Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

Application of Persistent Homology to the Study of Reticulate Evolution

Bachelor's Thesis

Naomi Rosenberg

5 July 2024

Advisor: Dr. Sara Kališnik Hintz
Department of Mathematics, ETH Zürich

Abstract

For a long time, evolution has been modeled by phylogenetic trees. Phylogenetic trees are genealogical trees from which the descendants of different species can be traced. They provide a linear representation in that they only record which random mutations occur over time, but they do not capture the merging of genetic material from different lineages. The fact that they cannot represent this naturally frequently occurring reticulate exchange of genomes is a decisive drawback. Since the 2000's, attempts have been made to address this problem using topology, in particular by applying persistent homology to genomic data in order to obtain a different, more suitable measure of evolutionary relationships. In this thesis, we develop theoretical foundations for this approach by formally introducing persistent homology and then defining temporal and topological novelty profiles – statistics quantifying how recombination contributes to genetic diversity. Based on the assumption that the goal of the topological approach to recombination is the estimation of novelty profiles, we prove a lower bound on them following [1]. In the proof, we resort to techniques from discrete Morse theory.

Acknowledgements

I would like to thank my advisor, Dr. Sara Kališnik Hintz, for introducing me to applied algebraic topology and for her invaluable advice during the writing of this thesis. Working under her supervision has been a highly instructive, inspiring and rewarding experience.

Contents

Contents **iii**

1 Introduction **1**

2 Persistent Homology **6**
2.1 Abstract and Geometric Simplicial Complexes 6
2.2 Persistent Homology 9
2.2.1 Persistence Vector Spaces 9
2.2.2 Decomposition Theorem 14
2.3 Visualizing Persistent Homology: Persistence Barcodes 23
2.3.1 Distances on Persistence Barcodes 23
2.3.2 Stability Results 25

3 Novelty Profiles **26**
3.1 Phylogenetic Graphs and Evolutionary Histories 26
3.2 Novelty Profiles 29
3.2.1 Temporal Novelty Profile 29
3.2.2 Topological Novelty Profile 31

4 Novelty Profiles on Galled Trees **37**
4.1 Galled Trees 37
4.2 Barcodes of Histories Indexed by Galled Trees 41
4.2.1 Metric Decomposition of an Evolutionary History 42
4.2.2 Vietoris-Rips Filtrations of Almost Linear Metric Spaces 46
4.2.3 Interference about Recombination from Barcodes 57

Bibliography **59**

Chapter 1

Introduction

In 1859, C. Darwin first introduced the idea of modelling evolution of phenotypic attributes using so-called *phylogenetic trees*, in his book "On the Origin of Species" [2]. Phylogenetic trees (as shown in Figure 1.1 a)) provide a visual representation of potential evolutionary relationships by tracing back from a small dataset of existing organisms. However, they have been considered outdated for quite some time since evolution does not only occur by random mutations over a number of generations *(vertical evolution)*, but also by the merging of genetic material between individuals of different lineages leading to so-called *reticulate events (horizontal evolution)*. C. Darwin's model fails at representing the latter. In particular, it fails at representing *recombination*, a process giving rise to genetic diversity by which the *genomes* (genetic information) of two parental organisms form a new genome. By uniting advantageous traits which have arisen in separate lineages and rescuing advantageous traits from otherwise disadvantageous backgrounds, recombination can hasten the pace at which genetic novelty arises. Directed acyclic graphs (as shown in Figure 1.1), which we call *phylogenetic graphs*, offer a way to visualize evolution including reticulate events. Phylogenetic graphs are defined in such a way that any *clone* (vertex with one incoming edge) inherits all mutations from its parent (and possibly some new mutations) and that for every mutation that occurs, there is a unique origin of the mutation in some organism. Finally, if both parents of a *recombinant* (vertex with two incoming edges) inherit a mutation, then the recombinant inherits it. Moreover, the definition stipulates that any mutation carried by a recombinant is inherited from a parent.

Figure 1.1: Two graphs modelling evolutionary relationships ((v, w) is an arc whenever v is a "direct ancestor", i.e. a "parent", of w). In a), we can see an example for vertical evolution as described by C. Darwin. In b), the black vertex corresponds to a recombinant which descends from two parents.

More than 150 years after C. Darwin's discoveries, in 2013, J. Chan et al. developed a framework which allows for an application of *persistent homology* to viral genomic datasets. The goal is to represent general evolutionary processes that may include reticulate events and extract abstract patterns in these processes by associating signatures, called barcodes, to genomic data sampled from an evolutionary history [3].

