal{H}(K) := \{H_0(K), H_1(K), ..., H_N(K)\} \qquad [6]$$

1.4 Persistent Homology

Simplicial homology can be a powerful tool when computing the number and dimension of holes in a simplicial complex. However, when we try to infer the existence of "holes" in a dataset embedded in $\mathbb{R}^n$, we do not have ready-made simplicial complexes to find holes in. This leads us to the concept of *persistent* homology, a technique which allows us to identify n-dimensional features which persist over a sequence of "nested" complexes, with the assumption that features which point to relevant geometric information correspond to "significant" holes.

1.4.1 Definition (ε-Fattening)

Let $A \subset X$ be a bounded set. The ε**-fattening** of A in X is defined as

$$A_\varepsilon := \bigcup_{a \in A} \{x \in X : d(x, a) \leq \varepsilon\}$$

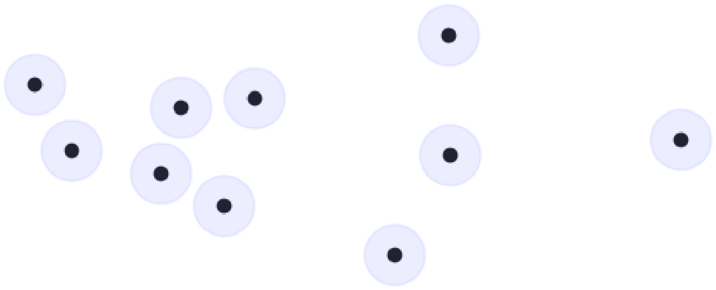


Figure 5: The ε-fattening of a set of points. Whereas the original set was restricted to the points themselves, the fattened set includes all points contained within a disk centered at a given point with radius ε.

Given a dataset, i.e. a finite set of points $A \subset \mathbb{R}^n$, the most straightforward way to quantify the existence of holes is to grow balls of increasing radius around each point in A and note the holes of each dimension at for each radius value. Formally, let $\epsilon \geq 0$. We would like to consider the homology of the ϵ-fattening A_ϵ for $\epsilon \in (0, \infty)$.

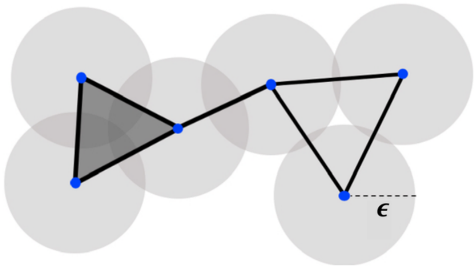


Figure 6: An ε-fattened set, denoted by the blue dots and grey disks, equipped with a simplicial complex. [12]

To compute the homology $\mathcal{H}(A_\epsilon)$, we will first compute the homology for the corresponding *Čech complex* of A at radius ϵ.

1.4.2 Definition (Čech complex)

For a finite set $A \subset \mathbb{R}^n$ and $\epsilon \geq 0$, the **Čech complex** $C(A, \epsilon)$ is the abstract simplicial complex such that a subset $\{v_0, v_1, ..., v_n\} \subseteq A$ is a simplex $\Delta^n = [v_0, v_1, ..., v_n] \in C(A, \epsilon)$ if and only if

$$\bigcap_{i=0}^{n} B(v_i, \epsilon) \neq \varnothing \qquad [6]$$

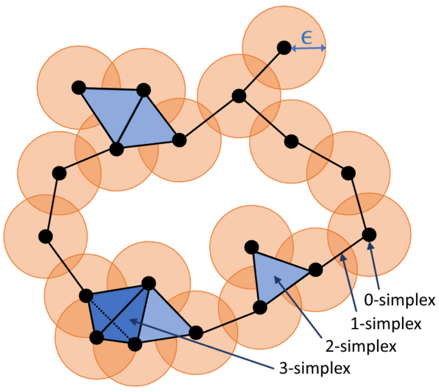


Figure 7: A set of points A, denoted by black dots; A_ϵ, denoted by the union of orange balls; and $C(A, \epsilon)$. [15]

The Čech complex is obviously of no good if its homology is not related to the homology of A_ϵ. Fortunately, we have the following lemma:

1.4.3 Nerve Lemma

Claim 1: For a finite set $A \subset \mathbb{R}^n$ and $\epsilon > 0$, the ϵ-fattening A_ϵ is homotopy equivalent (continuously deformable) to the Čech complex $C(A, \epsilon)$.
Claim 2: A_ϵ and $C(A, \epsilon)$ have the same homologies.

The reader will be directed to Appendix A.2 of [6] for more information on the proof of both claims.

If A has k data points (i.e. $|A| = k$), computing the simplices of a Čech complex $C(A, \epsilon)$ is more expensive than computing the simplices of another kind of complex, the Vietoris-Rips Complex. We define it here.

1.4.4 Definition (Vietoris-Rips Complex)

For a finite set $A \subset \mathbb{R}^n$ and $\epsilon \geq 0$, the Vietoris-Rips complex $VR(A, \epsilon)$ is the abstract simplicial complex such that a subset $\{v_0, v_1, ..., v_n\} \subseteq A$ is a simplex $\Delta^n = [v_0, v_1, ..., v_n] \in VR(A, \epsilon)$ if and only if, for the Euclidean distance function $d : \mathbb{R}^n \to \mathbb{R}$ and any $v_i, v_j \in \{v_0, v_1, ..., v_n\}$,

$$d(v_i, v_j) \leq 2\epsilon \qquad [6]$$

Although the Vietoris-Rips complex is not always equivalent to the Čech complex with the corresponding parameters, it is a computationally friendly approximation ([6]). Moreover, the Vietoris-Rips complex is often a good approximation, as the next lemma shows.

