Brains and Bugs: Two Applications of Persistent Homology in the Life Sciences

Daniel Collin

April 27, 2021

Abstract

Persistent homology is a way of giving a topological summary of a data-set. We give an introduction to the construction of persistent homology, including a proof of its algebraic decomposition as a finitely generated graded module with respect to a polynomial ring over a field. This proof is constructive and yields an algorithm for computing persistent homology in a practical sense.

In order to provide examples of the practical applications of persistent homology, we present two case studies. In the first case study we investigate the relationship between size and persistent homology of the bumblebee *Bombus terrestris*. We find that the difference in size between samples is in some ways highly correlated with differences in persistent homology. In the second case study we analyze the the persistent homology of a synthetic model of the striatum, a part of the basal ganglia in the brain. Here we find that the syntetic model differs in size and complexity from a number of control models when viewed through the lens of persistent homology and the accompanying theory.

Acknowledgements

First of all I would like to thank my supervisor Yishao Zhou for supporting my ideas, helping me find data for the case studies and helping me acquire computational resources to complete this project. I would also like to thank the Insect Sensory Ecology and Cognition Lab at the Department of Zoology (Stockholm University) and the ? Group at the Department of (?) (KTH) for providing me with the data for the two case studies. In particular, I would like to thank Johannes Hjorth and Emily Baird for giving me insight into their respective domains. Finally I would like to thank my girlfriend Mira, who put up with me through all of this.

The computations of in Section 4.2 were enabled by resources provided by the Swedish National Infrastructure for Computing (SNIC) at the High Performance Computing Center North (HPC2N) partially funded by the Swedish Research Council through grant agreement no. 2021-22166.

Contents

1 Introduction

Although ordinary statistical analysis and machine learning continue to see great success, the ever-changing modern digital landscape of data suggests that there is some value in exploring other avenues in mathematics for understanding data. One such avenue is Topological Data Analysis (TDA), an umbrella term for data analysis achieved through topological methods. While topology and algebraic topology in particular might be seen as something relegated to realms of mathematics, the perhaps most popular technique of TDA, persistent homology, has been successfully applied in areas such as neuroscience [1], biology [2] and material science [3].

In persistent homology where we approximate a non-trivial topological space, often a simplicial complex, on the data of interest. From this complex "holes" in the resulting space can be found, and these holes are what constitute homology. Now this approximation is not perfect and there are multiple complexes we can define on a space, so instead we define a sequence of complexes ordered by inclusion and then we compute for how many complexes in the sequence a hole persists.

While the high-level idea is not very complicated, the devil is in the details. In order to rigorously define this notion of persistent homology, as well as keep it flexible for other complexes than simplicial ones, we need to build a robust framework. We do this by defining a sort of complex of complexes from which we retrieve the holes that persist when going from one complex to another.

Our goal with this thesis is partly to provide an introduction to persistent homology as we would have liked it before we started this journey which is done in Chapters 2 and 3. As such, we have tried to keep a balance between the older material in the field that is foundational and newer material that is more up-to-date. Most of the definitions and results are accompanied by commentary, hopefully providing help along the way. We try to make the theoretical part somewhat self-contained, but some familiarity with linear algebra, category theory, commutative algebra and module theory is needed.

The other part of the thesis is given by Chapter 4 and consists of two case studies. In the first study we analyze the eyes of the bumblebee *Bombus terrestris*, in the second study we analyze a synthetic microcircuit of the striatum in the basal ganglia of the brain. Our goal is that these two case studies, although small, show that persistent homology has potential as a tool in the toolbox of data analysis. We have taken care to conduct our analysis in such a way that we highlight how persistent homology enables our approaches.

We owe a lot to a variety of sources that are cited throughout the thesis. The algebraic framework that we present in Chapter 3 was first developed in [4], although it borrows heavily from the more computational view presented in [5]. The articles [6], [7] have been of extra importance, as they provide clear overviews of the theory generalized to modules. Our novel contributions are our methodologies in the two case studies, although we it would not surprise us if similar approaches have been tried before in other domains.

2 Homology

Before go into what *persistent* homology is, it is well worth our time to clearly state what we mean by homology. Broadly speaking, homology is an invariant of topological spaces which is concerned with cycles in the space which are not boundaries. Or more abstractly, homology captures the notion n-dimensional holes in the space.

In *persistent* homology we generally work without predefined topological spaces and start with a basic data-set which at most contains some metric structure which we endow with some form of complex. The main complex we will be working with is the simplicial complex, since computationally we can approximate a such a complex on a set of data-points. We will also review homology of cubical complexes as this relates to one of the case studies. Therefore the classical definitions involving concepts such as singular homology are not something we will dwell on, but rather we refer the reader to Hatcher's excellent exposition in [8].

2.1 Simplices

We start with what will constitute our atoms in simplicial homology, namely the simplices.

Definition 2.1.1 ([9, p. 62])**.** An **n-simplex** is the smallest possible convex set in $\mathbb{R}^m$ containing $n+1$ points $v_0, \ldots, v_n$ such that the vectors $v_1 - v_0, \ldots, v_n - v_0$ are linearly independent. The points $v_0, \ldots, v_n$ are known as the **vertices** of the simplex. The number n is the **dimension** of a simplex.

As seen in Figure 2.1 the $0, 1, 2$ and 3-dimensional simplices are familiar shapes consisting of vertices, edges, triangles and tetrahedrons.

Definition 2.1.2 ([9, p. 62])**.** A **face** of a simplex is the convex hull of a subset of its vertices. If τ is a face of σ we write $\tau \subseteq \sigma$.

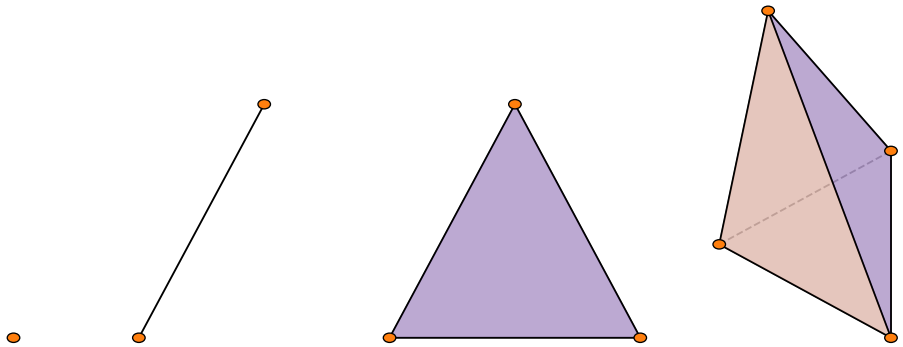


Figure 2.1: *0-simplex (left), 1-simplex (middle left), 2-simplex (middle right) and 3-simplex (right).*

Since higher dimensional simplices are made up of simplices of lower dimensions we can always decompose a simplex into its faces, in other words the lower dimensional simplices that make up the simplex. For example, a 2-simplex can be decomposed into the three edges that make up the triangle.

2.2 Simplicial complex

By gluing together simplices at their faces as seen in Figure 2.2 we can construct higher-order objects which we call simplicial complexes.

Definition 2.2.1 ([9, p. 63])**.** A **simplicial complex** K is a finite collection of simplices such that

- $\sigma \in K$ and $\tau \subset \sigma$ implies that $\tau \in K$
- $\sigma_1, \sigma_2 \in K$ implies that $\sigma_1 \cap \sigma_2$ is either empty or a face of both.

The first requirement tells us that a simplicial complex contains the faces of all its simplices. The second requirement tells us that the simplices are only glued together at common faces, in other words we do not allow simplices to intersect other than at their boundary.

We will refer to the construction in Definition 2.2.1 as a **geometric simplicial complex**. This is to distinguish it from a simplicial complex where we discard the geometric connotations. It is possible to define a combinatorial simplicial

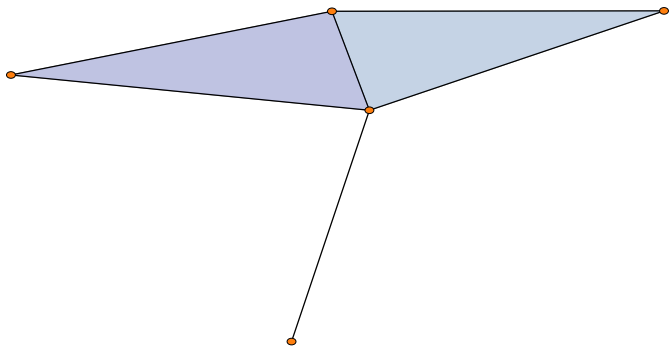


Figure 2.2: *Example of a simplicial complex consisting of two 2-simplices glued together with an attached 1-simplex.*

complex by only considering the ordering of the vertices and what higher-dimensional simplices they make up.

Definition 2.2.2 ([9, p. 62])**.** An **abstract simplicial complex** K is a finite collection of ordered sets such that $\alpha \in A$ and $\beta \subseteq \alpha$ implies that $\beta \in A$.

This abstract definition coincides with the geometric definition by calling the elements of K its simplices. The simplices of K are no longer geometric objects in metric space, but simply combinatorial objects consisting of vertex sets.

Definition 2.2.3. Given a geometric simplicial complex K, we can construct an abstract simplicial complex A by translating each simplex $\sigma_i \in K$ to the ordered set $[v_0, \ldots, v_n]$ where each v_i is an algebraic object simply denoting the presence of a vertex. We call K the **geometric realization** of A.

Hence, we can always transition from a geometric simplicial complex to an abstract simplicial complex. There is an elementary result which allows for the construction in other direction.

Theorem 1 ([9, Geometric Realization Theorem, p. 64])**.** *Every abstract simplicial complex of dimension n has a geometric realization in $\mathbb{R}^{2d+1}$.*

Of course, for a given abstract simplicial complex there might be many different ways of giving a geometric realization for it, but at least Theorem 1 guarantees the existence of one such realization.

From here on we will simply refer to an abstract simplicial complex as a simplicial complex unless stated otherwise. This allows us to state our definitions and work with simplices solely as combinatorial objects.

2.3 Simplicial homology

Roughly speaking, in topology when two spaces homeomorphic they share the same topological properties. One invariant that is the algebraic structure of the higher dimensional holes in the space. Homology can generally thought of as being the characterization of these holes in a topological space. Beyond that however, it is a way of associating algebraic objects to topological spaces. In the context of simplicial complexes, we first need some algebraic machinery in order to define precisely what we mean by holes in a simplicial complex.

Definition 2.3.1 ([4])**.** The kth **chain module** $C_k(K)$ on a simplicial complex K is the free module with basis given by the k-dimensional simplices in K with coefficients in some ring R with additive unit 0 and multiplicative unit 1. In other words, the elements of $C_k(K)$ are formal sums

$$\sum_{i=1} r_i \sigma_i$$

where $r_i \in R$ and σ_i is a k-dimensional simplex in K.

Definition 2.3.2 ([10, p. 2])**.** A **chain complex** C_* over a ring R is a family of R-modules $\{C_k\}_{k\in\mathbb{Z}}$ with R-module maps $\partial_k : C_k \to C_{k-1}$ such that the composition $\partial_{k-1} \circ \partial_k$ is the zero map. We call the maps ∂_k the **differentials** of the chain complex.

Theorem 2. *Given a simplicial complex K and a ring R the sequence of chain modules*

$$\ldots \xrightarrow{\partial_{k+1}} C_k(K) \xrightarrow{\partial_k} C_{k-1}(K) \xrightarrow{\partial_{k-1}} C_{k-2}(K) \xrightarrow{\partial_{k-2}} \ldots \xrightarrow{\partial_1} C_0(K)$$

with differentials defined as

$$\partial_k : C_k(K) \to C_{k-1}(K)$$
$$\partial_k(\sigma) = \sum_{i=0}^{k} (-1)^i [v_0, \ldots, \hat{v}_i, \ldots, v_k]$$

is a chain complex.

Proof. All we need to show is that $\partial\partial\sigma = 0$ for any simplex σ and it extends to arbitrary chains since ∂ is a homomorphism. By linearity of the homomorphism

we get

$$\begin{aligned}\partial\partial\sigma =& \sum_i (-1)^i \partial[v_0, \dots, \hat{v}_i, \dots v_n] \\ =& \sum_{j<i} (-1)^i (-1)^j [\dots, \hat{v}_j, \dots, \hat{v}_i, \dots] \\ &+ \sum_{i<j} (-1)^i (-1)^{j-1} [\dots, \hat{v}_i, \dots, \hat{v}_j, \dots] \\ =& 0.\end{aligned}$$

The first sum comes from when $j < i$ since if remove v_i the position of v_j is unchanged in the resulting simplex, whereas the second sum comes from the other possible case where $i < j$ and so removing v_i causes the position of v_j to shift by one. Hence, the sums cancel out and so $\partial\partial = 0$. □

Example 2.3.1. Given a simplicial complex K consisting of a triangle without interior as in Figure 2.3, a chain in $C_1(K)$ would be a linear combination of edges. For example, an element of $C_1(K)$ is $[v_0, v_1]+[v_1, v_2]$ which is highlighted in green.

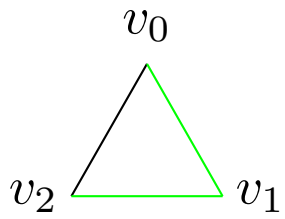


Figure 2.3: *A simplicial complex in which the 1-chain* $[v_0, v_1] + [v_1, v_2]$ *is highlighted in green.*

Example 2.3.2. Given a 2-simplex as in Figure 2.4 we get the differential of the interior of the simplex as

$$\partial_2([v_0, v_1, v_2]) = [v_1, v_2] - [v_0, v_2] + [v_1, v_2]$$

which geometrically is the boundary of the simplex. For this reason we refer to the differential of a chain complex of a simplicial complex as the **boundary map** or **boundary operator**.

The notion of the differential being a map from a simplex to its boundary motivates the following definition.

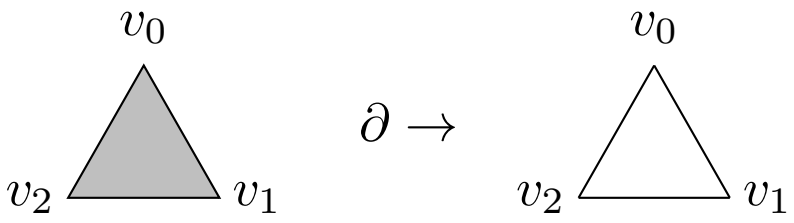


Figure 2.4: *Illustration of how the differential maps a 2-simplex to its boundary.*

Definition 2.3.3 ([10, p. 4])**.** Given a chain complex C_* the k**-cycles** Z_k and the k**-boundaries** B_k of K are the R-modules

$$Z_k := \ker \partial_k$$

$$B_k := \text{im } \partial_{k+1}.$$

Just as for a chain complex we will sometimes refer to the collection of Z_k and B_k as Z_* or B_* respectively.

Our overarching goal is to characterize k-cycles which are not k-boundaries. Hence, a vital result is this corollary to Theorem 2.