Persistent homology is an adaptation of homology to sequences of nested topological spaces called *filtrations*. On point clouds – finite metric spaces – we can get such filtrations, for example, by approximating the "shape" of the space we obtain by continuously "thickening" the points, as sketched in green in Figure 1.2 (when the points are in Euclidean space), depending on a *filtration parameter*. The space obtained from the union of the balls with the vertices as centres can be approximated by a simplicial complex. In the plane, a simplicial complex is the collection of vertices, edges and triangles glued along common edges. In higher dimensions, it may include higher dimensional convex hulls of finitely many points, called *simplices*, glued along boundaries. On a point cloud we build it by adding all vertices and adding a simplex whenever the diameter of its vertices is at most $2r$, as indicated in red in Figure 1.2. For every r, we call this construction *Vietoris-Rips complex* and we refer to the filtration as *Vietoris-Rips filtration*.

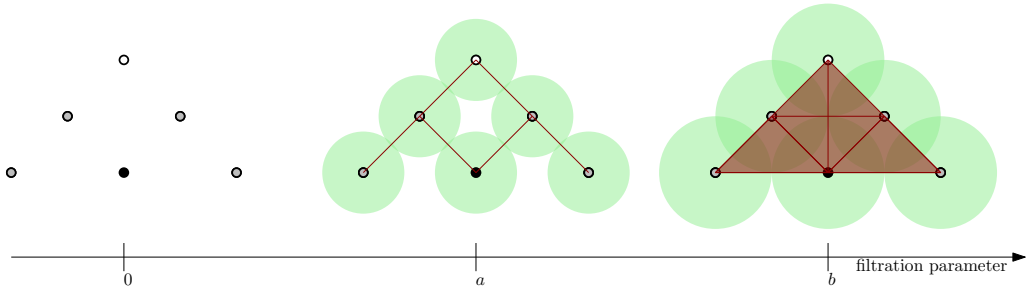

Figure 1.2: Approximation of the shape in Figure 1.1 b) embedded into $\mathbb{R}^2$. The green balls symbolize the "thickening" of the points and the collection of red simplices represents the approximation of the shape.

Given a filtration, persistent homology computes the number of connected components in dimension 0, the number of holes in dimension 1, and the number of higher dimensional voids in dimensions greater than 1. This information can be captured by *persistence barcodes*. The i^{th} persistence barcode contains an interval $[s, t)$ for every i-dimensional cavity (connected component if $i = 0$) which is *born* at time s and *dies* at time t. For the filtration shown in Figure 1.2, the 0^{th} and 1^{st} barcodes are shown in Figures 1.3 and 1.4.

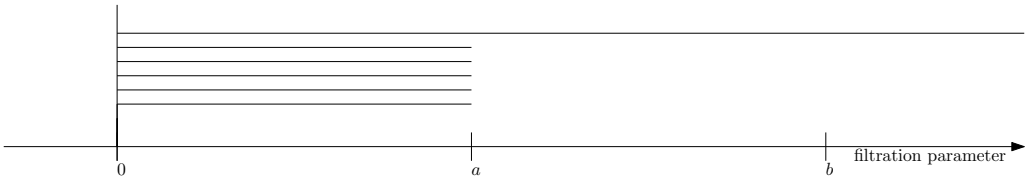

Figure 1.3: 0^{th} barcode corresponding to the red filtration in Figure 1.2, under the assumption that the first image corresponds to 0 on the x-axis, the middle one to a and the right one to b.

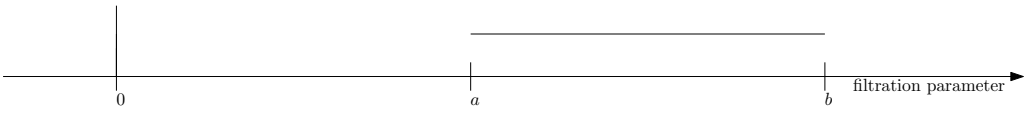

Figure 1.4: 1^{st} barcode corresponding to the red filtration in Figure 1.2, under the assumption that the first image corresponds to 0 on the x-axis, the middle one to a and the right one to b.

This idea of measuring the "shape" of a space can be applied to genomic data sets consisting of sets encoding the genomes of organisms in terms of the difference between each genome and some fixed (unspecified) reference genome.

As reference genome, we could for example have the sequence of bases $O_0 : ATGCCCAG$. As three other genomes we compare this sequence to, take

$$O_1 : \ ATGCCCAA, \ O_2 : \ TTACCCAG, \ \text{and} \ O_3 : \ TTGCCCAA,$$

where differences in the bases to the reference genome are marked in red. Then O_0 and O_1 differ at one position whereas O_0 and O_2 and O_0 and O_3 differ at two positions. We denote the differences as follows:

$$\begin{aligned} k: &\ \text{on position } 8 : G \mapsto A, \\ s: &\ \text{on position } 1 : A \mapsto T, \\ c: &\ \text{on position } 3 : G \mapsto A. \end{aligned}$$

We now define a metric space $\mathcal{E}$, called an *evolutionary history* indexed by a phylogenetic graph, to be a phylogenetic graph together with a set $\mathcal{E}_v$ for each vertex v. The distance between any two sets $\mathcal{E}_v$ and $\mathcal{E}_w$ in $\mathcal{E}$ is the cardinality of their symmetric difference.