1.4.5 Lemma: Čech and Vietoris-Rips Sandwich Lemma

Claim: For a finite set $A \subset \mathbb{R}^n$ and $\epsilon > 0$, we have

$$C(A, \epsilon) \subset VR(A, \epsilon) \subset C(A, \epsilon\sqrt{2}) \qquad [6]$$

Proof. The inclusion $C(A, \epsilon) \subset VR(A, \epsilon)$ is trivial. Consider a simplex $\Delta^n = [v_0, v_1, ..., v_n] \in C(A, \epsilon)$. Then by definition, $\bigcap_{i=0}^{n} B(v_i, \epsilon) \neq \varnothing$. Hence for any $v_i, v_j \in \Delta^n$, we have that $B(v_i, \epsilon) \cap B(v_j, \epsilon) \neq \varnothing$, which implies that there is a point $z \in B(v_i, \epsilon) \cap B(v_j, \epsilon)$. Since $d(v_i, z) \leq \epsilon$ and $d(v_j, z) \leq \epsilon$, we have $d(v_i, v_j) \leq d(v_i, z) + d(v_j, z) \leq 2\epsilon$, for any $v_i, v_j \in \Delta^n$. Hence by definition, $\Delta^n \in VR(A, \epsilon)$, and hence $C(A, \epsilon) \subset VR(A, \epsilon)$.

(The rest of the proof is adapted from a co-author's self-response to their question on Math Overflow ([8]).)

The inclusion $VR(A, \epsilon) \subset C(A, \epsilon\sqrt{2})$ can be easily deduced from Jung's Theorem, which states that the maximum Euclidean distance D between any two points x, y in a *compact* (e.g. finite) set $A \subset \mathbb{R}^n$ is related to the circumradius R of A (the smallest closed n-dimensional sphere containing A) by

$$D \geq R\sqrt{\frac{2(n+1)}{n}} \qquad [10]$$

The inclusion $VR(A, \epsilon) \subseteq C(A, \epsilon\sqrt{2})$ is equivalent to the following statement:

"For any $\{v_0, v_1, \ldots, v_n\} \subseteq A$, if $\{v_0, v_1, \ldots, v_n\}$ forms a simplex in $VR(A, \epsilon)$, then $\{v_0, v_1, \ldots, v_n\}$ forms a simplex in $C(A, \epsilon\sqrt{2})$."

Which is again equivalent to

"For any $\{v_0, v_1, \ldots, v_n\} \subseteq A$, if all distances $d(v_i, v_j) \leq 2\epsilon$ then the union of closed balls $\bigcap_{i=0}^{n} B(v_i, \epsilon\sqrt{2}) \neq \varnothing$."

Suppose that $\{v_0, v_1, \ldots, v_n\} \subseteq A$ and $d(v_i, v_j) \leq 2\epsilon$. Then $D := 2\epsilon$ is the diameter of the set $\{v_0, v_1, \ldots, v_n\}$. By Jung's Theorem, there exists a point c and radius $R \leq D\sqrt{\frac{n}{2(n+1)}} = \epsilon\sqrt{2}\sqrt{\frac{n}{n+1}}$ such that the closed ball $B(c, R) \supseteq \{v_0, v_1, \ldots, v_n\}$. This implies that $d(v_i, c) \leq r \leq \epsilon\sqrt{2}\sqrt{\frac{n}{n+1}} < \epsilon\sqrt{2}$ for any $v_i \in \{v_0, v_1, \ldots, v_n\}$. Hence for any $v_i \in \{v_0, v_1, \ldots, v_n\}$, the closed ball $B(v_i, \epsilon\sqrt{2})$ always contains the point c. Hence $c \in \bigcap_{i=0}^{n} B(v_i, \epsilon\sqrt{2})$, implying $\bigcap_{i=0}^{n} B(v_i, \epsilon\sqrt{2}) \neq \varnothing$. □

This lemma demonstrates that the Vietoris-Rips complex is a satisfactory approximation of the Čech complex. For subsequent theoretical discussions, we will continue to refer to the Čech complex; however, computational examples will use the Vietoris-Rips complex.

Regardless of whether one uses the Čech complex or the Vietoris-Rips complex, they both describe a simplicial complex given one fixed parameter ϵ. However, the original purpose of persistent homology is to describe the homology $\mathcal{H}(A_\epsilon)$ (equivalently $\mathcal{H}(C(A, \epsilon))$, approximated by $\mathcal{H}(VR(A, \epsilon))$) as ϵ increases from 0 to ∞. We describe this changing homology in terms of filtrations.

1.4.6 Definition (Filtration)

An arrangement of subcomplexes of a simplicial complex K in order of inclusion

$$\varnothing = K^0 \subseteq K^1 \subseteq \cdots \subseteq K^m = K$$

is called a **filtration** of K. The simple complex K given this arrangement is called a filtered complex. For the generality of the index, we make these conventions:

1. If $i \geq m$, then $K^i = K^m = K$
2. If $i \leq 0$, then $K^i = K^0 = \varnothing$

The subcomplex of a subcomplex K_i is still a subcomplex of K. Repeating this process, we eventually get to the empty set $\varnothing$. Subcomplexes obtained through this process will have an inclusive relationship in one direction. In other words, filtration is a method where we list the subsets of the simplicial complex in an increasing sequence to express how the current simplicial complex was created.

For example, let the simple complex K be represented by a tetrahedron. Several subcomplexes of K are shown below.