Corollary 2.1. *The k-boundaries are a submodule of the k-cycles.*

Proof. Let $\sigma \in B_k = \text{im } \partial_{k+1}$ then for some $\tau \in C_{k+1}$ we have that $\partial_{k+1}(\tau) = \sigma$. Hence,

$$\partial_k(\sigma) = \partial_k \partial_{k+1}(\tau) = (\partial_k \circ \partial_{k+1})\tau = 0$$

and so $\sigma \in \ker \partial_k = Z_k$. □

This tells us that if we can find the cycles and ignore those which are just boundaries, then we have identified the holes. This motivates the following definition of homology.

Definition 2.3.4. Given a simplicial chain complex C_* the homology module H_k is defined as

$$H_k(K) := \ker(\partial_k)/\text{im}(\partial_{k+1})$$

Hence, the kth homology group captures precisely those cycles which are not in the image of the higher dimensional differential. In other words, the non-trivial equivalence classes represent the cycles which are not boundaries.

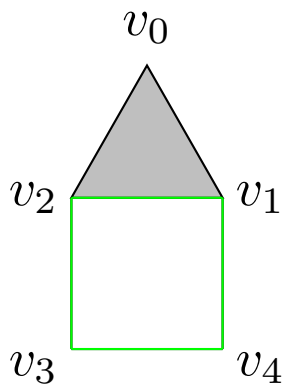


Figure 2.5: *Illustration of a simplicial complex with a 1-cycle in green consisting of a square without interior.*

Example 2.3.3. Consider the chain complex over $\mathbb{Z}_2$ resulting from the simplicial complex in Figure 2.5. There are two 1-cycles, the boundary of the filled triangle and the square. Since the boundary of the filled triangle is in $\text{im}(\partial_2)$ we get that this element belongs to the trivial class in H_1. But there is one non-trivial element given by the square without interior. So H_1 contains one non-trivial element generated by the cycle in green.

To quantify the number of linearly independent generators in a homology module H_k we say that the k**th Betti number** is denoted β_k where $\beta_k = \text{rank } (H_k)$.

2.4 Cubical Homology

The definition we gave of a chain complex is general enough that we need not limit ourselves to chain complexes arising from simplicial complexes. Another type of complex which we can define chain complexes on, and so by extension homology, are cubical complexes. As the atoms of simplicial complexes are simplices, the atoms of cubical complexes are cubes.

Definition 2.4.1 ([11, Definition 2.1, p. 40])**.** An elementary interval is a unit interval $[k, k+1]$ or a degenerate interval $[k, k]$ for $k \in \mathbb{N}$.

Definition 2.4.2 ([11, Definition 2.3-2.4, p. 40])**.** An **elementary cube** Q is the cartesian product of n elementary intervals

$$Q = \prod_{i=0}^{n} I_i \subset \mathbb{R}^n$$

and where n is known as the **embedding number** $emb(Q)$.

Definition 2.4.3 ([11, Definition 2.4, p. 41])**.** The **dimension** $dim(Q)$ of an elementary cube Q is the number of non-degenerate intervals in the product of elementary intervals that make up Q.

Note that under this definition a 0-dimensional cube is a degenerate elementary interval and a 1-dimensional cube is a non-degenerate elementary interval. This corresponds to our notion of vertices and edges in simplicial complexes.

We let $\mathcal{K}^n$ denote the set of all elementary cubes in $\mathbb{R}^n$. The set of all possible elementary cubes is then denote $\mathcal{K}$ defined as

$$\mathcal{K} := \bigcup_{n=1}^{\infty} \mathcal{K}^n$$

Additionally, we define

$$\mathcal{K}_k := \{Q \in \mathcal{K} \mid \dim Q = k\}$$

$$\mathcal{K}_k^n := \{\mathcal{K}_k \cap \mathcal{K}^n\}$$

where it is important to note that $K_d \neq K^d$ since K^d contains any elementary cube embedded in $\mathbb{R}^d$. For instance $Q := [0,0] \times [1,1] \times [2,2] \in \mathcal{K}^3$, but $Q \notin K_3$ since Q only consists of degenerate intervals and so $\dim Q = 0$.

Definition 2.4.4 ([11, Definition 2.9, p. 43])**.** A cubical set X is a finite union of elementary cubes. Given a cubical set X a cubical complex $\mathcal{K}(X)$ is defined as

$$\mathcal{K}(X) := \{Q \in \mathcal{K} \mid Q \subset X\}.$$

Additionally we define the k-skeleton of the cubical complex as

$$\mathcal{K}_k(X) := \{Q \in \mathcal{K}(X) \mid \dim Q = k\}.$$

Much like in the case with the simplicial complexes, working with the actual geometric objects can be unwieldy when doing algebraic calculations. For this reason we introduce an abstract object for each elementary cube, called an elementary chain.

Definition 2.4.5 ([11, p. 48])**.** Given a cube $Q \in \mathcal{K}_k^n$ the elementary k-chain $\hat{Q}$ is a map

$$\hat{Q} : \mathcal{K}_k^n \to \mathbb{Z}$$

Furthermore, since elementary cubes consist of finite cartesian products of elementary intervals, we need a corresponding notion on elementary chains.

Definition 2.4.6 ([11, Definition 2.23, p. 51])**.** Given two elementary cubes P, Q The cubical product is defined as

$$\hat{P} \diamond \hat{Q} := \widehat{P \times Q}$$

Fix some ring R with additive unit 0 and multiplicative unit 1 and then we can proceed as for the simplicial case.

Definition 2.4.7 ([11, Definition 2.27, p. 53])**.** The k**th chain module** of a cubical complex $\mathcal{K}(X)$ is the free R-module $C_k(\mathcal{K}(X))$ whose elements are formal sums

$$\sum \alpha_i Q_i$$

known as k**-chains** where $\alpha_i \in R$ and $Q_i \in K_k(X)$.

The cubical product extends to k-chains in the following way.

Definition 2.4.8. The cubical product $\diamond$ of two k-chains c_1, c_2 of a chain module on a cubical complex is

$$c_1 \diamond c_2 = \sum_i \sum_j \alpha_i \beta_j \hat{P}_i \diamond \hat{Q}_j$$

It can be shown with relative ease that the cubical product on k-chains is distributive, associative and is equal to 0 if one of its arguments is 0, see [11, Proposition 2.25, p. 51].

All we need to do in order for the machinery of homology to be applicable is to define a boundary operator on the elements of the chain modules of a cubical complex which has the property that $\partial\partial = 0$.

Definition 2.4.9 ([11, Definition 2.31, p. 54])**.** The boundary map of an elementary cube is defined inductively on the embedding number cube in the following way. Let Q be an elementary cube, we then have that

$$\partial\hat{Q} = \begin{cases} 0 & Q = [k,k] \\ \widehat{[k+1,k+1]} - \widehat{[k,k]} & Q = [k,k+1] \\ \partial\hat{I} \diamond \hat{P} + (-1)^{\dim \hat{I}} \hat{I} \diamond \partial\hat{P} & Q = I \times P, emb(Q) \geq 2 \end{cases}$$

which extends linearly on k-chains.

Theorem 3 ([11, Proposition 2.37, p. 58])**.** *The composed boundary map $\partial\partial$ is the zero map.*

Proof. Let Q be an elementary cube. If $emb(Q) = 0$ then by the definition of the boundary map $\partial(\partial Q) = \partial(0) = 0$. If $emb(Q) = 1$ then

$$\begin{aligned}\partial\partial Q =& \partial([k+1, k+1] - [k, k]) \\ =& \partial([k+1, k+1]) - \partial([k, k]) \\ =& 0 - 0 \\ =& 0\end{aligned}$$

We prove it for higher embedding numbers by induction. Assume it holds for $emb(Q) = n$, we want to prove that it also holds for $emb(Q) = n + 1$.

$$\partial\partial Q = \partial(\partial\hat{I} \diamond \hat{P}) + \partial((-1)^{\dim I}\hat{I} \diamond \partial\hat{P})$$

Now assume that $\dim I = 0$ then I is a degenerate interval and so we get

$$\begin{aligned}\partial\partial Q =& \partial(0 \diamond \hat{P}) + \partial((-1)^0\hat{I} \diamond \partial\hat{P}) \\ =& 0 + \partial(\hat{I} \diamond \partial\hat{P}) \\ =& (\partial\hat{I} \diamond \partial\hat{P} + (-1)^{\dim I}\hat{I} \circ \partial\partial\hat{P} \\ =& 0 + \hat{I} \diamond \partial\partial\hat{P} \\ =& 0 + 0 \\ =& 0\end{aligned}$$

where $\partial\partial\hat{P} = 0$ follows from the induction hypothesis since $\dim \hat{P} = n+1-1 = n$. Now, let us assume the other case, namely that $\dim I = 1$ then we get

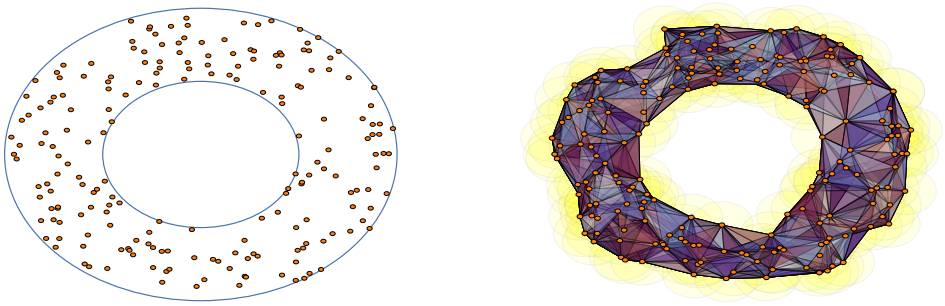
$$\begin{aligned}\partial\partial Q =& \partial(\partial\hat{I} \diamond \hat{P}) - \partial(\hat{I} \diamond \partial\hat{P}) \\ =& \partial\partial\hat{I} \diamond \hat{P} + (-1)^{\dim \partial I}\partial\hat{I} \diamond \partial\hat{P} - (\hat{\partial}I \diamond \partial\hat{P} - \hat{I} \diamond \partial\partial\hat{P}) \\ =& 0 + (-1)^{\dim \partial\hat{I}}\partial\hat{I} \diamond \partial\hat{P} - (\hat{\partial}I \diamond \partial\hat{P} - 0) \\ =& \partial\hat{I} \diamond \partial\hat{P} - \partial\hat{I} \diamond \partial\hat{P} \\ =& 0\end{aligned}$$

where $\dim \partial\hat{I} = 0$ since I is an elementary interval. Since we have now shown the induction hypothesis holds for $n + 1$ given it holds for n, this concludes the proof by induction. □

We now have everything we need to construct a chain complex on cubical complexes: a sequence of R-modules, the k-chain modules on cubical complexes, and module homomorphisms ∂ which have the property of $\partial\partial = 0$. Since our prior definitions with regards to *homology* were entirely stated in terms of chain complexes we can be assured that homology on chain complexes given by cubical complexes is well-defined.

3 Persistence

In the world of data we rarely have a topological description of the space our dataset lives in. At most, we could consider a set of data points as having the discrete topology but that is not very informative. What if there is an underlying topological space with a non-trivial topology? If so, figuring out properties of this space could provide us with indications of how the data is related *globally*. Consider for example the points sampled from an annulus in Figure 3.1a.



(a) **(b)**

Figure 3.1: *Imposing a simplicial complex* ***(b)*** *on data sampled from an annulus* ***(a)****.*

If we were ignorant of the fact that the underlying space has a the shape of an annulus, which is the situation we more often than not would have in a real-world scenario, being able to deduce the topological properties of this space would tell us that data only lies around a hole. This is where persistent homology comes in, a way of approximating the homology of a space without anything other than the data itself.

The basic principle is quite simple. Using the theory of simplicial homology we can impose a simplicial complex on our dataset as in Figure 3.1b. A natural

way of doing this is defining some form of measure of distance on our data-set, such that when samples are sufficiently close to each other we say they belong to the same simplex.

However, there is a problem with the idea in its naive form. How large is "sufficiently close"? If we use too large of a distance we end up with all points in a single simplex and retrieve no valuable homological information. On the other hand, if the distance is too small we end up with a simplicial complex with very few connections between vertices and this too could prove uninformative. As we will see, persistent homology circumvents this problem by simply considering *all* possible distances and encoding the homology of the resulting simplicial complexes in a single mathematical object.

3.1 Endowing a space with a complex

A data-set can often be considered as a set of points in Euclidean space. A natural way of endowing a space of points in $\mathbb{R}^n$ with a simplicial structure is the following construction.

Definition 3.1.1 ([9, p. 72])**.** For a family of points $X = \{x_\alpha\}_\alpha$ in some Euclidean space $\mathbb{R}^n$ the **Čech complex** $\check{\mathrm{C}}_\epsilon$ is given by the abstract simplicial complex whose k-simplices are given by a subfamily of $k+1$ points $\{x_{\alpha_i}\}$ such that

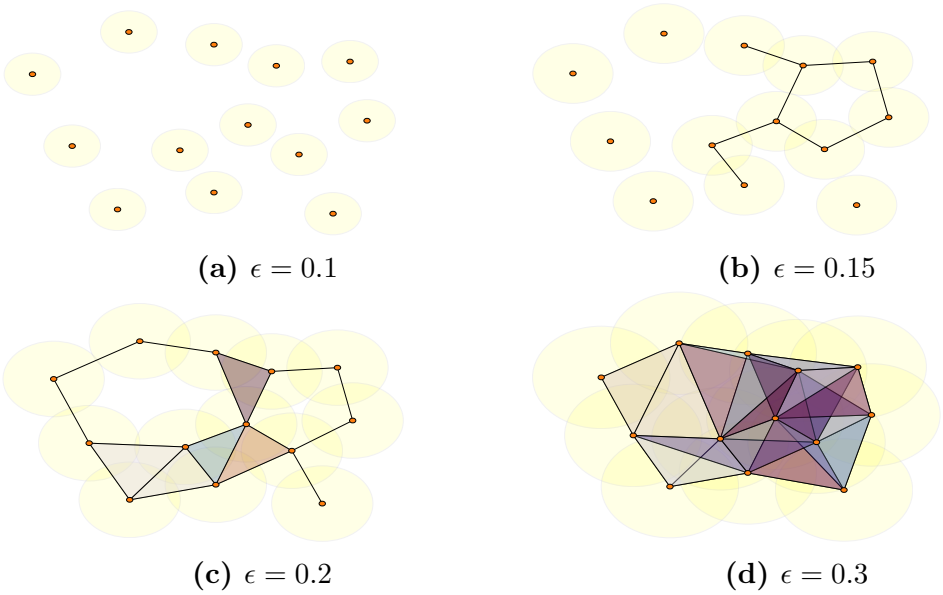
$$\bigcap_{i=0}^{k} B_{\epsilon/2}(x_{\alpha_i}) \neq \emptyset$$

where $B_r(x)$ is the closed ball of radius r centered at x.

The Čech complex is a special case of something called the nerve of a topological space, which guarantees that it has the same homology modules as the union of balls centered at the points [9, p. 71].