If we now look again at the example from before, we can consider O_i as v_i, for $i = 0, 1, 2$ and O_3 as r. The evolutionary history in Figure 1.5 then provides us with a suitable model for the evolving population.

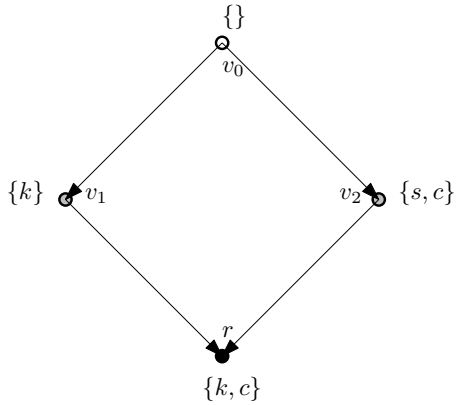

Figure 1.5: Evolutionary history indexed by a phylogenetic graph. Clones are drawn in grey and the unique recombinant is drawn in black.

For the purpose of measuring the influence of recombination on evolution, we define two types of *novelty profiles* on evolutionary histories. A novelty profile of an evolutionary history is simply a list of k monotonically decreasing numbers, where k is the number of recombination events in the history. The *temporal novelty profile* measures the *temporal novelty* of every recombinant r. The latter is defined to be the minimum of the symmetric distances between between $\mathcal{E}_r$ and $\mathcal{E}_v$ over all vertices v which correspond to organisms that appear earlier in the sense that there exists a path from v to r in the phylogenetic graph. Since there is only one recombinant the history in Figure 1.5, the temporal novelty profile corresponding to it is the list containing only the temporal novelty of r, which is (1). The *topological novelty profile* is defined to be the list of cardinalities of symmetric distances between $\mathcal{E}_u$ and $\mathcal{E}_v$, where u and v are the directed edges in the graph $T \setminus F^G$, constructed thusly: F^G is obtained by first removing all vertices from G pointing to recombinants, and T is the minimum spanning tree of the complete graph with the same vertex set as G, with edge weights determined by genetic distances, which contains F^G. The temporal novelty profile is constructed more intuitively, however, the topological novelty profile fulfils stronger stability results. The topological novelty profile of our example is (1) – just like the temporal novelty profile. On histories indexed by general phylogenetic graphs, the topological and temporal novelty profiles are not always the same. Nevertheless, the fact that they are the same here is not a coincidence, given that the underlying graph is a so-called *galled tree* on which they are always equal. A galled tree is a modification of a tree that may include some well-behaved cycles. This construction allows for modeling evolutionary relationships in a *low recombination regime*, in which only isolated recombination events occur.

In [1], M. Lesnick et al. suggest that the central inference problem in topological approaches to evolution is the estimation of novelty profiles. The findings presented in [3] suggest that the barcode of the Vietoris-Rips complex on an evolutionary history encodes information about both the number of recombination events and the contribution of recombination to genetic diversity, but the precise statistical nature of the relationship between barcodes

and recombination has not been made clear. In this thesis, we make progress towards understanding this relationship by applying persistent homology to the study of recombination. We focus on the findings published in [1] regarding reticulate evolution that can be modeled by galled trees. For the investigation of the latter, we introduce some notions from discrete Morse theory that allow us to reduce the problem of understanding the barcodes of galled trees to the problem of understanding the barcodes of so-called *almost metric spaces* – finite metric spaces P, with a distinguished point $p \in P$, such that $P \setminus \{p\}$ can be isometrically embedded into $\mathbb{R}$.

Finally, we prove our main result which states that for every evolutionary history indexed by a galled tree, the set of lengths of intervals in the first persistence barcode of the Vietoris-Rips complex provides a lower bound on the novelty profile. In particular, the number of intervals in the barcode is a lower bound on the number of recombination events in the history. Furthermore, all barcodes of the Vietoris-Rips complex for dimensions $i \geq 2$ are trivial.