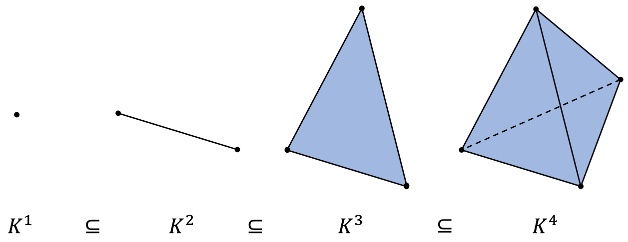


Figure 8: An example filtration of a tetrahedron. [16]

K^1 is defined as a vertex of K, K^2 is a subcomplex containing the vertices, K^3 is a subcomplex containing the sides, and K^4 is defined as K itself. The sequence

$$\varnothing = K^0 \subseteq K^1 \subseteq K^2 \subseteq K^3 \subseteq K^4 = K$$

is a filtration of K.

1.4.7 Choosing the Best Filtration

We now examine filtrations of Čech complexes for the purpose of persistent homology. (In practice, we will use Vietoris-Rips complexes in place of Čech complexes.)

For a dataset (set of points) $A \subseteq \mathbb{R}^d$ for some $d \in \mathbb{N}$, to observe the changes to $\mathcal{H}(A_\epsilon)$ as ϵ increases from 0 to ∞ requires computing the homology $\mathcal{H}(C(A, \epsilon))$ for every $\epsilon \in (0, \infty)$. Doing so being an impossible task, we instead try to compute the homology of $\mathcal{H}(C(A, \epsilon))$ for some finite set of ϵ_i's.

Let $0 \leq \epsilon_1 < \epsilon_2$ be parameters of two different Čech complexes on the same data set A. It is easily observed that $C(A, \epsilon_1)$ is a subcomplex of $C(A, \epsilon_2)$. Indeed, for closed balls centered at the same point v we have $B(s, \epsilon_1) \subset B(s, \epsilon_2)$; therefore, for a simplex $\Delta^n = [v_0, v_1, \ldots, v_n] \in C(A, \epsilon_1)$, we have

$$\begin{aligned} &\bigcap_{i=0}^{n} B(v_i, \epsilon_1) \neq 0 \\ \implies &\bigcap_{i=0}^{n} B(v_i, \epsilon_2) \neq 0 \\ \implies &\Delta^n \in C(A, \epsilon_2) \end{aligned}$$

Let $\epsilon_0 = 0, \epsilon_1, \epsilon_2, \dots, \epsilon_n$ be an increasing sequence of real numbers, where

$$\epsilon_n := \min\{\epsilon \in [0, \infty) : \bigcap_{x \in A} B(x, \epsilon) \neq \varnothing\}$$

Then by the argument above,

$$C(A, \epsilon_0) = A \subseteq C(A, \epsilon_1) \subseteq C(A, \epsilon_2) \subseteq \dots \subseteq C(A, \epsilon_n)$$

which gives us a filtration of $C(A, \epsilon_n)$.

We see that $\forall \epsilon \geq \epsilon_n$, $C(A, \epsilon) = C(A, \epsilon_n)$, the truth of which we will supply a short proof here. For any subset $A' \subseteq A$, we have $\bigcap_{x \in A'} B(x, \epsilon_n) \supseteq \bigcap_{x \in A} B(x, \epsilon_n) \neq \varnothing$ due to the definition of ϵ_n; therefore, $A' \in C(A, \epsilon)$. This implies that $C(A, \epsilon_n) = \mathcal{P}(A)$. Since we have the sandwich inclusion relationship

$$\mathcal{P}(A) = C(A, \epsilon_n) \subseteq C(A, \epsilon) \subseteq \mathcal{P}(A) \quad \forall \epsilon \geq \epsilon_n$$

we see that for all $\epsilon \geq \epsilon_n$, $C(A, \epsilon) = C(A, \epsilon_n)$.

It is important to choose $\epsilon_1, \epsilon_2, \dots, \epsilon_{n-1}$ wisely. Ideally, they should be distributed such that for every $\epsilon_{i-1} < \epsilon_i, i \in \{1, 2, \dots, n\}$ we have that $VR(A, \epsilon_i)$ is different from $VR(A, \epsilon_{i+1})$ by exactly one feature, in the sense that there is exactly one hole that either "dies" or "is born" going into $VR(A, \epsilon_i)$. Although there is no right or wrong way of defining ϵ_i's, Fasy et al. provide the following convenient convention:

1. Let $r_0 = 0$, and $r_1, \dots, r_{n-1}$ be a finite increasing sequence such that there exists $t_i \geq 0$ such that $C(A, r_i) \neq C(A, r_i - \epsilon)$ for all $0 < \epsilon < t_i$.
2. Let $\epsilon_0 = 0$, and ϵ_n be defined as earlier. For $\epsilon_1, \dots, \epsilon_{n-1}$, let

$$\epsilon_i = \frac{r_i + r_{i-1}}{2}$$ [6]

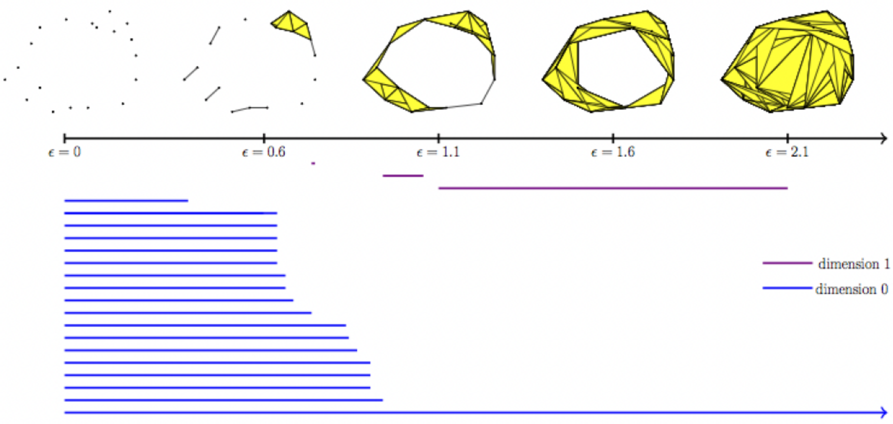


Figure 9: An example Vietoris-Rips filtration of a dataset with corresponding barcode (see 1.5.3). [17]

1.5 Persistence Diagram, Barcode

The purpose of persistent homology being data analysis, we would like to graphically see which holes "last longer" and hence have more important implications for the data. Besides plotting out a given dataset and drawing its Čech or Vietoris-Rips complexes for all the chosen ϵ_i found in 1.4.7, one can use other graphs that plot the "birth" and "death" of each hole in the homology of the dataset. Two of the most popular graphs of this type are the persistence diagram and the barcode.