However, the Čech complex is for practical purposes not feasible to compute [12]. The reason being that we need to keep the entire simplicial complex in memory and this can be quite large.

A sort of compromise is the Vietoris-Rips complex as seen in Figure 3.2. This complex is a simplification where we do not look for points in common between all balls, but rather say that if $k+1$ have balls that intersect *pairwise* they form a k-simplex.



(a) $\epsilon = 0.1$ **(b)** $\epsilon = 0.15$

(c) $\epsilon = 0.2$ **(d)** $\epsilon = 0.3$

Figure 3.2: *The Vietoris-Rips complex at different ϵ-values.*

Definition 3.1.2 ([9, p. 74])**.** For a family of points $X = \{x_\alpha\}_\alpha$ in some Euclidean space $\mathbb{R}^n$ the **Vietoris-Rips complex** R_ϵ is given by the abstract simplicial complex whose k-simplices are given by a subfamily of $k+1$ points $\{x_{\alpha_i}\}$ such that for any two points in the collection $x_{\alpha_i}, x_{\alpha_j}$ we have that

$$B_{\epsilon/2}(x_{\alpha_i}) \cap B_{\epsilon/2}(x_{\alpha_j}) \neq \emptyset$$

where $B_r(x)$ is the closed ball of radius r centered at x.

The Vietoris-Rips complex does not come with the same guarantee of fidelity to the underlying space as the Čech complex does. However, it is entirely defined by the vertices and the edges of the simplicial complex, allowing it to be stored as a graph.

Given a monotonically increasing sequence of real numbers $(\epsilon_i)_i^n$ we can for each ϵ_i associate to a finite set of points X the Vietoris-Rips complex R_{ϵ_i}. Then as illustrated in Figure 3.2 we have inclusions

$$R_{\epsilon_1} \xhookrightarrow{\iota} R_{\epsilon_2} \xhookrightarrow{\iota} \ldots \xhookrightarrow{\iota} R_{\epsilon_{n-1}} \xhookrightarrow{\iota} R_{\epsilon_n}.$$

For $i < j$ the inclusion $\iota : R_{\epsilon_i} \to R_{\epsilon_j}$ induces a map $\iota_* : H_k(R_{\epsilon_i}) \to H_k(R_{\epsilon_j})$ and the image of this map tell us which equivalence classes in H_k survive when going from R_{ϵ_i} to R_{ϵ_j}, in other words the homological features that persist going from resolution ϵ_i to resolution ϵ_j in the Vietoris-Rips construction. The following

result then lends some credibility to the Vietoris-Rips complex through its relationship with the Čech complex.

Lemma 4 ([9, Vietoris-Rips Lemma, p. 74])**.** *Given $\epsilon > 0$ there is a chain of inclusions*

$$R_\epsilon \hookrightarrow C_{\epsilon\sqrt{2}} \hookrightarrow R_{\epsilon\sqrt{2}}$$

Hence, any cycle that persists through the induced map $H_k(R_\epsilon) \to H_k(R_{\epsilon'})$ for $\epsilon' \geq \sqrt{2}\epsilon$ is also present in the Čech complex $\check{\mathrm{C}}_{\epsilon'}$.

The insight that the homological features that persist tell us more than the individual homology groups themselves are central to the idea of persistent homology. Before we give the formal definition of persistent homology we must first generalize the concept of endowing a space with a complex.

Definition 3.1.3. A **filtration** F of a simplicial (or cubical) complex K is a totally ordered set of subcomplexes $K^i \subseteq K$ for $i \in \mathbb{N}$ such that if $i \leq j$ then $K^i \subseteq K^j$.

Note that the Čech and Vietoris-Rips constructions over a sequence of resolutions are two instances of filtrations, but with this definition we are not restricted to them alone. With this definition in place, we can state the formal definition of persistent homology.

Definition 3.1.4. For $p > 0$ the **p-persistent kth homology module** of filtration F is given as

$$H_k^{i,p} = Z_k^i / (B_k^{i+p} \cap Z_k^i)$$

where Z_k^i, B_k^i are the cycle and boundary modules of the resulting chain complexes $C_*(F_i)$

This module is well-defined since the inclusion $K^i \hookrightarrow K^{i+p}$ induces inclusions $C_k^i \hookrightarrow C_k^{i+p}$ hence we have inclusions $Z_k^i \hookrightarrow C_k^i \hookrightarrow C_k^{i+p}$ and so Z_k^i is a submodule of C_k^{i+p}. Furthermore, it captures precisely what we have been alluding to earlier: the p persistent homology modules are exactly the equivalence classes that survive up to some filtration p.

3.2 Persistence Module

While Definition 3.1.4 serves as a sufficient framework for persistent homology, it is still particular in the sense that it is stated in terms of filtrations and chain complexes arising from them. It is possible to make the notion of persistence

even more general which allows us to understand its algebraic structure. This does not mean that we should entirely discard our anchoring of persistent homology in the realm of simplicial complexes, as it relates closely to how we will do persistent homology in practice, but rather we should let this more abstract approach serve as the theoretical underpinning which opens up the possibility of other types of approximation of data than simplicial complexes.

Definition 3.2.1 ([10, p. 2])**.** Let C_*, D_* be chain complexes over some ring R. A **chain map** $u : C_* \to D_*$ is a family of R-module homomorphisms $u_k : C_k \to D_k$ such that the following diagram commutes

$$\begin{array}{ccccccccc}
\dots & \xrightarrow{\partial_{k+2}} & C_{k+1} & \xrightarrow{\partial_{k+1}} & C_k & \xrightarrow{\partial_k} & C_{k-1} & \xrightarrow{\partial_{k-1}} & \dots \\
 & & \downarrow{\scriptstyle u_{k+1}} & & \downarrow{\scriptstyle u_k} & & \downarrow{\scriptstyle u_{k-1}} & & \\
\dots & \xrightarrow{\partial_{k+2}} & D_{k+1} & \xrightarrow{\partial_{k+1}} & D_k & \xrightarrow{\partial_k} & D_{k-1} & \xrightarrow{\partial_{k-1}} & \dots
\end{array}$$

Definition 3.2.2. A **persistence complex** is a family of chain complexes C^i_* together with chain maps $\iota^i : C^i \to C^{i+1}$ that go between them in the following way

$$\begin{array}{ccccccccc}
 & & \vdots & & \vdots & & & & \\
 & & \downarrow{\scriptstyle \partial_{k+2}} & & \downarrow{\scriptstyle \partial_{k+2}} & & & & \\
\dots & \xrightarrow{\iota^{i-1}} & C^i_{k+1} & \xrightarrow{\iota^i} & C^{i+1}_{k+1} & \xrightarrow{\iota^{i+1}} & \dots & & \\
 & & \downarrow{\scriptstyle \partial_{k+1}} & & \downarrow{\scriptstyle \partial_{k+1}} & & & & \\
\dots & \xrightarrow{\iota^{i-1}} & C^i_k & \xrightarrow{\iota^i} & C^{i+1}_k & \xrightarrow{\iota^{i+1}} & \dots & & \\
 & & \downarrow{\scriptstyle \partial_k} & & \downarrow{\scriptstyle \partial_k} & & & & \\
 & & \vdots & & \vdots & & & &
\end{array}$$

Definition 3.2.3 ([4])**.** A **persistence module** M is a family of R-modules M^k together with module homomorphisms $\phi : M^k \to M^{k+1}$.

With the definition of the persistence module we arrive at an alternate definition of persistent homology, the persistent homology of a persistence complex.

Definition 3.2.4. For $p > 0$ the p-persistent homology of a persistence complex (C_*, ι) is denoted H^p_* and is defined to be the images of the induced homomorphisms $\iota^{p-1}_* \circ \iota^{p-2}_* \circ \cdots \circ \iota^i_* : H_*(C^i_*) \to H_*(C^p_*)$.

In the light of this definition, we see that the p-persistent homology of a persistence complex is a persistence module where the module homomorphisms ϕ are the maps induced by the chain maps $\iota : C_*^i \to C_*^{i+1}$. The objects given in definitions 3.1.4 and 3.2.4 are in fact isomorphic.

Lemma 5. *Let $\iota_k^{i,p} : H_k^i \to H_k^p$ be the module homomorphism that takes a class in H^i to the class which contains that class in H^p. Then $Im(\iota_k^{i,p}) \simeq H_k^p$.*

Proof. Note that the kernel of $\iota^{i,p}$ are exactly those classes of cycles which become boundaries at some index $i, i+1, \ldots, p$, hence $\ker(\iota^{i,p}) = (B^{i+p} \cap Z^i)$. So by the first isomorphism theorem for modules we get that

$$Im(\iota^{i,p}) \simeq H_k^i / \ker(\iota^{i,p}) \simeq H_k^i / (B^{i+p} \cap Z_k^i) \simeq (Z_k^i / B_k^i) / (B_k^{i+p} \cap Z_k^i) \simeq H_k^{i,p}$$

where last isomorphism follows from the fact that $B_k^i \subseteq B_k^{i+p} \cap Z_k^i$. □

Definition 3.2.5 ([4])**.** We say a persistence module (M^k, ϕ^k) is of **finite type** if each component M^k is a finitely generated R-module and the maps ϕ^k are isomorphisms for $k > N$ for some integer N.

When we start with a finite simplicial complex K we get that $C_*(K)$ consists of finitely generated R-modules since the number of simplices in each dimension is finite, hence the resulting persistence complex and persistence modules are of finite type.

The most important theoretic result is just around the corner, but before that we need to recall some definitions regarding graded rings and modules.

Definition 3.2.6. Let R be a ring. We say R is a **graded ring** if it can be decomposed as

$$R = \bigoplus_i R_i$$

Note that given a ring R the polynomial ring $R[x]$ is always a graded ring, since it can be decomposed into $R[x] = Rx^0 \oplus Rx^1 \oplus \ldots$

Definition 3.2.7. A non-zero element r in a graded ring R is said to be homogeneous of degree n if $r \in R_n$ and $r \notin R_j$ for all $j \neq n$.

In other words, the homogeneous elements of a graded ring are those elements that are contained to a single component. Adding elements from different components give us elements that are not homogeneous.

Definition 3.2.8. Let $R = \bigoplus_i R_i$ be a graded ring and M an R-module. We say that M is a **graded R-module** if M decomposes as

$$M = \bigoplus_i M_i$$

where M_i are submodules of M, such that $R_i M_j \subseteq M_{i+j}$.

Most of the ordinary algebraic constructions on modules hold for graded modules as well. The only additional requirement is that they preserve homogeneous elements in the obvious way. For example, a morphism of graded modules is a morphism of modules such that it preserves degree. In other words, a morphism takes an element in a graded module of degree n to an element in another graded module of degree n. Similarly, a graded submodule of a module is simply a graded module such that each component is a submodule of the corresponding component in the parent module.

We can now see that if we have a persistence module M over some ring R and we give R a graded structure by considering $R[t]$ then a graded module structure on M is given by

$$\alpha(M) = \bigoplus_{k=0}^{\infty} M^k$$

The action t^p sends $M^k \to M^{k+p}$ by p repeated applications of t, in other words t shifts the elements up in the graduation by its power

$$t \cdot (m^0, m^1, m^2, \dots) = (0, \phi^0(m^0), \phi^1(m^1), \phi^2(m^2), \dots)$$

and so we get that $R[t]_p M^k = Rt^p M^k \subseteq M^{k+p}$ which satisfies the condition we gave in our definition of a graded module.

The map α is actually a functor between the category of persistence modules and graded modules [12] which becomes an isomorphism of categories when considering persistence modules of finite type over a field. Hence, for ease of notation we will simply consider a persistence module to be a graded module when the aforementioned conditions are fulfilled. This gives us a lot for free: we do not have to be afraid of taking quotients or direct sums of persistence modules as we know what objects they correspond to in the category of graded modules. For example, we write H for the direct sum of the persistence modules H_k and similarly we write C for the persistence complex given by the sum of the persistence modules C_k.

We now arrive at the result which fully characterizes the algebraic structure of persistent homology. This result sadly comes with the restriction that makes α an isomorphism of categories, namely that it only characterizes persistence modules of finite type over a field. The more general problem is under the functor α equivalent to characterizing graded $R[t]$ modules over an arbitrary ring R, which is known to be a hard problem [12].

Theorem 6 ([4])**.** *For a persistence module M of finite type over a field $\mathbb{F}$,*

$$M \cong \bigoplus_i t^{p_i} \cdot \mathbb{F}[t] \oplus (\bigoplus_j (t^{r_j} \cdot \mathbb{F}[t])/(t^{s_j})$$

While the restriction to a field $\mathbb{F}$ somewhat limits the usefulness of the decomposition, we often in practice prefer working in $\mathbb{Z}_2$ due to computational aspects and hence in most cases it poses no real problem.

The proof of Theorem 6 is constructive and ultimately leads to an algorithm for computing persistent homology in terms of linear algebra. Hence, persistent homology is computable in practical applications even for large data-sets. Due to the intimate relation between the proof and the computational aspects we will give it an appropriate treatment in the next section.

Theorem 6 has an intuitive explanation in terms of filtrations: when M is the persistent homology H given by some filtration F, the free part consists of generators which appear in the subcomplex F_{p_i} and continue to exist for all future filtrations. The torsional part consists of generators which appear at in the subcomplex F_{r_j} and disappear in $F_{r_j+s_j}$. Furthermore, the decomposition provides the p-persistent homology for all p and so we circumvent the problem of having to choose an optimal step of the filtration of which to compute homology.

We can make this association of the decomposition of a persistence module with intervals more precise through the following definitions.

Definition 3.2.9. For a persistence module M as in Theorem 6 we associate the interval $(i, j) \in \mathbb{N} \cup \{\infty\}$ to M as

$$Q(i, j) := t^i \cdot \mathbb{F}[t]/t^{j-i},$$

$$Q(i, \infty) := t^i \cdot \mathbb{F}[t].$$

Furthermore, given a multiset of intervals $S = \{(i_0, j_0), \ldots, (i_n, j_n)\}$ we say that

$$Q(S) := \bigoplus_k Q(i_k, j_k).$$

The **barcode $\mathbf{B_M}$** is the multiset of intervals $Q^{-1}(M)$.

In the obvious way, Q is a bijection which maps each summand in the decomposition of a persistence module to an interval. This gives us a correspondence between infinite intervals and generators of the free part and finite intervals and generators of the torsional part.

3.3 Proof of the Structure Theorem

As promised we will now derive the connection between linear algebra and persistence modules. For this part we will abuse the isomorphism of categories between persistence modules and graded modules over $\mathbb{F}[t]$ and readily switch between the two perspectives. A key observation is that for every (graded) module M there is an exact sequence

$$K \to G \to M \to 0$$

where G is the free module on the generators of M and K is the free module on the generators of $\ker G \to M$. By the first isomorphism theorem this exact sequence yields an isomorphism $M \cong G/\text{im}\,(K \to G)$, and hence we can characterize M by only knowing K, G and f. We call such a sequence a **presentation** of M.