Chapter 2

Persistent Homology

In order to analyze genetic relations between different organisms, we can view the different organisms as vertices forming a metric space where the distance between two organisms in this space is determined by how much the genetic material of the organisms differs. It has previously been hypothesized that applying the so-called persistent homology to an arbitrary sample P of an evolving population provides us with information about both the number of recombination events and the contribution of recombination to genetic diversity. However, the precise statistical nature of the relationship between persistent homology and recombination has not been made clear. In order to be able to make progress towards understanding this relationship, we introduce persistent homology in this chapter. We start by defining simplicial complexes before introducing filtrations, collections of nested topological spaces (in our application simplicial complexes), on point clouds. In homology, the inclusion map between two spaces in the filtration induces a homomorphism, so that in total, every filtration induces a sequence of homology groups with homomorphisms between them. We refer to this object as persistence vector space and we use it to compute persistent homology. Intuitively, persistent homology measures the shape of the space we obtain from the sample P interpreted as a subset of a metric space by continuously thickening the data points (so that we get balls) in terms of the number of voids in it. We show that every persistence vector space can be decomposed into a sum of easily understandable persistence vector spaces that we visually interpret by barcodes. Finally, we introduce distances on barcodes in order to state stability results that form a basis for some results in Chapter 3. We follow [4], [5], and [1]. We also consult [6] where it is marked.

2.1 Abstract and Geometric Simplicial Complexes

We consider persistent homology on simplicial complexes which we introduce in this section. We start by defining simplices before defining simplicial complexes which we obtain by, roughly speaking, gluing simplices together along their faces in compliance with some rules specified later.

Definition 2.1 *Let $P = \{x_0, x_1, \ldots, x_n\}$ denote a subset of an Euclidean space $\mathbb{R}^k$. We say that P is in **general position** if it is not contained in any affine hyperplane of $\mathbb{R}^k$ of dimension less than n.*

Definition 2.2 *Consider a set P of points in $\mathbb{R}^k$ in general position. We define the **simplex** spanned by P to be the convex hull $\sigma = \sigma(P)$ of P in $\mathbb{R}^k$. The points x_i are called **vertices** and the simplices $\sigma(T)$ spanned by non-empty subsets $T \subseteq P$ are called **faces** of σ. We say that a simplex is a **k-simplex** when it is spanned by $k+1$ vertices.*

Definition 2.3 *A **(finite geometric) simplicial complex** is a finite collection $\mathcal{X}$ of simplices in a Euclidean space so that the following conditions hold.*

1. *For any simplex σ of $\mathcal{X}$, all faces of σ are also contained in $\mathcal{X}$.*
2. *For any two simplices σ and τ of $\mathcal{X}$, the intersection $\sigma \cap \tau$ is a simplex, which is a face of both σ and τ.*

Let us have a look at some examples and counterexamples.

Example 2.4 *In Figure 2.1, Figure a) is a geometric simplicial complex. Figures b), c) and d) are not simplicial complexes. Indeed, in b), we can see a two-dimensional simplex, but not all of its edges are part of the structure. This violates the first condition. In c), there is an intersection of an edge and a vertex, and the intersection is not a face of the upper simplex and in d), we observe an intersection of edges which violates the second condition.*

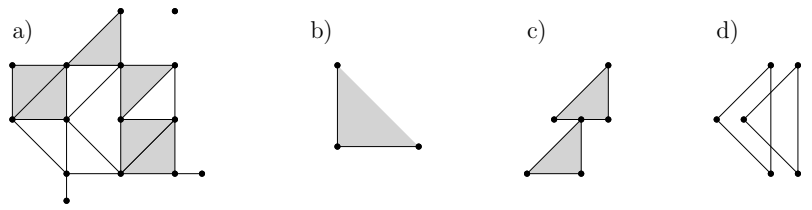

Figure 2.1: Figure a) represents a geometric simplicial complex, whereas b), c) and d) are not geometric simplicial complexes.

Keeping only the combinatorial information and omitting coordinates, we arrive at the following definition.

Definition 2.5 *An **abstract simplicial complex** is a pair $X = (V(X), \Sigma(X))$, where $V(X)$ is a finite set called the vertices of X, and $\Sigma(X)$ is a subset (called the simplices) of the collection of all non-empty subsets of $V(X)$, satisfying the conditions that if $\sigma \in \Sigma(X)$, and $\emptyset \neq \tau \subseteq \sigma$, then $\tau \in \Sigma(X)$. We say that an abstract simplicial complex is k-**dimensional** if the simplices of maximal dimension are k-dimensional.*

We note the following.

Observation 2.6 *Every geometric simplicial complex $\mathcal{X}$ determines an abstract simplicial complex whose vertex set is the set of all vertices of all simplices of $\mathcal{X}$, and where a subset*

of the vertices is in the collection of simplices if and only if the set is the set of vertices of some simplex of $\mathcal{X}$.