First, we need to quantify the birth and death times of a hole. Intuitively, they are the ϵ_i-values of a hole's appearance and disappearance.

1.5.1 Definition (Birth and Death of Hole)

Let A be a given dataset, and let $\epsilon_0, \epsilon_1, \dots, \epsilon_n$ be chosen according to 1.4.7.

- A hole h of dimension k *is born going into* $C(A, \epsilon_i)$ if h is nontrivial in $H_k(C(A, \epsilon_i))$ but the boundary of the hole does not exist in $C(A, \epsilon_r)$ for any $r < i$. ϵ_i is the *birth time* of h.
- A hole h of dimension k *dies going into* $C(A, \epsilon_i)$ if h is nontrivial in $H_k(C(A, \epsilon_{i-1}))$ but is trivial in $H_k(C(A, \epsilon_s))$ for all $s \geq i$. ϵ_i is the *death time* of h. [6]

1.5.2 Definition (Persistence Diagram)

The **persistence diagram** (PD) of dimension k for a dataset A is a planar diagram (over $\mathbb{R}^2$ where each k-dimensional hole that has ever been born is represented by the point (b, d), where b is the birth time and d is the death time of the hole. [1]

1.5.3 Definition (Barcode)

The persistence barcode (or just **barcode**) of dimension k for a dataset A is a bar graph with an axis labeled by ϵ_i-values. Each k-dimensional hole that has ever been born is represented by a bar starting at b and ending at d, where b is the hole's birth time and d is its death time. The bars may be unordered. [1]

Persistence diagrams and barcodes across several dimensions may be represented more conveniently by overlaying the diagrams together and using a legend to distinguish between points or bars representing holes of different dimensions.

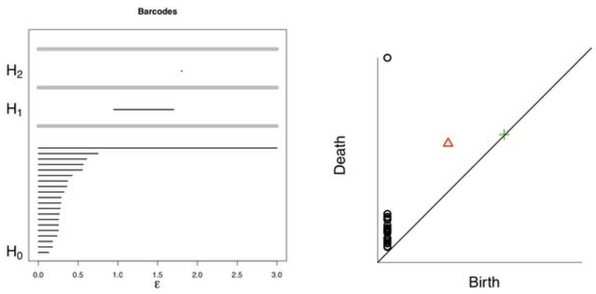


Figure 10: The barcode and persistence diagram of dimensions 0, 1, and 2 for the same dataset. [14]

2 Stability of Persistent Structures

We have seen how the topological features of a dataset can be visualized via persistence diagrams and barcodes. We now determine to what extent these visual representations are stable against perturbations of the data. If we want persistent homology to be a viable analysis tool, this will be important to know, since in practice most values in a dataset are measured more than once, and rarely will one measurement be completely identical to another.

First, some clarification: for two datasets corresponding to the same measured quantity, the former may differ from the latter in either the *number* of recorded data points or the *value* of a given data point. The criteria by which stability is measured must take both into account; as we will see, this can be done by defining the "distance" between datasets in a special way, and then relating it to another "distance" between persistence diagrams. We will now go into detail concerning the first of these.

2.1 Gromov-Hausdorff Distance

Previously, we defined via 1.1.2 the notion of a metric d between two *elements* of a set. We now want to quantify the distance between two *subsets* of a larger (but finite) set upon which a metric is specified. Recall that, in Figure 5, the ε-fattened version of a set contained points that were not originally in A proper. This suggests that it is possible for an ε-fattened set to "absorb" neighbouring sets, which leads us to the definition of the Hausdorff distance.

2.1.1 Definition (Hausdorff Distance)

The **Hausdorff distance** between two subsets $A, B \subset X$ is defined as

$$d_H(A,B) := \inf\{\varepsilon \geq 0 | A \subseteq B_\varepsilon \wedge B \subseteq A_\varepsilon\}$$

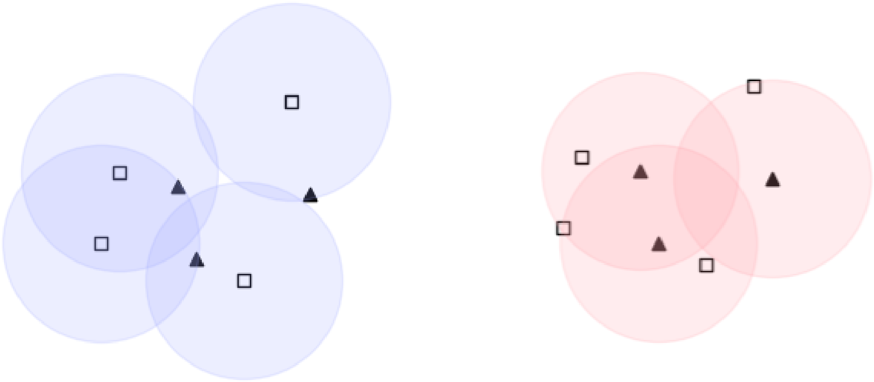


Figure 11: Two ε-fattened sets represented by triangular and square markers. The Hausdorff distance is illustrated here as the minimal radius (ε) of the open ball required for one set to contain the other.