Lemma 7 ([7, Theorem 7])**.** *For a finitely generated graded module M over $\mathbb{F}[t]$ there a presentation of M given by a short exact sequence*

$$0 \to K \to G \to M \to 0$$

where G, K are finitely generated.

Proof. First of all, note that since G is the free module on the generators of M and M has a finite set of generators so does G.

A standard result due to Hillbert (see for example [13, Theorem 4, p. 76]) is that any submodule of a finitely generated module over $\mathbb{F}[t]$ is finitely generated. This means in particular that $\ker(G \to M)$ is finitely generated, as it is a submodule of G over $\mathbb{F}[t]$.

We start by proving that $\ker(G \to M)$ is free. Suppose for contradiction that $\ker(G \to M)$ is not free, then for a finite set of generators $\{k_i\}_{i=0}^n$ of $\ker(G \to M)$ the assumption implies linear dependence such that

$$\sum_i r_i t^{a_i} k_i = 0 \iff t^b \sum_i r_i t^{a_i - b} k_i = 0 \implies \sum_i r_i t^{a_i - b} k_i = 0$$

where $b = \min\{a_i \mid r_i \neq 0\}, r_i \in \mathbb{F}$ and the last implication follows from that $\ker(G \to M)$ is a submodule of the free module G. Then for some term in the sum we have that $a_j = b$ and $k_j \neq 0$ which gives us that

$$r_j t^{a_j - b} k_j = r_j k_j \implies k_j = -r_j^{-1} \sum_{i \neq j} r_i t^{a_i - b} k_i.$$

But then k_j can be given as a linear sum of generators of $\ker(G \to M)$ so it can be excluded from the list of generators. Iterating the argument above then exhausts the finite set of generators until the set consists of linear independent generators, hence $\ker(G \to M)$ is free.

Since $\ker(G \to M) = \text{im } (K \to G)$ by the exactness of the presentation and K is the free module on the generators of this finitely generated, free image it follows that $K \cong \text{im } K \to G$ and so $K \to G$ is injective and K is finitely generated. □

When K, G are free and finitely generated the presentation yields a map $\mathbb{F}[t]^n \to \mathbb{F}[t]^m$, in other words it can for a given choice of bases be represented by a matrix in $\mathbb{F}[t]^{n \times m}$ which we call a **presentation matrix** of M. This is the vital connection between persistence modules and linear algebra. By reducing this matrix, much like in the process of Gaussian elimination, we end up with a matrix that describes compatible bases in K and G. Then M is generated by the basis in G with relations given by the basis in K.

Definition 3.3.1. A matrix with coefficients in $\mathbb{F}[t]$ is in **graded Smith normal** if its only non-zero entries are on the (possibly permuted) diagonal and these entries are homogeneous elements.

We will now describe an algorithm for computing the graded Smith normal form. Since the resulting matrix will consist of linearly independent columns and rows they define a basis for the column and row space.

Algorithm 1 Reduction to graded Smith normal form [4], [7]

1: **Input:** Matrix $\in \mathbb{F}[t]^{n\times m}$ with homogeneous entries with columns and rows sorted in order of ascending degree.
2: **Returns:** Matrix $\in \mathbb{F}[t]^{n\times m}$ in graded Smith normal form with columns and rows sorted in order of ascending degree.
3: **for** $i < m, j < n$ **do**
4: Eliminate all entries *rightwards* by column additions in row i.
5: Eliminate all entries *upwards* by row additions in column j.

Proof. (**Correctness of Algorithm 1.**) To show the correctness of the algorithm it suffices to show that we can eliminate entries and that such elimination does not change the degrees of the homogeneous basis elements given by the rows and columns. Since columns and rows are finite, we will eventually terminate.

Suppose we wish to eliminate some entry in row j with an element in row i. This implies that $i > j$ by the algorithm itself. Let the entries be denoted $a_i = c \cdot t^{r_i}$ and $b_j = d \cdot t^{s_j}$ respectively. Since the columns and rows are chosen to be homogeneous, we know that they can be written this way. The matrix is sorted in ascending degree order in both rows and columns, which implies that $\deg(a_i) \leq \deg(b_j)$, hence we can always eliminate b_j by adding $t^{s_j - r_i}\frac{-d}{c} \cdot a_i$. Furthermore, if γ_i, γ_j are the basis elements of rows i, j then we know from linear algebra that the addition causes a change of basis $\gamma_i' := \gamma_i - t^{s_j - r_i}\frac{-d}{c} \cdot \gamma_j$. We get that $\deg(\gamma_i') = \max\{\deg(\gamma_i), (s_j - r_i + \deg(\gamma_j)\}$, but since a_i, b_j are entries of the same homogeneous column this implies that $r_i + \deg(\gamma_i) = s_j + \deg(\gamma_j)$, hence $\deg(\gamma_i') = \deg(\gamma_i)$. An identical argument holds for if we wish to eliminate by column additions, with the exception that it is γ_j that changes basis. □

When a presentation matrix of M is put into graded Smith normal form by Algorithm 1 it gives us compatible bases for K and G and the isomorphism becomes

$$M \cong \frac{\text{basis elements given by rows}}{\text{basis elements given by non-zero rows times their entry}}$$

To understand how we can intrepret the presentation matrix as such, consider that the rows give a basis for G under the inclusion $K \hookrightarrow G$. The non-zero rows tell us which of those basis elements become basis elements of K when multiplied by their corresponding non-zero entry and thus define when they are killed in M.

We are now ready to prove **Theorem 6**.

Proof. The module M is of finite type so it has a presentation matrix P. By reducing P to graded Smith normal form with Algorithm 1 we get a matrix P'. Since $M \cong G/K$ by the first isomorphism theorem we have that M the free part is given by the basis elements of G associated with the zero rows of P', as these are basis elements which are not hit by the inclusion $K \hookrightarrow G$ in any part of the grading. Furthermore, the torsional part in the decomposition is given by the non-zero rows which define basis elements of G that when shifted by their entry are basis elements of K and thus are killed in M. This proves the existence of the decomposition.

What remains to show is that the decomposition is unique no matter our choice of basis for G and K. This actually follows from a much more general result, the Krull-Schmidt Theorem [14, p. 115], but we will give an elementary proof. Suppose that we have two different decompositions of M that we denote C, D. We claim that they are identical up to permutation of components and modulo trivial components. We have homomorphisms

$$f_{ij} := C_i \overset{\iota_{C_i}}{\hookrightarrow} C \cong \overset{\phi}{M} \cong D \overset{\pi_{D_j}}{\twoheadrightarrow} D_j,$$

$$g_{ij} := D_j \overset{\iota_{D_j}}{\hookrightarrow} D \cong \overset{\phi^{-1}}{M} \cong C \overset{\pi_{C_i}}{\twoheadrightarrow} C_i,$$

where the maps are the obvious inclusions and projections from and onto components and the isomorphism through the decomposition of M. Then for at least one pair i, j the homomorphism f_{ij} cannot be the zero map since $\bigoplus_{i,j} g_{ij} \circ f_{ij} : C \to C$ is an automorphism of C. Assume that $i = 1, j = 1$, if this is not the case we can just permute the summands. Then f_{11} (by symmetry all arguments below hold for g_{ij} as well) is an isomorphism $C_1 \cong D_1$ since the graded isomorphism $\bigoplus_{i,j} f_{ij}$ takes a homogeneous element $t^a \to c \cdot t^a$ for some $c \in \mathbb{F}$.

Looking at

$$f_{2_+} := \bigoplus_{i=2,j=2} f_{ij} : \oplus_{i=2} C_i \to \oplus_{j=2} D_j$$

we can deduce that it is also an isomorphism.

For injectivity, note that if $\hat{c} = (c_2, c_3, \dots) \in \ker f_{2+}$ then ϕ has to map $\oplus_{i=2} \iota_{C_i} \hat{c} = (0, c_2, c_3, \dots)$ to $(d_1, 0, 0, \dots)$ for some $d_1 \in D_1$ since the final map is just the projections $\oplus_{j=2} \pi_{D_j}$. By post-composing $\iota_{C_1} \phi^{-1}$ we get

$$\iota_{D_1}(d_1) = \phi \oplus_{i=2} \iota_{C_i}(\hat{c}) \iff$$
$$\pi_{C_1}\phi^{-1}\iota_{D_1}(d_1) = \pi_{C_1}\phi^{-1}\phi \oplus_{i=2} \iota_{C_i}(\hat{c}) \iff g_{11}(d_1) = 0$$

which gives us that $\hat{c} = 0$. Hence, f_{2_+} is injective.

For surjectivity take any element $\hat{d} := (d_2, d_3, \dots)$ in $\oplus_{j=2} D_j$. Since ϕ is an isomorphism there exists some $d_1 \in D_1$ such that

$$\hat{c} := (0, c_2, c_3, \dots) = \phi^{-1}(d_1, d_2, d_3, \dots)$$

for some $\hat{c} \in C$ and so we get

$$\oplus_{j=2}\pi_{D_j}\phi \oplus_{i=2} \iota_{C_i}(c_2, c_3, \dots) = f_{2+}(c_2, c_3, \dots) = \hat{d}.$$

So we get that f_{2_+} is an isomorphism and by symmetry g_{2+} is also an isomorphism. Hence, $g_{2+}f_{2_+}$ is an automorphism and we can repeat the same argument to find new indices i, j such that f_{ij} and g_{ij} are isomorphisms. Since the decomposition is a direct sum, there are a finite number of non-trivial modules and so we will eventually have exhausted all of them. The rest of the summands not covered by these isomorphisms must then consist of trivial modules and so we have proven the uniqueness of the decomposition up to permutation and trivial modules.

□

Now for explicitly computing persistent homology we have the presentation

$$0 \to B \to Z \to H \to 0$$

and so we could theoretically apply the proof above and get a decomposition of H, but this comes with one caveat: to give a presentation matrix of H requires us to first have a basis of Z, which we typically do not have. Instead we have the boundary map $\partial : C \to C$ which has as its image B and Z as its kernel. Hence, we must first compute the Smith graded normal form of the map ∂ which gives us a compatible basis in both B and Z. Then we compute the Smith normal form of the inclusion map $B \to Z$ from which we can derive the barcode $\mathbf{B}_H$.

Example 3.3.1. Consider the filtration given in Figure 3.3. For simplicity we work in $\mathbb{F} = \mathbb{Z}_2$. The persistence complex has basis given by the simplices

$v_0, v_1, v_2, v_{12}, v_{20}, v_{01}, v_{03}, v_{120}$. We get the following boundaries by evaluating the boundary map ∂ on the simplices which generate the chain complex

$$v_2 + v_1, t(v_1) + t^2(v_0), t(v_2) + t^2(v_0), v_3 + t^2 v_0, t^2(v_{12}) + t(v_{10}) + t(v_{20}).$$

Hence we have the following matrix representing the map ∂:

$$\begin{array}{c} \\ v_0 \\ v_1 \\ v_2 \\ v_3 \\ v_{12} \\ v_{20} \\ v_{01} \\ v_{03} \\ v_{120} \end{array} \begin{array}{c} \begin{array}{ccccccccc} v_0 & v_1 & v_2 & v_3 & v_{12} & v_{20} & v_{01} & v_{03} & v_{120} \end{array} \\ \begin{pmatrix} . & . & . & . & . & t^2 & t^2 & t^2 & . \\ . & . & . & . & 1 & . & t & . & . \\ . & . & . & . & 1 & t & . & . & . \\ . & . & . & . & . & . & . & 1 & . \\ . & . & . & . & . & . & . & . & t^2 \\ . & . & . & . & . & . & . & . & t \\ . & . & . & . & . & . & . & . & t \\ . & . & . & . & . & . & . & . & . \\ . & . & . & . & . & . & . & . & . \end{pmatrix} \end{array}$$

Reducing this matrix to graded Smith normal form while keeping track of basis changes gives us the matrix

$$\begin{array}{c} \\ v_0 \\ v_1 + tv_0 \\ v_2 + v_1 \\ v_3 + t^2 v_0 \\ v_{12} \\ v_{20} \\ v_{01} + tv_{12} + v_{20} \\ v_{03} \\ v_{120} \end{array} \begin{array}{c} \begin{array}{ccccccccc} v_0 & v_1 & v_2 & v_3 & v_{12} & v_{20} + tv_{12} & v_{01} + v_{20} + tv_{12} & v_{03} & v_{120} \end{array} \\ \begin{pmatrix} . & . & . & . & . & . & . & . & . \\ . & . & . & . & . & t & . & . & . \\ . & . & . & . & 1 & . & . & . & . \\ . & . & . & . & . & . & . & 1 & . \\ . & . & . & . & . & . & . & . & . \\ . & . & . & . & . & . & . & . & . \\ . & . & . & . & . & . & . & . & t \\ . & . & . & . & . & . & . & . & . \\ . & . & . & . & . & . & . & . & . \end{pmatrix} \end{array}$$

From the reduction to normal form we can read off a basis for Z given by the zero columns and the basis of B given by the non-zero rows together with their non-zero entry. This gives us a presentation matrix of H from the inclusion $B \hookrightarrow Z$

$$
\begin{array}{c}
\\
v_0 \\
v_1 \\
v_2 \\
v_3 \\
v_{01}+v_{20}+tv_{12}
\end{array}
\begin{array}{c}
\begin{array}{cccc}
v_2+v_1 & tv_1+t^2v_0 & v_3+t^2v_0 & tv_{01}+t^2v_{12}+tv_{20}
\end{array} \\
\begin{pmatrix}
. & t^2 & t^2 & . \\
1 & t & . & . \\
1 & . & . & . \\
. & . & 1 & . \\
. & . & . & t
\end{pmatrix}
\end{array}
$$

Reduction of this matrix to Smith normal form gives us

$$
\begin{pmatrix}
. & . & . & . \\
. & t & . & . \\
1 & . & . & . \\
. & . & 1 & . \\
. & . & . & t
\end{pmatrix}
$$

We see that there is one zero row given by v_0 which gives us the interval $[1, \infty)$. This is because v_0 is the 0-cycle in Figure 3.3 that eventually becomes the connected component of the entire simplicial complex, hence it never dies. We additionally have two intervals $[2, 3)$ and $[3, 4)$ from the rows with t as their entry. The first one is the other connected component given by $v_1 + v_2$ which is born at filtration step 2 and dies when it becomes part of the single connected component given by v_0 in filtration step 3. The interval $[3, 4)$ corresponds to the triangle without interior at filtration step 3 which later dies at filtration step 4 when the triangle is filled in.