In the sequel, we show that geometric and abstract simplicial complexes can actually be translated into one another. For this purpose, we need some more theoretical background.

Definition 2.7 *A geometric simplicial complex $\mathcal{X}$ is a **geometric realization** of some abstract simplicial complex X if there exists an embedding $i : V(X) \to \mathbb{R}^d$ that takes every abstract k-simplex $\{v_0, \ldots, v_k\}$ in X to the geometric k-simplex that is the convex hull of $i(v_0), \ldots, i(v_k)$. We denote the geometric realization of X by $|X|$.*

We can now state a proposition making sure that there exists a translation as described above.

Proposition 2.8 *Every abstract simplicial complex has a geometric realization.*

Proof Consider an arbitrary abstract simplicial complex X and denote its vertices by $V(X) = \{v_1, \ldots, v_n\}$. We claim that the abstract simplicial complex has a geometric realization in $\mathbb{R}^n$. Indeed, consider the following map

$$i : V \to \mathbb{R}^n,\ v_j \mapsto e_j \text{ for } j \in \{1, \ldots, n\},$$

where we denote by e_j the j^{th} standard basis vector in $\mathbb{R}^n$. By construction, the map i is an embedding that takes every abstract k-simplex in X to the geometric k-simplex that is the convex hull of the image of its vertices under i. □

A visualization of a geometric realization of an abstract simplicial complex is shown in Figure 2.2.

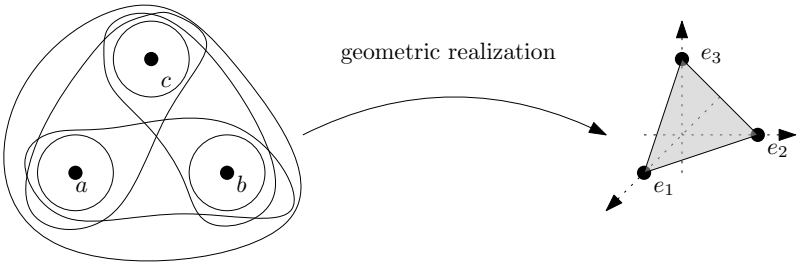

Figure 2.2: The abstract simplicial complex $\{\{\}, \{a\}, \{b\}, \{c\}, \{a, b\}, \{a, c\}, \{b, c\}, \{a, b, c\}\}$ on the left and a geometric realization on the right.

Remark 2.9 *Less computationally expensive realizations can also be found: for example, for any k-dimensional abstract simplicial complex, there exists a geometric realization in $\mathbb{R}^{2k+1}$. For a proof, we refer to [6, Lemma 5.4.2], which gives a construction from which the statement follows directly.*

We also want to have mappings between simplicial complexes.

Definition 2.10 *For two abstract simplicial complexes X and Y, a **map of abstract simplicial complexes** $f : X \to Y$ is a map of sets $f_V : V(X) \to V(Y)$ such that, for any simplex $\sigma \in \Sigma(X)$, the subset $f_V(\sigma)$ is in $\Sigma(Y)$.*

We would like to point out that the construction of the geometric realization is functorial in the sense that every map $f : X \to Y$ of abstract simplicial complexes induces a continuous map $|f| : |X| \to |Y|$, so that $|f \circ g| = |f| \circ |g|$ and $|id_X| = id_{|X|}$.

Since we can translate abstract and geometric complexes into one another, we only refer to simplicial complexes from now on.

2.2 Persistent Homology

We now use our knowledge of simplicial complexes to build such complexes on point clouds and thereby count the number of voids suggested by the point cloud.

2.2.1 Persistence Vector Spaces

What we get in the calculation of the persistent homology depends on how precisely we proceed in order to obtain a simplicial complex from a point cloud. In our observations, the data points in a point cloud P correspond to genomes and the nested family of spaces is constructed by "growing balls" around the points and then constructing simplicial complexes approximating the "shape" of our data set based on the space obtained from these balls. While "thickening" the points in the point cloud, and instead of considering single points, considering balls with centers the points from the point cloud and some radius r, we track the appearance of connected components, holes and higher dimensional voids as r increases. In order to examine the development of the "shape" over time, we define so-called filtrations.

Definition 2.11 *A **filtration** is a collection of topological spaces $\{\mathcal{F}_r\}_{r\in[0,+\infty)}$ such that $\mathcal{F}_r \subseteq \mathcal{F}_s$ whenever $r \leq s$. A **simplicial filtration** is a filtration where every $\mathcal{F}_r$ in the collection is a simplicial complex.*

Since, when analyzing data, we deal with point clouds, most of the time, the input data is not given in the form of a simplicial filtration. So we have to transform the data into one. In the process of transforming the data, we want the shape of the underlying data to be retained as much as possible. There exist several methods for constructing simplicial complexes on point clouds. We provide some intuition by introducing the so-called Čech complex.