Intuitively, the Hausdorff distance between two subsets quantifies the smallest ε such that one subset, when fattened by ε, contains the other. For the purpose of comparing topological features, however, using the Hausdorff distance is not always sufficient, since it may be that two datasets which are far apart in Hausdorff distance are nonetheless topologically similar. This is often the case when one set can be transformed into another via translation and rotation. To deal with this, we must "align" the sets we are dealing with in the ambient space; that way, our focus is not on the explicit location of a set, but on its configuration and corresponding properties.

2.1.2 Definition (Isometric Embedding)

Let $A \subset X$ be a bounded set. A function $f : A \to X$ is called an **isometric embedding** of A if, $\forall a_i, a_j \in A$, we have $d(a_i, a_j) = d(f(a_i), f(a_j))$.

2.1.3 Definition (Gromov-Hausdorff Distance)

Let $A, B \subset X$ be bounded. Let α, β be the sets of all isometric embeddings of A and B into X, respectively. If $f \in \alpha$ and $g \in \beta$, the **Gromov-Hausdorff distance** between A and B is given by

$$d_{GH}(A,B) := \inf_{\substack{f \in \alpha, \\ g \in \beta}} \{d_H(f(A), g(B))\}$$

Rather than measuring the Hausdorff distance between A and B, d_{GH} is defined to be the minimal[3] Hausdorff distance over *all* isometric embeddings of A and B into the ambient space. This implies that, if $d_{GH}(A,B) = 0$, A and B can identically mapped in the space ($f(A) = g(B)$).

---

[3] Because it is in some sense the "shortest path" between two isometry classes, the Gromov-Hausdorff distance can be thought of as geodesic, although its existence as such is beyond the scope of this paper. See [5] for further reading.

2.2 Bottleneck Distance

Having quantified the distance between two sets of points A and B in a metric space, we also need a way of quantifying the distance between the corresponding persistence diagrams PD(A) and PD(B). This can be done in the following way.

2.2.1 Definition (Chebyshev Distance)

Let $a = (x_a, y_a), b = (x_b, y_b) \in \mathbb{R}^2$. Then the **Chebyshev distance** L^∞ between a and b is given by

$$||a - b||_\infty := \max\{|x_a - x_b|, |y_a - y_b|\}$$

L^∞ is defined along the coordinate axes in $\mathbb{R}^2$. In the context of the n^{th} persistence diagram, these correspond to the values of ϵ at which an n-simplex is born and subsequently dies.

Now, let U and V be finite sets, with $|U|$ not necessarily equal to $|V|$. As with the Gromov-Hausdorff metric, we would like to use the geometry of U and V to allow the Chebyshev distance $||u - v||_\infty$ between individual elements in each set to inform the distance between the sets themselves. If $U = V$, we also require the distance between them to be 0, which suggests that we want to match each $u \in U$ to some $v \in V$ such that their Chebyshev distance is *minimal*.

The naive approach would be to find a mapping ϕ such that the Chebyshev distance is minimized between $u \in U$ and its image under ϕ. If we let Φ be the set of all functions $\phi : U \to V$, then we could define a distance

$$d(u, v) \overset{?}{:=} \min_{\phi \in \Phi}\{||u - \phi(u)||\}_\infty$$

To find the distance between U and V, we would then take the largest such distance, i.e.

$$d(U, V) \overset{?}{:=} \min_{\phi} \max_{u \in U}\{||u - \phi(u)||\}_\infty$$

However, this approach does not take the size of either set into account, as there may not be a bijective mapping of U onto V. The solution is given by the geometry of the persistence diagram itself: since in practice many points in a persistence diagram live on the diagonal D due to random sampling noise, adding a sufficient number of points on the diagonal to both U and V such that the resultant sets have the same number of points would allow a bijection from one onto the other.

2.2.2 Definition (Bottleneck Distance)

Let U, V be finite sets in $\mathbb{R}^2$, and let $D = \{(x, y) \in \mathbb{R}^2 | x = y \geq 0\}$. Let D_U be the orthogonal projection of U onto D and define D_V analogously. Let Φ be the set of all bijections $\phi : U \cup D_V \to V \cup D_U$. Then the **bottleneck distance** between U and V is given by

$$d_B(U, V) := \inf_{\phi \in \Phi} \sup_{u \in U} ||u - \phi(u)||_\infty{}^{4}$$

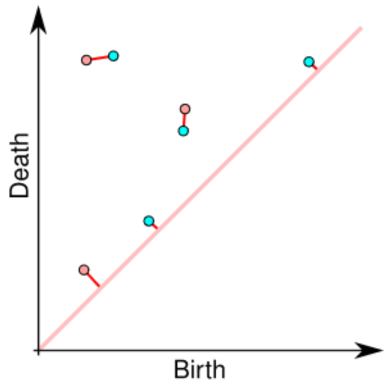


Figure 12: Two overlaid persistence diagrams–the set of orange points and the set of green points–with the optimal matching between points represented as a red line. Note that the red lines do not represent distances between points, as the bottleneck distance is defined to be parallel to one of the coordinate axes. [6].

In essence, the bottleneck distance quantifies the maximum amount by which we would have to adjust either the birth or death time of some $u \in U$ for it to identically match its optimal pairing $v \in V$. If $|V| = 0$, the bottleneck distance is measured from U to its orthogonal projection on D.

We now state without proof the stability theorem for finite sets in $\mathbb{R}^n$, which relates the Gromov-Hausdorff distance between sets to their corresponding bottleneck distance.