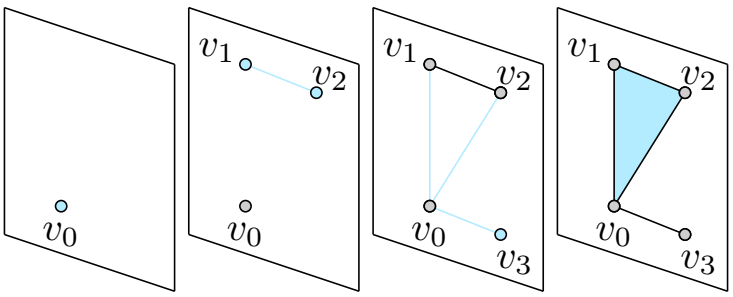


Figure 3.3: *Filtration of a simplicial complex showing the intermediate simplicial complex at each filtration step. A blue simplex indicates the simplex was added in that filtration step.*

Algorithm 1 for a simplicial complex has, just like ordinary Gaussian elimination over fields, worst case time complexity $O(m^3)$ where m is the number of simplices [4]. However, as seen in Example 3.3.1 the boundary matrix is sparse. Furthermore, the decomposition of H can be read entirely from the reduced boundary matrix without constructing an explicit presentation matrix. These are some areas where the algorithm usually is made more efficient, but dwelling on such optimizations is outside the scope of this thesis. For more in-depth treatments on this subject see [9] for the theoretical underpinnings and [15] for the de facto solution on which many software libraries are based.

3.4 Visualizing Persistence

The persistent homology of a space is not a very easy algebraic object to work with in practical terms. Even when considered under the bijection with intervals it is a multiset of intervals and as such helpful visualizations allow us to analyze and compare persistent homology. There are two principal ways of visualizing the decomposition of a persistence module: barcode diagrams and persistence diagrams.

3.4.1 Barcodes

A **barcode diagram** is a visual depiction of $\mathbf{B}_H$ where each bar depicts the start end and end of an interval, or equivalently the birth and death of a particular generator in one of the homology modules.

In Figure 3.4 we see a barcode diagram generated from points sampled from an annulus. Note that for small values of ϵ there are many generators of H_0, this is because the vertices have not been connected into a single component yet.

We see that there some short intervals appearing for H_1 at around $\epsilon = 0.3$ and we can see that these are not the hole that would represent the annulus, but rather noise that appears before ϵ has become large enough. At around $\epsilon = 0.6$ the simplicial complex now captures the shape of the annulus and indeed the barcode diagram shows that we have one generator of H_0, the only connected component, and one generator of H_1 which is the hole in the middle of the annulus.

Note how this hole in the middle of the annulus is gone when $\epsilon = 1$ which highlights that it is difficult to find an optimal ϵ.

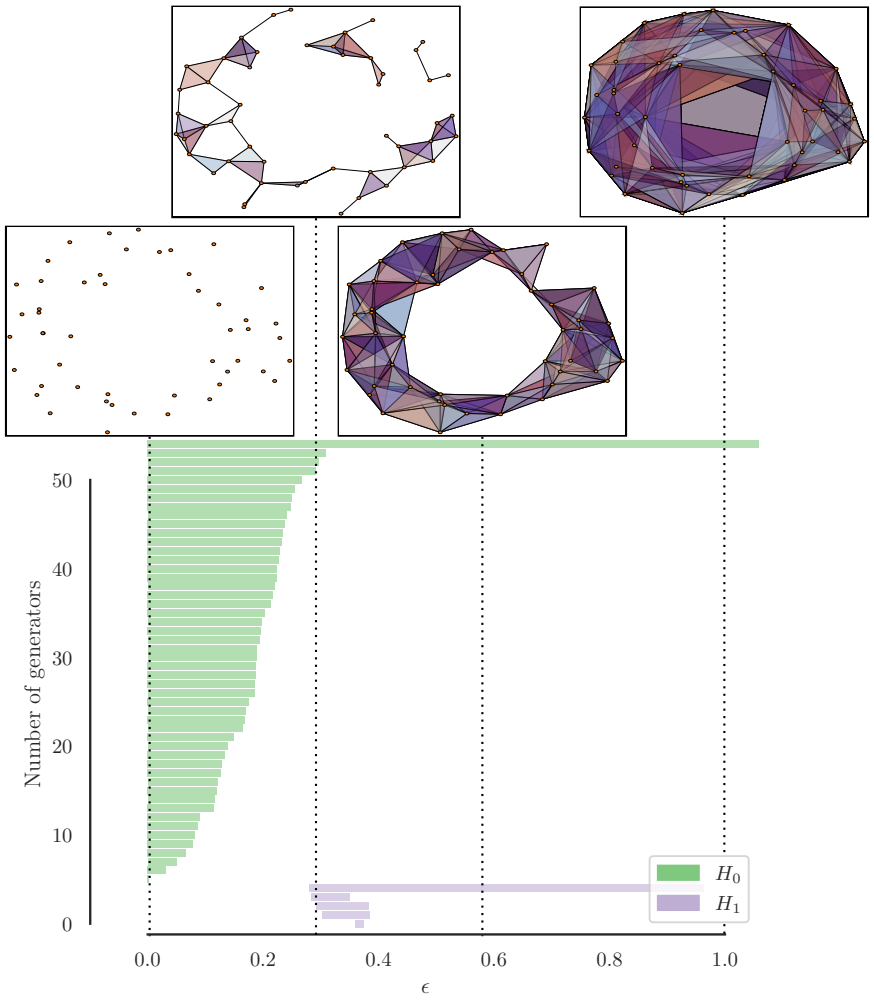


Figure 3.4: *Persistence barcode showing the birth and death of generators in the homology groups of a Vietoris-Rips complex approximated from points sampled from an annulus at different ϵ.*

3.4.2 Persistence Diagrams

Another way of illustrating persistent homology is the persistence diagram as seen in Figure 3.5.

Definition 3.4.1. The **persistence diagram** X of a persistence module M is a multiset of points in $\mathbb{R}^2 \cup \{\infty\}$ defined as

$$X := \{(x, y) \in \mathbb{R}^2 \mid [x, y) \in \mathbf{B}_M\} \cup \{(x, x) \mid |x \in \mathbb{R}\}.$$

In other words, it is the set of *(birth,death)* pairs given by the intervals associated with the decomposition of M together with all points on the diagonal.

When visualized as in Figure 3.5 it serves alternative to the barcode in Figure 3.4 where we instead plot the ϵ-value on both axes and for each generator we draw a point given by its corresponding interval. When we have a lot of intervals this is a preferable way of visualizing the persistent homology, since unlike the barcode it does not grow vertically with the number of intervals. Generators that never die are mapped at a line representing infinity.

Just like in the barcode in Figure 3.4 we can see in Figure 3.5 that the only two generators that live for a considerable amount of time is a single connected component in H_0 and a single hole in H_1. This is consistent with the topology that we expect from an annulus. At around $\epsilon = 0$ we see a lot of H_0 generators being born and dying at almost the same time. Since the number of generators of H_0 tells us the number of connected components in the topology this clearly illustrates how the sampled points go from being isolated islands to being incorporated in a larger simplex.

3.5 Metrics

As persistent homology is often used as a topological summary of some data, it can be beneficial to be able to compare two different data samples with respect to their persistent homology. There are two commonly used metrics for doing this, the bottleneck distance and the Wasserstein distance.

Definition 3.5.1. The **bottleneck distance** between two persistence diagrams X, Y is

$$W_\infty(X, Y) = \inf_{\beta: X \to Y} \sup_{x \in X} ||x - \beta(x)||_\infty$$

where β is a bijection from X to Y.

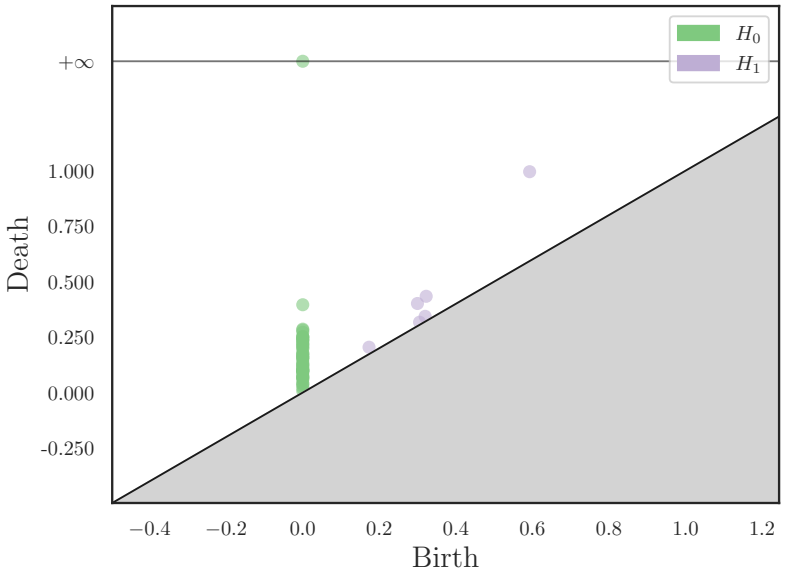


Figure 3.5: *A persistence diagram over the birth and death of generators in the homology groups of a Vietoris-Rips complex approximated from points sampled from an annulus. The closer a point is to the diagonal line the shorter it lived.*

In other words, the bottleneck distance finds a matching, in the space of possible matchings, between the two persistence diagrams such that the largest distance in the matching is the smallest one possible. Since there could be more intervals in one persistence diagram, any point can also be matched with an infinite number of points on the diagonal which are included in the persistence diagram. Its name is derived from the fact that there is only one matching of points which contributes to the actual value of the distance, the largest one, and hence the distance is "bottlenecked" by that matching.

One disadvantage of the bottleneck distance is that it is quite coarse, it does not tell us very much about the other distances between other matched points. An alternative is the q-Wasserstein distance which instead incorporates all distances in the best matching.

Definition 3.5.2. The q-Wasserstein distance between two persistence diagrams X, Y is

$$W_q(X,Y) = (\inf_{\beta: X \to Y} \sum_{x \in X} ||x - \beta(x)||_\infty^q)^{\frac{1}{q}}$$

Both the Wasserstein and bottleneck distance have their role as the bottleneck distance can be considered more robust to noise, since small changes in the matchings are ignored in favor of the largest matching.

A desired quality of these metrics are stability, ideally we want the metrics to actually reflect difference in the underlying spaces. There are stability theorems that under varying conditions fulfill a meta-theorem which guarantee this.

Definition 3.5.3. A filtration function $X \to \mathbb{R}$ is a function from a simplicial or cubical complex X such that the sublevel set $\{f^{-1}(-\infty, a) \mid a \in \mathbb{R}\}$ is a filtration.

Theorem 8 (Stability meta-theorem [6])**.** *For a nice enough space X and nice enough filtration functions $f, g : X \to \mathbb{R}$, a nice enough norm of the difference of $f - g$ serves as an upper bound to the distances between the persistence diagrams given by f, g.*

In other words, a small perturbation of the filtering functions will at most be as large as the difference between the functions themselves. The term *nice* is intentionally vague, since these conditions vary. For an overview of the particular scenarios in which the meta-theorem is applicable see [6], including a formulation which allows for general persistence modules under certain conditions. In our practical applications we are only considering finitely many sublevel sets from a filtration function which motivates the following definition.

Definition 3.5.4. A filtration is called **tame** if the persistence complex arising from the filtration function is of finite type.

Then we have the following corollary of the meta-theorem.

Corollary 8.1. *If f, g given as in Theorem 8 are tame then the meta-theorem holds for the bottleneck distance with norm given by the L_∞-norm and for the q-Wasserstein distance with the L_q-norm.*

As the proofs of Corollary 8.1 require a lot of tedious and technical details which would detract from the overall theme of this thesis, we instead refer the reader to elementary proofs regarding the q-Wasserstein distance in [16] and the bottleneck distance in [17].

4 Two Applications of Persistent Homology

Since our purpose with thesis is not only to give an introduction to persistent homology in terms of theory, but also display how it can be used with actual real-world data, we illustrate this pipeline two different case studies where persistent homology serve as our main tool for data analysis.

In the first case we quantify differences in morphology between different-sized individuals of the bumblebee *Bombus terrestris* by computing the persistent homology of 3D volumes of their corneas. To our knowledge this is the first use of persistent homology in data pertaining to insects, although materials [3], [18], reconstructions of 3D volumes [19], [20] and plants [2] have been investigated with approaches that are similar in spirit.

In the second case we try to understand the network structure of the striatum, a part of the basal ganglia in the brain. Due to the sheer computational power needed to compute persistent homology for this data our analysis is more of a holistic summary of the resulting simplicial structure rather than focusing solely on persistent homology. Our approach is largely inspired by [1], in which a similar analysis is done but for a different part of the brain.

The application of persistent homology to a data-set is not entirely trivial. In order to compute persistent homology we need to construct a persistence complex on the data at hand. This can be done in a multitude of ways, but the basic pipeline can be seen in Figure 4.1.

A central part of this pipeline is the formulation of a filtration function which yields a sequence of complex and hence a persistence complex. The filtration provides the translation of data into an algebraic object we can compute persistent homology of. This means that when we analyze results from persistent homology, perhaps by comparing metrics between two barcodes or reading directly of a persistence diagram, the semantic meaning of those results is intimately connected to how the filtration creates the resulting complex. Dually, this means that in order to formulate the filtration we need to have a

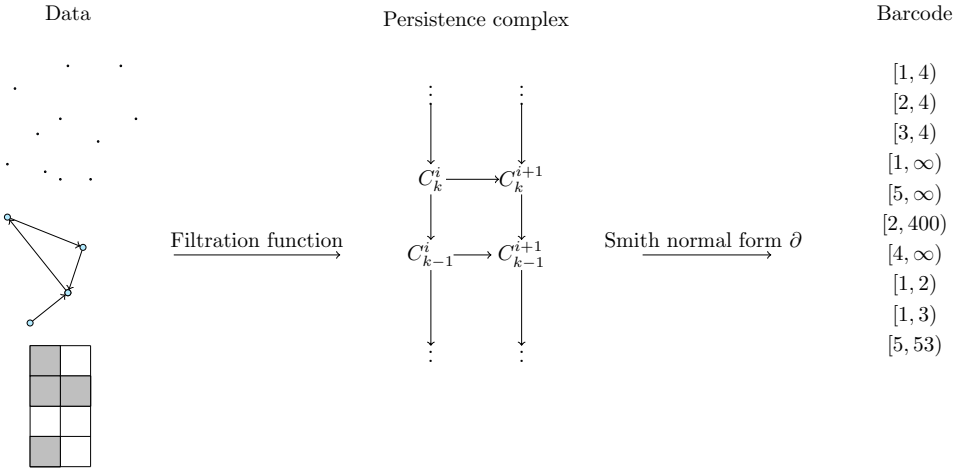


Figure 4.1: *The pipeline of computing persistent homology of data.*

strong understanding of the data itself so that our filtration captures essential properties of the data. The complex constructed on the data-set is an approximation of it topologically, but since it is not the "true" space in which the data lives care has to be taken so that any conclusions made from the barcode are meaningful.

4.1 Corneas of Bombus terrestris

It has been found that the size of individuals of the species *Bombus terrestris* affects aspects of their visual capabilities [21]. By applying persistent homology we can investigate whether this difference in size also translates to a difference in persistent homology, and so by proxy a difference in topology. If so, this could serve to strengthen the hypothesis that larger individuals have superior, or at the very least different, visual capabilities than smaller individuals. Persistent homology is a good candidate for this purpose as metrics on persistence diagrams are indifferent to differences in scale but rather measures differences in shape.