Definition 2.12 (Čech complex) *Let (M, d) be a metric space and let P be a finite subset of M. Given a real $r > 0$, the **Čech complex** $\mathcal{C}(P)_r$ is defined to be the simplicial complex with vertex set P obtained by adding a simplex with vertices $a_1, \ldots, a_n$ to the complex whenever the balls with radius r and centers $a_1, \ldots, a_n$ intersect.*

It has been shown that the Čech complex $\mathcal{C}(P)_r$ at radius r has the same homotopy type as the union of balls grown around the data points with radius r (at least for some metric spaces including the Euclidean space). This follows directly from a result referred to as Nerve Theorem [7]. In this thesis, we work with a slightly different and computationally more efficient complex, the so-called Vietoris-Rips complex $\mathbb{VR}(P)_r$ at radius r on a point cloud P. Another decisive advantage over the Čech complex is that the points of the data set do not have to be embedded into the Euclidean space. The Vietoris-Rips complex only approximately measures the presence of voids in the union of balls around data points.

Definition 2.13 (Vietoris-Rips Complex) *Let (P, d) be a finite metric space. Given a real $r > 0$, the **Vietoris-Rips complex** is the simplicial complex $\mathbb{VR}(P)_r$, where a simplex σ is in $\mathbb{VR}(P)_r$ if and only if $d(p, q) \leq 2r$ for every pair (p, q) of vertices of σ.*

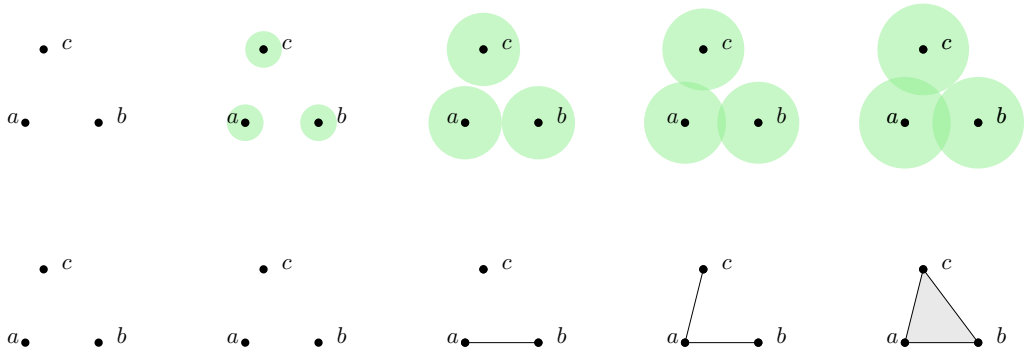

Figure 2.3: The Vietoris-Rips complex for different radii (times) of the point cloud consisting of the points a, b and c in the plane.

Example 2.14 *Consider a space P, consisting of three non-collinear points in $\mathbb{R}^2$, as shown in Figure 2.3. By "growing balls" of radius r around these points, we can read off which simplices are included in the Vietoris-Rips complex $\mathbb{VR}(P)_r$.*

Note that this example shows that the Vietoris-Rips complex and the Čech complex are generally different complexes. In fact, the Čech complex, which belongs to the rightmost image, is an unfilled triangle because the intersection of all three balls is empty.

Although the two complexes introduced might differ, we obtain the following inclusions:

$$\mathcal{C}(P)_r \subseteq \mathbb{VR}(P)_r \subseteq \mathcal{C}(P)_{2r}.$$

Consequently, the Vietoris-Rips complex gives a reasonable approximation for the number of holes in the "thickened" point cloud. So enough about why it makes sense to work with the Vietoris-Rips complex.

We would like to point out that visualizing Vietoris-Rips complexes does only work with the image of growing balls when the points are in Euclidean space with the standard metric. An example of a filtration where the image of growing balls cannot be used is shown in Figure 2.4.

Example 2.15 *Consider the set of points on the left of Figure 2.4 and the metric induced by some weighted graphs (edges and distances are not specified in the figure). Then the filtration given in the figure may represent a Vietoris Rips filtration, although the classic image of "growing balls" cannot be drawn in the usual way since the metric on the set of points is not assumed to be the one given by the shown embedding of points in the plane.*

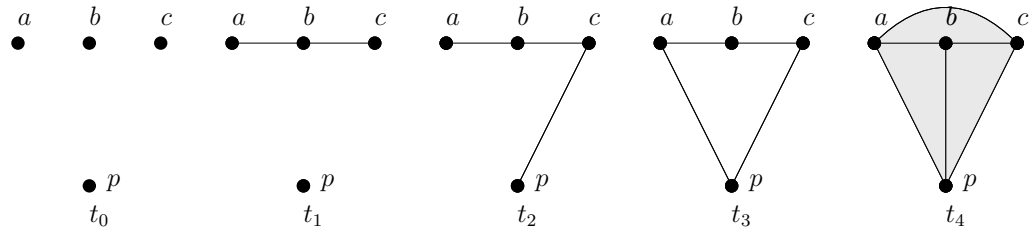

Figure 2.4: The Figure shows the Vietoris-Rips filtration for of the point cloud presented on the left with a metric which is not the one given by the embedding of points in the plane presented in the image.