---

[4] Note that, since U and V are finite sets in $\mathbb{R}^2$, the infimum over Φ and supremum over U will always be attained as an explicit minimum and maximum, respectively. We include this notation for generality.

2.3 Theorem (Stability theorem)

Let $X, Y \in \mathbb{R}^n$ be finite sets. Let U be the persistence diagram of the Vietoris–Rips filtration of X, and let V be the persistence diagram of the Vietoris–Rips filtration of Y. Then

$$d_B(U, V) \leq 2d_{GH}(X, Y)$$

For a detailed proof of this result, see [4]. The stability theorem suggests that the difference between representations of the topological features of two distinct sets is bounded from above by how far apart the sets are in the Gromov-Hausdorff distance. Consequently, given two sets X and Y where the behaviour of one set in the ambient space is known (e.g. spatial coordinates, phase space, etc.), the behaviour of the other set may be reasonably inferred from that of the first if the Gromov-Hausdorff distance between them is small, as this sets an upper bound on how much the topological features of the two sets can differ.

This description may be somewhat abstract, so the following section provides some concrete examples of persistent homology, and discusses some of its myriad applications.

3 Computation and Applications

3.1 Computation (Distance Matrix)

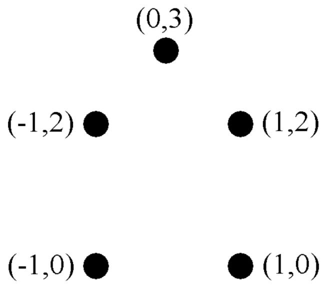
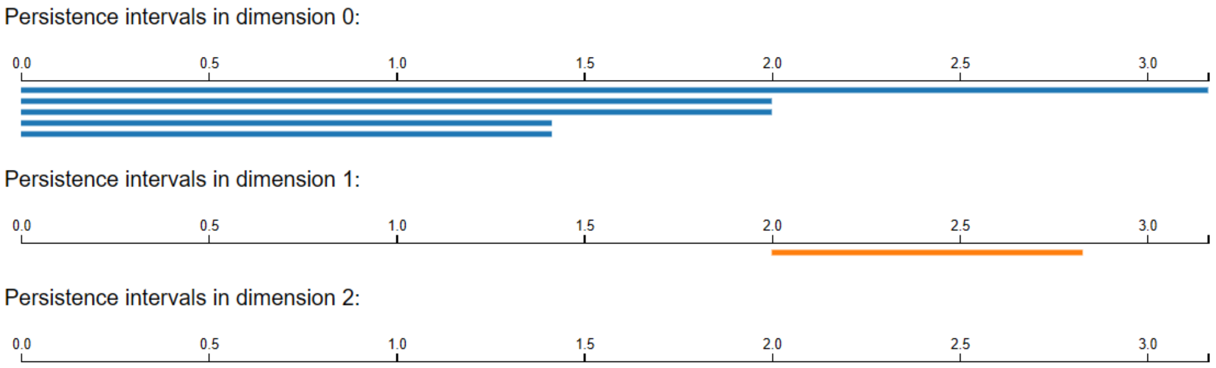


Figure 13: 5 points on a plane containing a hole

Using Ripser[5], we compute the persistent barcode of the Vietoris-Rips complex of the following 5 points in a plane: (0.3), (1,2), (-1,2), (-1,0), (1,0). The input, a 5x5 distance matrix corresponding to the distances between the 5 points, is fed to the program.



According to the persistent barcodes, there is a single relatively persisting 1-dimensional feature (a hole) born at 2, and dying at $2\sqrt{2}$.

3.2 Computation (Molecular Configuration, Cyclo-octane)

This time, we compute the persistent barcode of a real cyclooctane (C_8H_{16}) molecule conformations, again using Ripser.

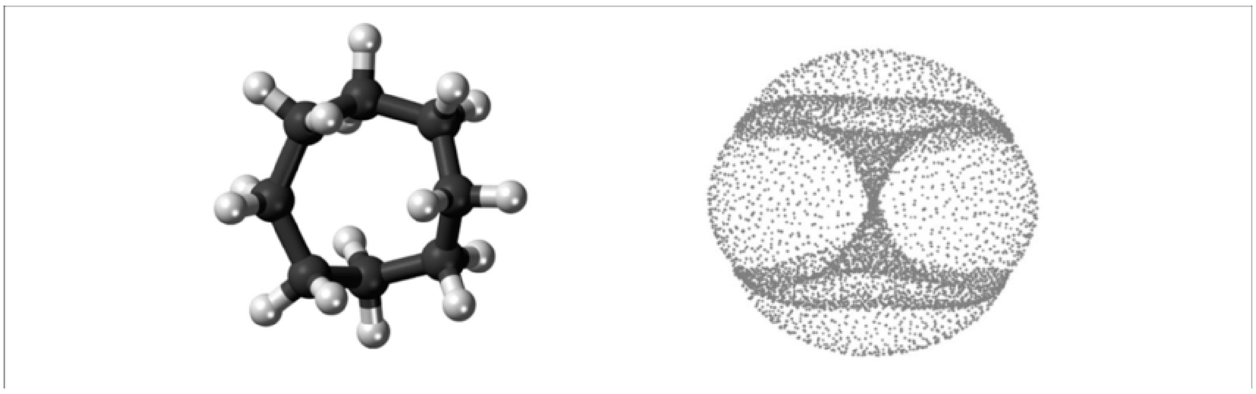


Figure 14: (Left) cyclo-octane molecule consisting of a ring of 8 carbon atoms and 16 hydrogen atoms. (Right) A PCA projection of a dataset of different conformation of the cyclo-octane molecule. [1]