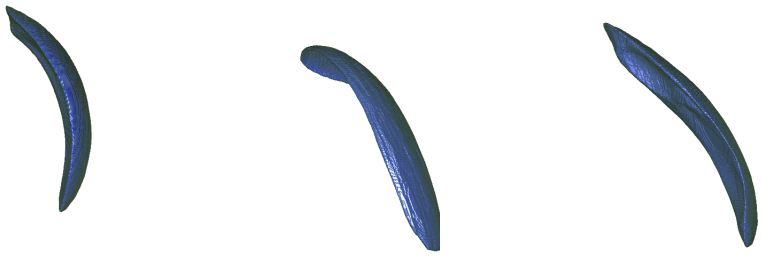
Our questions we wish to investigate in this case study are

1. Is there a correlation between the size of the bumblebees and their persistent homology?

2. Can we with persistent homology identify subgroups of bumblebees, and if so are these subgroups related to their size?

4.1.1 Data

The data consists of binary 3-dimensional volumes (see Figure 4.2 for renderings of some of the samples) of the corneas acquired by micro-CT scans of the samples described in Table 4.1. The main focus of the analysis will be on samples from the bumblebee *Bombus terrestris*, but in total there are 20 samples belonging to 8 different species of insects. The additional samples from other species will be used to verify our topological findings.



(a) *TA_60204* **(b)** *MQ_60209* **(c)** *BT_77970*

Figure 4.2: *Example renderings of cornea volumes.*

4.1.2 Methodology

Since the data we are working with are 3-dimensional volumes a natural choice is to endow it with the structure of a cubical complex. Our strategy is similar to the Vietoris-Rips complex, but instead of working with distances between points we work with adjacent cubes. Each voxel can be considered as a degenerate interval (or vertex) in a cubical complex, where 4 pairwise adjacent voxels give rise to a square and 8 pairwise adjacent voxels give rise to a cube.

In order to compute persistent homology we need a filtration which defines subcomplexes of the cubical complex. Since the volumes are binary, we give the volume a bit more structure by giving each voxel the distance to the closest point on the boundary of the volume.

ID	ITW	Species
AM_60185	2.90	Apis mellifera
AM_60186	2.95	Apis mellifera
BT_77967	5.42	Bombus terrestris
BT_77970	4.00	Bombus terrestris
BT_77971	4.02	Bombus terrestris
BT_77973	1.97	Bombus terrestris
BT_77974	2.97	Bombus terrestris
BT_77976	5.47	Bombus terrestris
MB_60160	3.25	Melipona bicolor
MB_60161	3.25	Melipona bicolor
MQ_60208	3.64	Melipona quadrifasciata
MQ_60209	3.64	Melipona quadrifasciate
PR_60164	1.49	Plebia remota
PR_60206	1.49	Plebia remota
TA_60204	1.17	Tetragonista angustula
TA_78016	1.17	Tetragonista angustula
TC_60166	1.94	Tetragona clavipes
TC_60167	1.94	Tetragona clavipes
TS_60163	2.10	Trigona spinipes
TS_60203	2.10	Trigona spinipes

Table 4.1: *Table over the data samples used in the analysis. The ID column gives a unique ID to each sample and the ITW column gives the intertegular width of each sample.*

Definition 4.1.1. Given a subset $Y \subset \mathbb{R}^n$ we define the Euclidean Distance Transform, or EDT, as

$$EDT(x) = \inf_{y \in \partial Y} ||x - y||_2$$

where ∂Y is the boundary of Y.

Our filtration is a sequence of cubical complexes K_i given by including voxels of at most value ϵ_i as degenerate intervals. Higher dimensional cubes such as edges, squares and geometric cubes are, similarly to the Vietorix-Rips complex, included whenever there is a sufficient amount of pairwise adjacent voxels.

Example 4.1.1. To calculate the EDT of the binary image in Figure 4.3 we simply calculate the difference vector from a pixel of value 1 to the closest

pixel with value 0. For example, to get $\sqrt{5}$ in the top left corner we need to walk one step in to the left and two steps upwards which translates to the vector $(-1, 2)$ which has Euclidean norm $\sqrt{1+4} = \sqrt{5}$. We then compute the filtration of the transformed image, which gives us cubical complexes for each value of ϵ. Since there are five distinct values of the pixels, we find five subsequent cubical complexes ordered by inclusion.

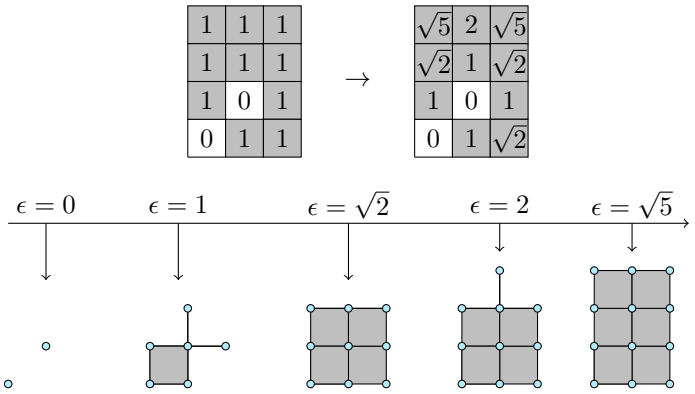
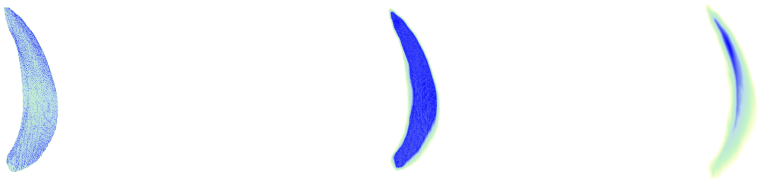


Figure 4.3: *Transformation of a binary image into a filtration of cubical complexes based on the Euclidean Distance Transform.*

Our filtration on the volumes will describe the structure of the cornea starting at it the void surrounding it, then including the hollow shell which is its boundary, and then as the threshold increases the cubical complex will include more and more of the denser parts within the volume. An illustration of the thresholding at different values is seen in Figure 4.4.



(a) $\epsilon < 2$ **(b)** $\epsilon < 10$ **(c)** $\epsilon < 100$

Figure 4.4: *EDT thresholding of BT_77976. Cooler colors indicate denser parts of the volume relative to the rest of the volume.*

The resulting topological summaries we get are persistence diagrams. While these are in themselves interesting, in order to answer whether there is any

relation between the size of an individual and the persistent homology of its cornea, we compute a distance matrix giving the distance of each sample to another. Since we have a number of dimensions of homology to compare and two different metrics (1-Wasserstein and Bottleneck) we get a total of 6 distance matrices. For each of these distance matrices, we divide them into two submatrices, one group consisting of the submatrix with entries from *Bombus terrestris* and the other group consisting of all the other samples to act as a control group on the first.

Our hypothesis is that there is some relationship between the distances given by the absolute difference in ITW and their persistent homologies. But since different homology dimensions and metrics provide different summaries of the objects we first need to figure out which one of them is most suited for our applications.

We determine which metric we will rely on by doing a so called **Mantel test**. A Mantel test is a non-parametric test of distance matrices, in which we compute the Spearman rank correlation of the two matrices under the null hypothesis that the matrices are uncorrelated. We can then derive a test statistic by permuting one of the matrices in both rows and columns and computing correlations for each such permutation. The results of a Mantel test is a correlation coefficient indicating the strength of the correlation and p-value indicating how likely it is that this coefficient would appear in a random permutation of one of the matrices.

We then perform clustering of the persistent homology distance matrix based on **hierarchical clustering**. It is a simple algorithm where we first consider each sample as its own cluster, and then group together clusters depending on the distance between them. The distance between two clusters is given as the minimum distance between any two samples between the two clusters. Our hope then is that the final clustering reflects the ITW of the *Bombus terrestris* samples.

4.1.3 Results

The Mantel tests in Table 4.2 reveal that the highest correlation is given by the bottleneck distance on H_2. Perhaps this is not too surprising, our objects are volumes and the most distinguishing aspects of volumes will be how they encode voids. Interestingly, the H_1 bottleneck distance matrix shows a very high p-value indicating that the largest distance between holes is not very telling in drawing a conclusion about correlation between size and persistent

homology. This could be explained again by the fact that our object is a volume of a single connected component namely a cornea, and so any existence of holes will at best be local geometric information and at worst simply noise.

Group	Metric	$H_1\rho$	H_1 p-value	$H_2\rho$	H_2 p-value
BT	Bottleneck	-0.069	0.84	0.86	0.0083
BT	Wasserstein	0.59	0.053	0.49	0.080
Others	Bottleneck	0.22	0.030	0.33	0.0073
Others	Wasserstein	0.23	0.023	0.26	0.013

Table 4.2: *Table displaying the statistics computed in the Mantel test of the pairwise distances in different dimensions of persistent homology and the ITW for the species Bombus terrestris. The symbol ρ denotes the Spearman rank correlation coefficient computed between the two distance matrices.*

The 1-Wasserstein metric does not provide a low enough p-value for us to draw any conclusions from the tests when it comes to *Bombus terrestris*, this perhaps indicates that the sample size of the *Bombus terrestris* submatrix is too small. It is possible that the more sensitive nature in the 1-Wasserstein metric, since it records not only the largest differences in the persistence diagrams but all of them, makes it less robust to a smaller sample size.

Since the bottleneck distance on H_2 has a strong correlation with the ITW distance matrix we proceed with a hierarchical clustering on its distance matrix as seen in Figure 4.5. We see that there are two groups formed where one of the groups contains an additional cluster. The first group, which does not contain a subgroup, consists of two samples BT_77967 and BT_77976. These are the largest samples in terms of ITW and the remaining group contains two subgroups both in which the ITWs are smaller. It is worth repeating that this clustering is done without knowledge of the actual ITW, these clusterings are purely based on the bottleneck distance of the H_2 persistent homology and as such a differences here indicate differences based solely on the fact that their persistence diagrams differ.

In order to further clarify in what way the persistence diagrams of the *Bombus terrestris* are differing we can look at visualizations of the bijections done under the bottleneck distance and which pair of generators are the ones to determine the metric.

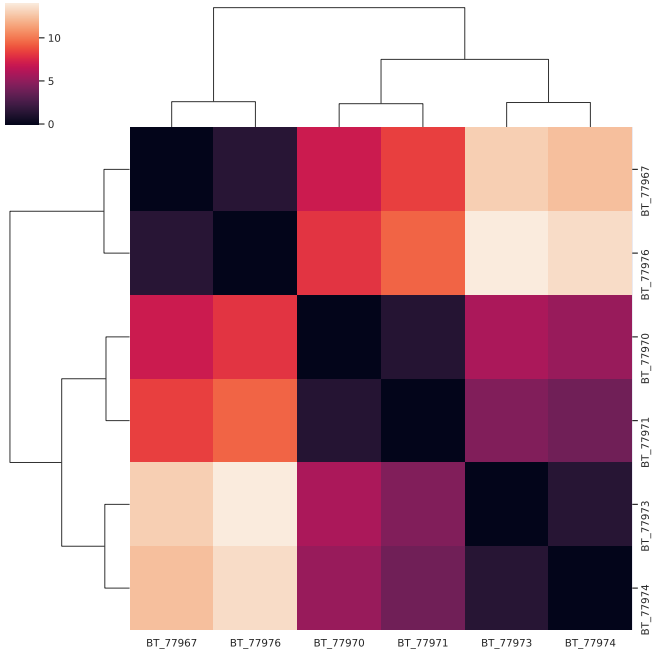


Figure 4.5: *Hierarchical single-link clustering of the bottleneck distance matrix derived from the persistent homologies of the Bombus terrestris in H_2.*

In Figure 4.6 we see the matchings produced for the elements *within* each of the identified clusters. We see that within the group the optimal matching always gives the largest distance as a matching between a point and the diagonal.

On the other hand, if we look at the distances from samples *between* groups in Figure 4.7 we see that there is one generator that is responsible for the bottleneck distance in all of these matchings. That is the longest living generator which is born at filtration value 0. At filtration value 0 the only voxels in the volume are the empty voxels constituting the background, which means that the void is left by the space where the cornea will be at a higher filtration value. As the filtration value increases the volume gets filled in with denser and denser parts of the cornea, but as seen in Figures 4.6 and 4.7 it lives for a long time before the cornea is entirely filled out. This difference in lifetime, which can somewhat be translated to the density of the cornea, appears to be the distinguishing factor between the three identified clusters of samples.

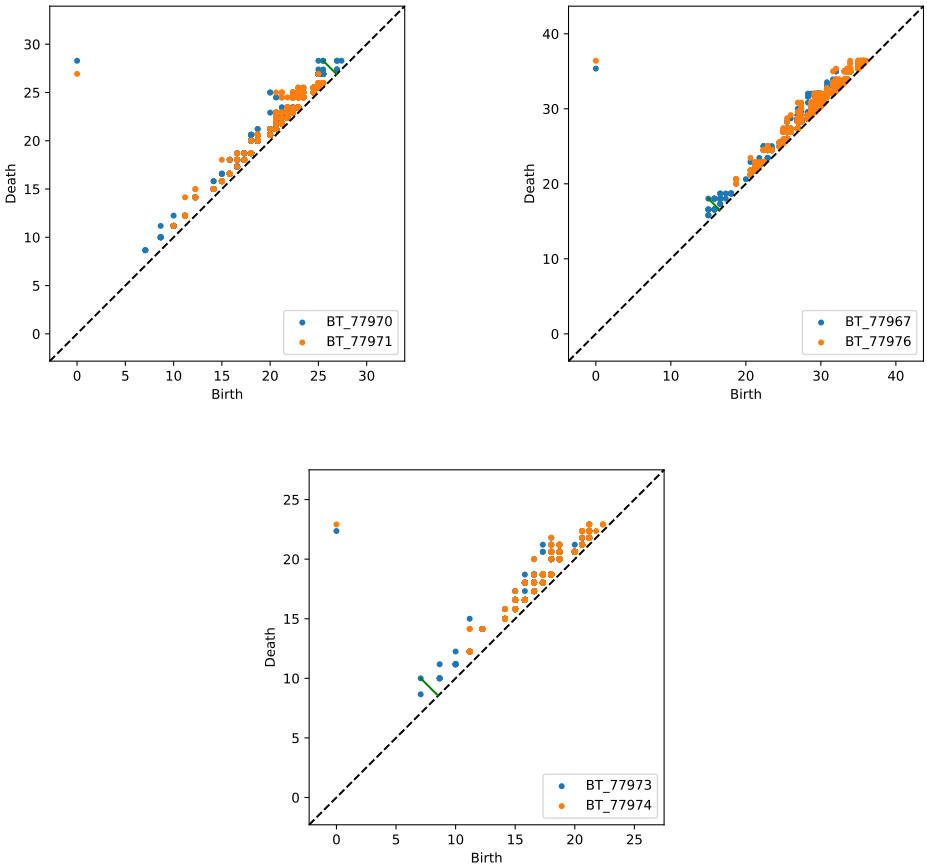


Figure 4.6: *Visualisations of bottleneck distances within clusters on persistence diagrams of H_2.*