In the following, we present the general theory, which we then apply to the Vietoris-Rips complex.

Definition 2.16 (Morphism of Filtrations) *A **morphism** $f : \mathcal{F} \to \mathcal{G}$ **of filtrations** is a collection of continuous maps $\{f_r : \mathcal{F}_r \to \mathcal{G}_r\}_{r\in[0,+\infty)}$ such that the following diagram commutes for all $r \leq s$:*

$$\begin{array}{ccc} \mathcal{F}_r & \hookrightarrow & \mathcal{F}_s \\ {\scriptstyle f_r}\downarrow & & \downarrow{\scriptstyle f_s} \\ \mathcal{G}_r & \hookrightarrow & \mathcal{G}_s \end{array}$$

*We say that f is an **objectwise homotopy equivalence** if each f_r is a homotopy equivalence.*

Intuitively, if two filtrations are connected by an objectwise homotopy equivalence, we can think of them as topologically equivalent.

If a continuous map g is a homotopy equivalence, then $g_* = \tilde{H}_i(g)$ is an isomorphism, where we denote by $\tilde{H}_i$ the i^{th} reduced homology. This implies the following lemma.

Lemma 2.17 *If a morphism of filtrations $f : \mathcal{F} \to \mathcal{G}$ is an objectwise homotopy equivalence, then for any $i \geq 0$, $f_* : \tilde{H}_i(\mathcal{F}) \to \tilde{H}_i(\mathcal{G})$ is an isomorphism.*

Definition 2.18 (Persistence Vector Space) *A **persistence vector space** M consists of a collection of vector spaces $\{M_r\}_{r\in[0,+\infty)}$, together with a collection of linear maps $\{m_{r,s} : M_r \to M_s\}_{r\leq s}$ such that*

1. *for all $r \leq s \leq t$ the following diagram commutes:*

$$\begin{array}{ccc} M_r & & \\ {\scriptstyle m_{r,s}}\downarrow & \searrow^{m_{r,t}} & \\ M_s & \xrightarrow{m_{s,t}} & M_t \end{array}$$

2. *$m_{r,r} = id_{M_r}$ for all r.*

From now on, we work over the field $K = \mathbb{F}_2$ whenever we compute examples. The nice thing when considering homology with $\mathbb{F}_2$-coefficients is that we can ignore orientations. The construction works similar to homology with $\mathbb{Z}$-coefficients, but instead of computing the homology by considering the sequence

$$\ldots \xrightarrow{\partial_{n+1}} C_n(X) \xrightarrow{\partial_n} \ldots \xrightarrow{\partial_2} C_1(X) \xrightarrow{\partial_1} C_0(X) \xrightarrow{\partial_0=0} 0$$

for a simplicial complex X with homology groups $H_i(X) = \ker(\partial_i)/\operatorname{im}(\partial_{i+1})$, we consider the sequence

$$\ldots \xrightarrow{\partial_{n+1}\otimes id} C_n(X) \otimes \mathbb{F}_2 \xrightarrow{\partial_n \otimes id} \ldots \xrightarrow{\partial_2 \otimes id} C_1(X) \otimes \mathbb{F}_2 \xrightarrow{\partial_1 \otimes id} C_0(X) \otimes \mathbb{F}_2 \xrightarrow{\partial_0 \otimes 0 = 0} 0$$

with homology groups $H_i(X; \mathbb{F}_2) := H_i(X \otimes \mathbb{F}_2) = \ker(\partial_i \otimes id)/\operatorname{im}(\partial_{i+1} \otimes id)$. One can show that this construction satisfies the Eilenberg-Steenrod Axioms and thus defines a homology theory. For more on this construction, reference is made to [8, Chapter 5.1] and [9, Chapter 2.2, "Homology with coefficients"]. Abusing notation, we write $H_i(X)$ when actually meaning $H_i(X; \mathbb{F}_2)$, and similarly in the adapted persistent setting.

The general theory works for any field.