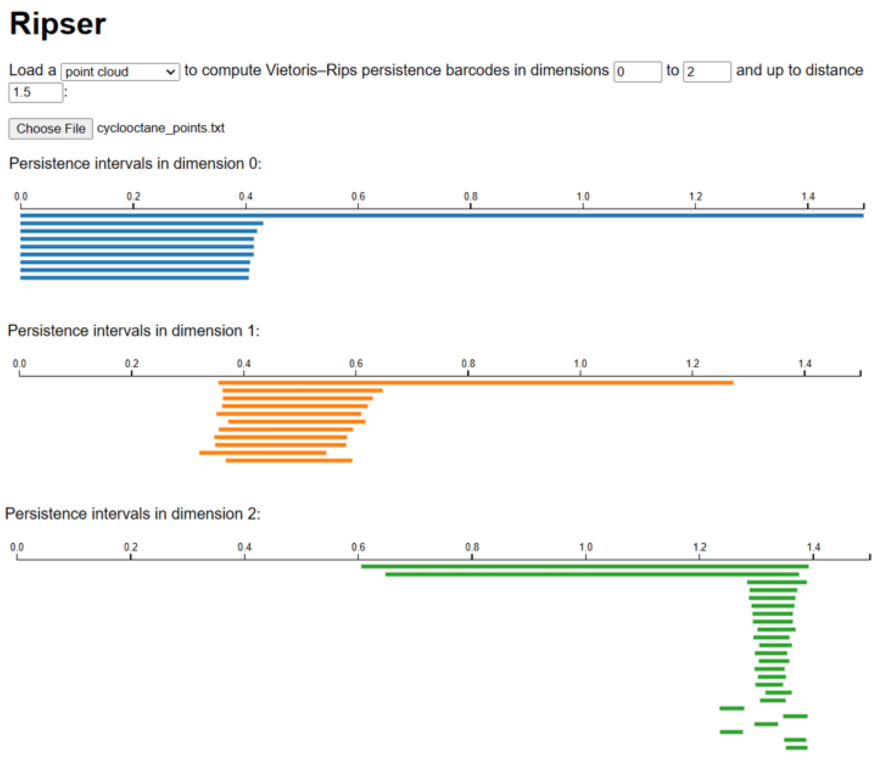
Each molecule conformation can be mapped to a point in $\mathbb{R}^{24=8*3}$ as there are a total of 8 carbons in each molecule and each carbon is represented by 3 coordinates x, y, and z. However, this is not easily imaginable nor physically realiseable. Thus, we employ principal component analysis (PCA), a dimensionality-reduction method, to project the conformation space of $\mathrm{C} * 8\mathrm{H} * 16$ onto $\mathbb{R}^3$, as demonstrated by [2].

As seen in Figure 15, the resulting Betti numbers are $Betti_0 = Betti_1 = 1$, and $Betti_2 = 2$, consistent with the homology groups of the shape: a union of a hollow sphere with an hourglass inside.

This is a great example of efficiently recovering homology groups of the conformation space using persistent homology. For further details, refer to [2].

---

[5] a C++ code made for VR complex persistence barcode computation



Ripser

Load a point cloud to compute Vietoris–Rips persistence barcodes in dimensions 0 to 2 and up to distance 1.5:

Choose File cyclooctane_points.txt

Persistence intervals in dimension 0:

Persistence intervals in dimension 1:

Persistence intervals in dimension 2:

Figure 15: Persistent barcodes in dimensions 0, 1, and 2

3.3 Application (Biological Aggregations)

Most real-life datasets unfortunately do not come in perfect shapes such as spheres or torus. Instead, they often contain noises or lack certain components. For biological aggregations, like a flock of birds or a school of fish, the issue worsens due to the high dimensionality of data and the local interactions between individual agents that form the aggregation. Nevertheless, TDA serves as a very useful tool to extract both the local and global topological features.

Figure 16: A flock of birds has a lot of both local and global geometrical features. [11]

Examining the topological features reveals the shapes and structures for bird flocks that emerge from their collective behavior. Computation of persistent homology identifies and measures these features, helping us understand the flock's organization and dynamics over time.

The local features are just as important. Formation patterns usually follow simple local rules, such as maintaining a certain distance from other birds or aligning their direction. Local features help identify these rules and patterns, which are fundamental to how the flock forms and maintains its shape. In addition, they show how birds adjust their positions and speed in response to changes, providing insights into the flock's adaptability.

Betti numbers, introduced in Definition 1.3.5., are useful for counting connected components, trapped volumes, and more. These homological measures distinguish simulations that traditional methods cannot, improving the classification accuracy, as shown in [3].

To summarize, TDA provides insight into mechanisms that govern flock formation and maintenance, through underlying topological features and their persistence across scales. This leads to a better understanding of the individual-level interactions that result in the observed collective—global — behavior.

3.4 Application (Biological Neural Network)

The main idea underlying TDA is that the clusters of data points, or 'point cloud' of data may reveal important properties that may not be obvious from other approaches. In the case of biological neural networks, TDA has been found useful in investigating brain network structures with possible abnormalities and cognitive performance through brain-information flow, compared to graph theory-based measures.