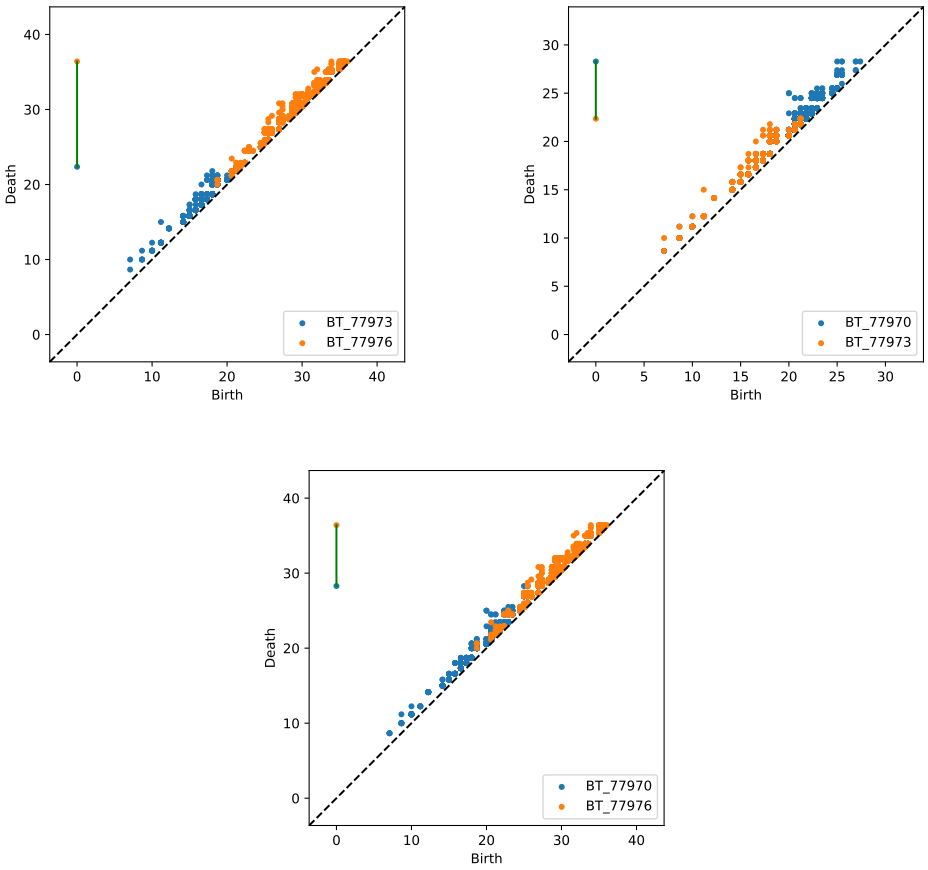


Figure 4.7: *Visualisations of bottleneck distances between clusters on persistence diagrams of* H_2.

While the purpose of this analysis is mostly to showcase persistent homology in the wild, these results do support to the idea that different sized *Bombus terrestris* do not only have larger eyes, but also that they are topologically different. Our clustering results on H_2 group the larger individuals together and we find a strong correlation with bottleneck distance in H_2 and ITW. Furthermore, we find that there is one generator born at the beginning of the filtration that is responsible for the "bottleneck" when comparing samples from the different clusters whereas within the same cluster other generators with much shorter lifetime are the contributors.

4.2 The Simplicial Structure of the Striatum

There is much reason for considering simplicial complexes when it comes to networks related to the brain. An extensive field of study when it comes to network analysis of the brain are motifs, which at a high-level simply means a repeated pattern in a network which has some form of semantic meaning to the network. Simplices capture such patterns at a micro-level through the connectivity of its faces. Furthermore, homology captures another type of patterns at a meso-level through cycles of simplices. Armed with persistent homology we arrive at a comprehensive summary of the network through the lens of the filtration of our choice.

In this analysis we follow [1] by observing the high-dimensional simplicial structure of the brain by constructing a particular simplicial complex on the connectivity matrix of a synthetically generated microcircuit of the striatum, a part of the basal ganglia in the brain. By computing the resulting persistent homology we hope to discover distinguishing features of our network from a few selected control models. Furthermore, our aim is that this analysis serves as yet another example of the breadth of possible scenarios in which persistent homology is applicable outside the realm of theory. While we only investigate the mentioned microcircuit, our methodology is general enough for it to be applicable in any scenario where we data can be interpreted as directed graphs.

4.2.1 Data

Recall that a directed graph $G = \{V, E\}$ consists of a set of **vertices** V and a set of **edges** E, where an edge is an ordered set (v_i, v_j) for some vertices $v_i, v_j \in V$. The **degree** of a vertex in a directed graph is the sum of the number of outgoing and incoming edges from and to the vertex.

In this case analysis our main object of study is a synthetic network of generated based on empirical findings regarding the micro-circuitry of the striatum realized as a directed graph, see [22] for further details. We compare this network to a three different models of directed graphs that all have different qualities common to networks of the brain.

Definition 4.2.1 ([23])**.** The **Erdős–Rényi (ER) model** generates a directed graph through the choice of two parameters: the number of vertices n and the number of edges m. The graph is then selected uniformly from the set of all graphs with n vertices and m edges.

Definition 4.2.2 ([24])**.** The **Watts-Strogatz (WS)** model is parameterized by the number of vertices n, the average out-degree m and a rewiring probability p. A directed graph is constructed by first creating a graph of n vertices, such that for every vertex there is an outgoing edge to its m closest neighbours modulo n, meaning that the graph is circular. Then finally each edge has a probability p of being rewired to a different, uniformly selected, endpoint.

Definition 4.2.3 ([25])**.** The **Barabási–Albert (BA) model** is parameterized by the number of vertices n and m, the average out-degree of each vertex. It is constructed by first adding m vertices and successively adding vertices until there are n vertices. For each new vertex v_i we add an edge $\{v_i, v_j\}$ with a probability of

$$\frac{\deg(v_j)}{\sum_k \deg(v_k)}$$

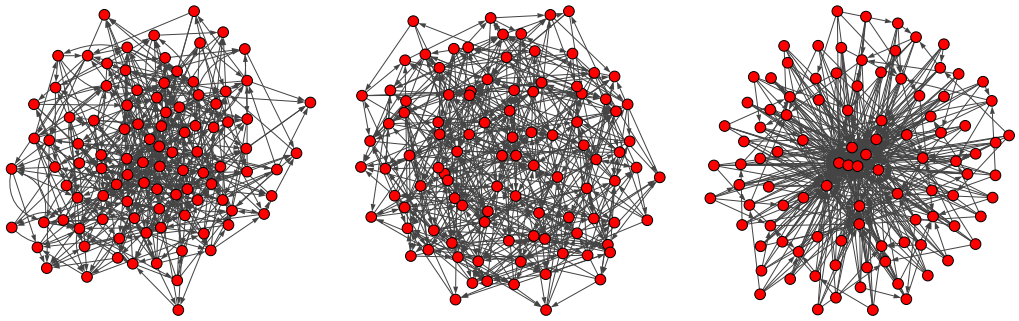
Hence, at each addition of a vertex an older vertex with a high degree has a higher chance of having its degree increased.

ER can be considered the baseline model, since it is an arbitrary random graph among all possible graphs. WS is said to have **small-world properties**, which means that there are clusters of highly connected nodes and that the average path between two vertices is short. BA is said to be **scale-free** which means that most vertices have a low degree, but some "hubs" have a high degree. Both scale-free and small-world properties have been observed in brain networks [26]. Figure 4.8 gives an example of each model generated with 100 vertices.

In this analysis we compare we will compare the three models above with the synthetic network generated from the striatum, which we will refer to as ST. In Table 4.3 we can see our choice parameters for the models and the resulting number of edges and vertices. For ER and WS our parameter choices were made to match the number of edges in ST. However, for the BA model we instead choose parameters in order to match the number of dimensions in the simplicial complex on ST as seen in Figure 4.10.

4.2.2 Methodology

It is not entirely clear how we should define a simplicial complex on a directed graph. The asymmetry is important as connections between neurons in the brain do not necessarily go both ways. While we could simply consider the network as a undirected graph by adding missing edges it is likely that important qualities of the network might be lost. There are a number of different



(a) *Erdos-Rényi* **(b)** *Watts-Strogatz* **(c)** *Barabási-Albert*

Figure 4.8: *An example of the ER, WS and BA models on 100 vertices.*

Network	Vertices	Edges	Parameters
ST	50000	12298074	-
ER	50000	12298074	$n = 50000, m = 12298074$
WS	50000	12300000	$n = 50000, m = 247, p = 0.5$
BA	50000	849864	$n = 500000, m = 17$

Table 4.3: *Breakdown of the ST, ER, WS and BA graphs compared in the analysis.*

ways to go about defining a similar complex on a directed graph . In our case we follow [1] and construct something called the *directed flag complex* of a directed graph.

Definition 4.2.4 ([27])**.** A **directed clique** is a directed graph $G = (V, E)$ such that every vertex has at least an outgoing or incoming edge to every other vertex in the graph.

Recall that a simplicial complex includes a partial order that descends to a linear order on each simplex. Hence, if we can construct such a partial order on our graph then we can define a simplicial complex from there.

Definition 4.2.5 ([27])**.** Let G={V,E} be a directed graph. The **directed flag complex** dFl(G) is defined to be the simplicial complex whose k-simplices are all directed cliques with vertices $v_0, \ldots, v_k$ such that $\forall i : v_i \in V$ and $\forall i, j : i < j \implies (v_i, v_j) \in E$. The vertices v_0, v_k are called the source and the sink of a k-simplex.

This construction is essentially the same construction given by the Vietoris-Rips complex where pairwise intersection is given by an edge. However, there are some notable difference due to the fact that the underlying graph is directed. One difference is that we only create a simplex from cliques of simplices whose edges flow "upwards" in the order. In Figure 4.9 we see how the left clique has a source and sink and thus defines a 3-simplex in the directed flag complex, but the right clique has one edge going in the wrong direction hence it is not a 3-simplex.

Another difference given by the fact that we have directions in the underlying graph, which for example results in that the simplices $[v_0, v_1]$ and $[v_1, v_0]$ given by a bidirectional edge in the graph are two different simplices and hence make up a cycle $[v_0, v_1] + [v_1, v_0]$.

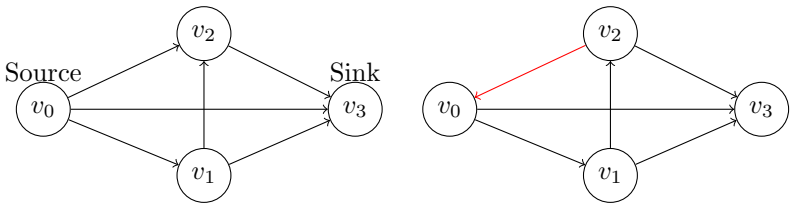


Figure 4.9: *Two directed cliques where the left clique does create a 3-simplex in the directed flag complex and the right clique does not.*

We investigate the graphs ST, ER, WS, and BA in three different ways: the number of simplices in each dimension, their Betti numbers and their persistence diagrams.

For computing persistent homology, we need to define a filtration on the directed flag complexes. We use a simple filtration function

$$f(\sigma) = \begin{cases} -\deg(\sigma) & \text{if } \dim(\sigma) = 0 \\ \max_{\tau \text{ is a face of } \sigma}\{f(\tau)\} & \text{if } \dim(\sigma) > 0 \end{cases}$$

where we add vertices in the reverse order of their degrees and their resulting higher-dimensional simplices whenever all their vertices have been included in the complex. The degree of the vertex indicates how central a vertex is to the graph, which means that the filtration will describe the network in its most basic building blocks and then add less and less important vertices.

4.2.3 Results

First we note the large number of simplices in ST as seen in Figure 4.10, where we find as much as over a hundred billion simplices in dimension $7, 8$ and 9. This phenomena was also observed in [1], however not of the same magnitude. Furthermore, we see that the number of dimensions is significantly larger in ST compared to ER and WS. The graph BA was chosen so that its directed flag complex had the same dimension as ST's, but we see that the number of simplices in each dimension is notably lower. So while the BA model seems to capture the complexity of connectivity in ST it does not capture the magnitude. ER and WS present a lot fewer simplices and a lot fewer dimensions than ST. In this sense ER is the simplest model, but this is to be expected since it is a random graph we expect there to be less of simplicial complexity.

One thing that has to be mentioned is that the computation of homology and even more so persistent homology becomes extremely computationally expensive due to the number of simplices in ST. As seen in Section 3.3 reducing the boundary matrix in a dimension is at worst case given by a cubical amount of field operations in proportion to the number of simplices in one dimension above. Hence, computing something like β_7 for ST would involve a computation of the magnitude $10^{3 \cdot 11}$ field operations. Even worse, persistent homology is not only affected by the number of simplices in the resulting simplicial complex, but also the number of simplicial complexes in the filtration. For this reason we only present persistence diagrams and Betti numbers for H_0, H_1 and additionally the Betti number for H_2.

In Table 4.4 we see that ST has a lower β_1 than any of the other graphs. Curiously, even though BA has a lot fewer edges it still has a higher Betti number in dimension 1. It is possible that the low amount of 1-cycles is a feature that distinguishes brain networks from other types of networks, but no such result is presented in [1] since β_1 was never computed due to computational limitations. If we turn to β_2 we see that ST has the largest change from β_1, while the other graphs are pretty close to their β_1 however with a larger error. This dramatic increase from dimension 1 to dimension 2, both in Betti number and number of simplices, could also be seen as a feature of ST compared to the other models. The distribution of simplices in Figure 4.10 indicates that this dramatic increase continues for several dimensions.