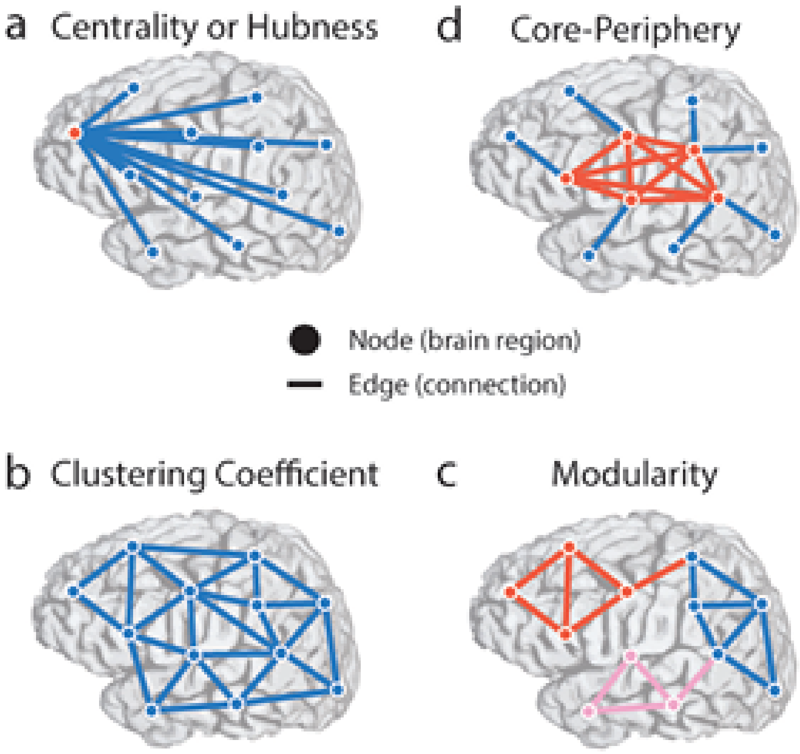


Figure 17: Examples of types of brain network [13]

Once the networks are extracted from scans obtained by imaging data (i.e., fMRI; functional magnetic resonance imaging), we can identify topological features such as cavities and connections. Their possible utility encompasses connectivity estimation, disease diagnosis and monitoring, and rehabilitation and treatment. Schizophrenia, autism, and Alzheimer's disease are examples of diseases where the alterations in the brain network could serve as diagnostic biomarkers. Research on network reorganization induced by memory rehabilitation treatment and seizure therapy supports this potential efficacy. All possible uses of this algebraic topological tool are yet to be fully discovered, however, continued discussion and research will continue to open doors for a more holistic and robust understanding of neural systems.

4 Acknowledgements

We would like to give our deepest thanks to our supervisor, Julien Cheng, without whom this project would not be possible. Special thanks also goes to Dr. Vukasin Stojisavljevic for his knowledge on the topic. While many sources were used in the compilation of this project, much of its foundational material comes from Adams, Bush, and Mirth [1].

References

[1] Henry Adams, Johnathan Bush, and Joshua Mirth. *Data Science For Mathematicians.* Chapman and Hall/CRC, 1st edition, 2020.

[2] Henry Adams and Michael Moy. Topology applied to machine learning: From global to local. *Frontiers in Artificial Intelligence 4:668302*, May 2021.

[3] Tom Halverson Chad M. Topaz, Lori Ziegelmeier. Topological data analysis of biological aggregation models. *PLoS ONE*, May 2015.

[4] Frederic Chazal and David Cohen-Steiner. Proximity of persistence modules and their diagrams. 2008.

[5] Samir Chowdhury and Facundo Mémoli. Explicit geodesics in gromov-hausdorff space. 2018.

[6] Brittany Terese Fasy, Fabrizio Lecci, Alessandro Rinaldo, Larry Wasserman, Sivaraman Balakrishnan, and Aarti Singh. Statistical inference for persistent homology: Confidence sets for persistence diagrams. *Annals of Statistics*, 2014.

[7] Allen Hatcher. *Algebraic Topology.* Cambridge University Press, 1st edition, 2001.

[8] Kindness Chen (https://mathoverflow.net/users/527960/kindness chen). Why is the vietoris–rips complex VR(s, ϵ) a subset of the Čech complex Čech$(s, \epsilon\sqrt{2})$? MathOverflow. URL:https://mathoverflow.net/q/474145 (version: 2024-06-29).

[9] Jiří Matoušek. *Using the Borsuk–Ulam Theorem: Lectures on Topological Methods in Combinatorics and Geometry.* Springer-Verlag, 2003.

[10] Eric W. Weisstein. Jung's theorem, 2024. Last accessed 30 June 2024.

Images

[11] B.A.C. A flock of starlings in denmark. flocks of this size and larger are common in north america following the successful introduction of the species in the 19th century., 2006. Last accessed 19 June 2024.

[12] Omer Bobrowski. A Čech complex generated by a set of points in r2, 2022. Last accessed 30 June 2024.

[13] Ankit N. Khambhati1 Danielle S. Bassett and Scott T. Grafton. Emerging frontiers of neuroengineering: A network science of brain connectivity. *Annual Review of Biomedical Engineering*, 19, 03 2017.

[14] Brittany Terese Fasy, Fabrizio Lecci, Alessandro Rinaldo, Larry Wasserman, Sivaraman Balakrishnan, and Aarti Singh. Statistical inference for persistent homology: Confidence sets for persistence diagrams, 2014.

[15] Martina Flammer. Example of a Čech complex, taken from [10], 2023. Last accessed 19 June 2024.

[16] Peter Grindrod. 2: Computation of ph for a point cloud using the vietoris rips complex., 2015. Last accessed 19 June 2024.

[17] Nina Otter, Mason Porter, Ulrike Tillmann, Peter Grindrod, and Heather Harrington. A roadmap for the computation of persistent homology. *EPJ Data Science*, 6, 06 2015.

[18] platypeanArchcow. A diagram illustrating the great-circle distance (in cyan) and the straight-line distance (in red) between two points p and q on a sphere., 2022. Last accessed 19 June 2024.

[19] Peter Prevos. Transforming a coffee mug into a doughnut., 2018. Last accessed 19 June 2024.

[20] Trevorgoodchild. Simplicial complex example., 2007. Last accessed 19 June 2024.