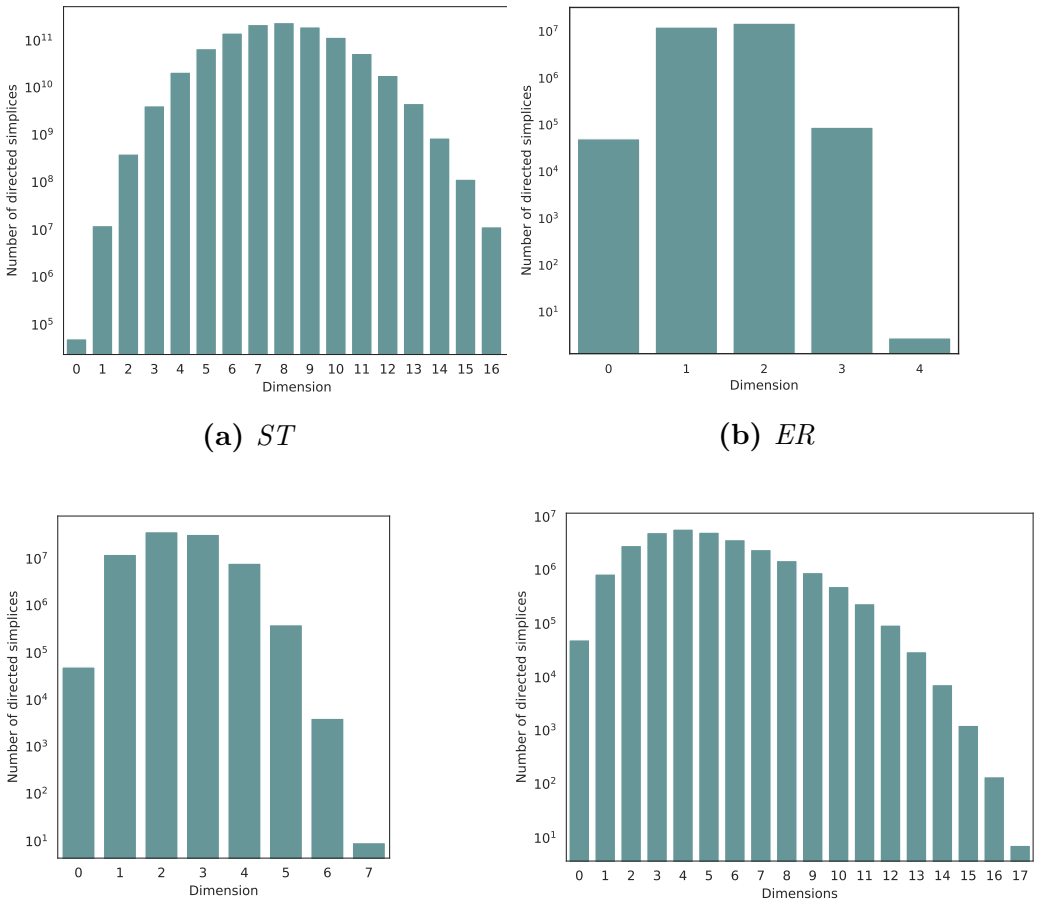
We see in Figure 4.11 that ST is markedly different from all other models in terms of persistent homology. The persistence diagram of BA is not very informative due to all of its vertices essentially having the same degree. However,

WS and ER display similar persistence diagrams in which the distribution of holes is concentrated to a small interval of filtration steps from degree 500 to degree 400. In comparison, ST shows a much larger spread of both birth and death of holes, with some holes even being born or dying close to degree 0.

To conclude, we have found that directed flag complex on *ST* differs in several ways from the other three graphs: it has a much larger number of simplices in each dimension, it has simplices in high dimensions, it has lower β_1 than all models and lower β_2 than all models but BA. Furthermore, the persistence diagram of ST shows a distinct spread of the birth and death of generators of H_1, whereas for BA they are all born at the same time due to nature of the filtration and for ER and WS they are all restricted within a small interval of degrees.

Network	β_1	β_2	Computations
ST	5128 ± 1219	293013 ± 79867	1
ER	369770.18 ± 687.69	3470380.91 ± 2774.23	100
WS	334754.93 ± 625.38	3756595.22 ± 3709.97	100
BA	11982.51 ± 956.18	12946.77 ± 1734.00	100

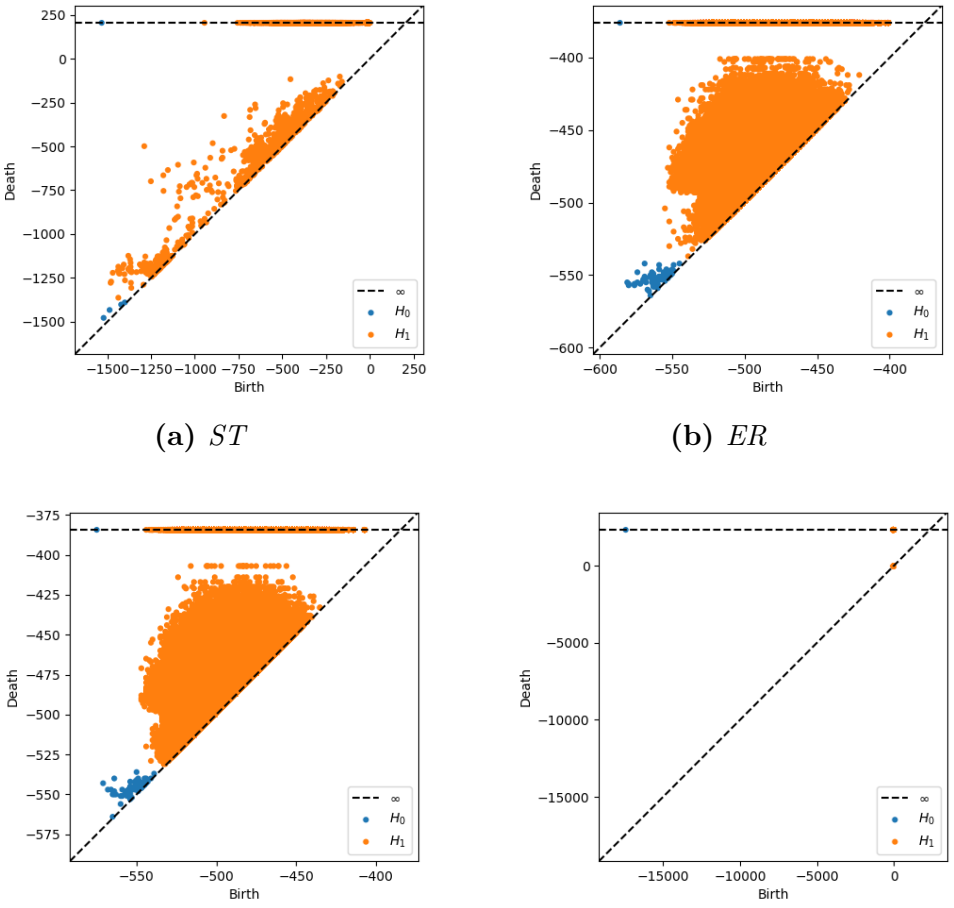
Table 4.4: *Table over the first and second Betti numbers for the graphs ST, ER, WS and BA. The computation column denotes the number of instances of the model over which the result was averaged, hence the Betti numbers are given by an average value and its standard deviation. For ST the approximation seen in [27] was used in order to reduce computational time in which some columns were not reduced to Smith normal form. Every such non-reduced column can at most subtract or add one Betti number.*



(a) *ST* **(b)** *ER*

(c) *WS* **(d)** *BA*

Figure 4.10: *The total number of simplices in each dimension for the directed flag complex on ST, ER, WS and BA. For ER, WS and BA the counts are the result of the mean number of simplices in each dimension over 100 computations.*



(a) *ST* **(b)** *ER*

(c) *WS* **(d)** *BA*

Figure 4.11: *Persistence diagrams over H_0 and H_1 for ST,ER,WS and BA given by inclusion of vertices and their resulting higher order simplices in the negative order of their degrees.*

5 Conclusion

Our goal with this thesis is to provide both an introduction to theory of persistent homology, as well as examples of applications to real-world data. We believe this is achieved.

We provide an exposition of persistent homology through the concept of a persistence module. We then state and finally prove the unique decomposition of persistence modules into a direct sum of free and torsional parts. Furthermore, the proof of this theorem yields a concrete algorithm for computing persistent homology of a given filtration through the computation of graded Smith normal form. By associating the decomposition with barcodes, and further on persistence diagrams, we illustrate how persistent homology can be visualized. Furthermore, we review the bottleneck distance and q-Wasserstein distance which allows us to compare different barcodes with each other.

On the application side, we present two case studies. These studies are not to be seen as stand-alone results in their respective domains, but rather as examples of how persistent homology can be applied to achieve fruitful insights into data. By using these non-traditional ways of exploring data we hope that we show there is some merit to considering persistent homology as a way of enhancing a traditional data analysis.

In the first case study we analyze 3D scans of the corneas of the bumblebee *Bombus terrestris*. This analysis shows how persistent homology can be applied to volumetric data and how it can be used to perform a clustering and similarity analysis. Furthermore, we are able to find that the persistent homology, specifically the barcode of H_2 compared across samples with the bottleneck distance, reinforces the already shown hypothesis in [21], namely that the shape of the eyes of *Bombus terrestris* differs between smaller and larger individuals. We also interpret this result as implying that it is the density of the cornea which is a distinguishing factor between differently-sized individuals.

In the second case study we analyze a synthetically generated network made to mimic the micro-circuitry of the striatum. By interpreting the network as

a directed graph, we show how persistent homology can be used to compare real-world data given as graphs to control models generated in multiple ways. We reinforce the already established result in [1], that the resulting directed flag complex on the brain network displays a much richer simplicial structure in terms of dimensions and number of simplices compared to control models. Furthermore, we establish that β_1 of the micro-circuitry is much lower than any of the control models. Finally, we see that the distribution of the persistent homology in H_1 of the micro-circuitry is spread across the entire spectrum of possible degrees, whereas control models only have holes in small intervals of degrees. These observations could act as points of differentiation when it comes to characterizing the striatum.

For some further directions in the first case study, one could extend the methodology to see if H_2 is always the distinguishing factor within and between different species of insects. However, this would likely require a larger amount of samples.

In the second case study a potential lane of investigation is whether other persistence diagrams of filtrations than the degree based filtration are as unique to the micro-circuitry compared to the other models. Some other filtrations that can be used with the exact same methodology are other measures of importance in a graph, such as the number of shortest path through an edge or the number of neighbors which are neighbors to each other. This is something we wanted to do, but the computational demands together with time restraints made it unfeasible.

Additionally, it could prove fruitful to see whether micro-circuitry in the actual biological striatum displays a similar signature to the synthetic model in terms of persistent homology. Should this be the case, then it strengthens the result of the case study as a signature for the networks of neurons in the striatum. If it is not the case, then persistent homology could perhaps be a parameter to take into account when further calibrating the synthetic model.

Bibliography

[1] M. W. Reimann, M. Nolte, M. Scolamiero, K. Turner, R. Perin, G. Chindemi, P. Dłotko, R. Levi, K. Hess, and H. Markram, "Cliques of neurons bound into cavities provide a missing link between structure and function," *Frontiers in Computational Neuroscience*, vol. 11, p. 48, 2017.

[2] M. Li, K. Duncan, C. Topp, and D. Chitwood, "Persistent homology and the branching topologies of plants," *American Journal of Botany*, vol. 104, 2017.

[3] C. Moon, S. A. Mitchell, J. E. Heath, and M. Andrew, "Statistical inference over persistent homology predicts fluid flow in porous media," *Water Resources Research*, vol. 55, no. 11, pp. 9592–9603, 2019.

[4] A. Zomorodian and G. Carlsson, "Computing persistent homology," *Discrete & Computational Geometry*, vol. 33, no. 2, pp. 249–274, 2005.

[5] H. Edelsbrunner, D. Letscher, and A. Zomorodian, "Topological persistence and simplification," *Discrete Computational Geometry*, vol. 28, no. 4, pp. 511–533, 2002.

[6] M. Vejdemo-Johansson, "Sketches of a platypus: Persistent homology and its algebraic foundations," in *Algebraic Topology: Applications and New Directions*, 2014, pp. 295–320.

[7] P. Skraba and M. Vejdemo-Johansson, "Persistence modules: Algebra and algorithms," *CoRR*, vol. abs/1302.2015, 2013. arXiv: `1302.2015`.

[8] A. Hatcher, *Algebraic topology*. Cambridge New York: Cambridge University Press, 2002.

[9] H. Edelsbrunner and J. L. Harer, *Computational topology: an introduction*. Providence, R.I: American Mathematical Society, 2010.

[10] C. Weibel, *An introduction to homological algebra*. Cambridge England New York: Cambridge University Press, 1994.

[11] T. Kaczynski, K. Mischaikow, and M. Mrozek, *Computational homology*. New York: Springer, 2004, vol. 157.

[12] R. Ghrist, "Barcodes: The persistent topology of data," *Bulletin (New Series) of the American Mathematical Society*, vol. 45, 2008.

[13] D. Cox, J. Little, and D. O'Shea, *Ideals, Varieties and Algorithms: An Introduction to Computational Algebraic Geometry and Commutative Algebra*. Springer, 1991.

[14] N. Jacobson, *Basic algebra*. Mineola, N.Y: Dover Publications, 2009.

[15] U. Bauer, *Ripser: Efficient computation of vietoris-rips persistence barcodes*, Preprint, 2021. eprint: `1908.02518v2`.

[16] P. Skraba and K. Turner, *Wasserstein stability for persistence diagrams*, 2021. arXiv: `2006.16824 [math.AT]`.

[17] ——, *Notes on an elementary proof for the stability of persistence diagrams*, 2021. arXiv: `2103.10723 [math.AT]`.

[18] O. Delgado-Friedrichs, V. Robins, and A. Sheppard, "Morse theory and persistent homology for topological analysis of 3d images of complex materials," in *IEEE International Conference on Image Processing (ICIP)*, 2014, pp. 4872–4876.

[19] A. Gutierrez, D. Monaghan, M. J. Jiménez, and N. E. O'Connor, "Persistent homology for 3D reconstruction evaluation," in *Computational Topology in Image Context*, M. Ferri, P. Frosini, C. Landi, A. Cerri, and B. Di Fabio, Eds., Berlin, Heidelberg: Springer Berlin Heidelberg, 2012, pp. 139–147.

[20] A. Gutierrez, M. J. Jimenez, D. Monaghan, and N. E. O'Connor, "Topological evaluation of volume reconstructions by voxel carving," *Computer Vision and Image Understanding*, vol. 121, pp. 27–35, 2014.

[21] G. Taylor, P. Tichit, M. Schmidt, A. Bodey, C. Rau, and E. Baird, "Bumblebee visual allometry results in locally improved resolution and globally improved sensitivity," *eLife*, vol. 8, 2019.

[22] J. J. J. Hjorth, A. Kozlov, I. Carannante, J. Frost Nylén, R. Lindroos, Y. Johansson, A. Tokarska, M. C. Dorst, S. M. Suryanarayana, G. Silberberg, J. Hellgren Kotaleski, and S. Grillner, "The microcircuits of striatum in silico," *Proceedings of the National Academy of Sciences*, 2020.

[23] P. Erdos and A. Renyi, "On the evolution of random graphs," *Publications of the Mathematical Institute of the Hungarian Academy of Sciences*, vol. 5, pp. 17–61, 1960.

[24] D. Watts and S. Strogatz, "Collective dynamics of 'small-world' networks," *Nature*, no. 393, pp. 440–442, 1998.

[25] A.-L. Barabasi and R. Albert, "Emergence of scaling in random networks," *Science (New York, N.Y.)*, vol. 286, pp. 509–12, 1999.

[26] O. Sporns, D. R. Chialvo, M. Kaiser, and C. C. Hilgetag, "Organization, development and function of complex brain networks," *Trends in Cognitive Sciences*, vol. 8, no. 9, pp. 418–425, 2004.

[27] D. Luetgehetmann, D. Govc, J. Smith, and R. Levi, "Computing persistent homology of directed flag complexes," *Algorithms*, vol. 13, no. 1, 2020.