Example 2.19 *We want to give an example of persistence vector spaces. Consider the filtration in Figure 2.4. We define persistence vector spaces with basis the i-chains in the filtration. For the purpose of simplifying the notation, we write simplices by concatenating letters designating vertices (e.g. by ab, we mean the 1-simplex $[a, b]$ with vertices a and b). By convention, we work with $\mathbb{F}_2$-coefficients. By $\langle \cdots \rangle$, we denote the span with $\mathbb{F}_2$-coefficients. For $i = 0$, the basis of the persistence vector space is given by the set of vertices $P = \{a, b, c, p\}$ for all $r \in [0, +\infty)$, so for all $r \in [0, +\infty)$*

$$(C_0(P))_r = \langle a, b, c, p \rangle.$$

For $(C_1(P))_r$, we obtain

$$(C_1(P))_r = \begin{cases} 0 & \text{if } r \in [t_0, t_1), \\ \langle ab, bc \rangle & \text{if } r \in [t_1, t_2), \\ \langle ab, bc, cp \rangle & \text{if } r \in [t_2, t_3), \\ \langle ab, bc, cp, ap \rangle & \text{if } r \in [t_3, t_4), \\ \langle ab, bc, cp, ap, ac, bp \rangle & \text{if } r \in [t_4, +\infty). \end{cases}$$

Moreover, assuming that acp is contained in the simplicial complex obtained at time t_4, it follows from the filtration that

$$(C_2(P))_r = \begin{cases} 0 & \text{if } r \in [t_0, t_4), \\ \langle abc, abp, bcp, acp \rangle & \text{if } r \in [t_4, +\infty). \end{cases}$$

Similar to the way we have already done it for filtrations, we also want to define morphisms between persistence vector spaces.

Definition 2.20 (Morphism of Persistence Vector Spaces) *A **morphism** or a **linear transformation** $f : M \to N$ **of persistence vector spaces** is a collection of linear maps $\{f_r : M_r \to N_r\}_{r \in [0,\infty)}$ such that for all $r \leq s$, the following diagram commutes:*

$$\begin{array}{ccc} M_r & \xrightarrow{m_{r,s}} & M_s \\ {\scriptstyle f_r}\downarrow & & \downarrow{\scriptstyle f_s} \\ N_r & \xrightarrow{n_{r,s}} & N_s \end{array} \quad .$$

*We say that f is an **isomorphism** if each of the maps f_r is an isomorphism.*

Example 2.21 *Consider again the persistence vector spaces from Example 2.33. For $i \in \{1, 2\}$, a linear transformation between $(C_i(P))_r$ and $(C_{i-1}(P))_r$ is given by the usual boundary map*

$$(\partial_i)_r : (C_i(P))_r \to (C_i(P))_r, [v_0, \ldots, v_i] \mapsto \sum_{j=0}^{i} (-1)^i [v_0, \ldots, \hat{v_j}, \ldots, v_i],$$

known from algebraic topology [9]. Note that in the undirected setting, i.e. when working $\mathbb{F}_2$-coefficients, this reduces to

$$(\partial_i)_r : (C_i(P))_r \to (C_i(P))_r, \{v_0, \ldots, v_i\} \mapsto \sum_{j=0}^{i} \{v_0, \ldots, \hat{v_j}, \ldots, v_i\}.$$

We observe that the notion of a quotient space can be extended to persistence vector spaces.

Definition 2.22 *If $N \subseteq M$ is a sub-persistence vector space, i.e., a choice of K-subspaces $N_r \subseteq M_r$, for all $r \in [0, \infty)$, so that $m_{r,s}(M_r) \subseteq M_s$ for all $r \leq s$, then we call the persistence vector space $M/N = \{M_r/N_r\}_r$ the **quotient space** of the persistence vector space M and one of its sub-persistence vector spaces N. In this case, the linear transformation from M_r/N_r to M_s/N_s is given by sending the equivalence class $[v]$ to the equivalence class $[m_{r,s}(v)]$ for every $v \in M_r$.*

Next up, we look at sets with filtrations.

Definition 2.23 *Let X be any set, equipped with a function $\rho : X \to [0, +\infty)$. Such a pair (X, ρ) is called an $\mathbb{R}_+$-**filtered set**.*

Definition 2.24 *Let X be an $\mathbb{R}_+$-filtered set. Denote by $V_K(X)$ the K-linear span of all elements in X. We define the **free persistence vector space** on the pair (X, ρ) to be the persistence vector space $V_K(X, \rho) = \{V_K(X, \rho)_r\}_r$ with $V_K(X, \rho)_r \subseteq V_K(X)$ equal to the K-linear span of the set $X[r] = \{x \in X | \rho(x) \leq r\} \subseteq X$.*

Example 2.25 *The persistence vector spaces $(C_i(P))_r$ for $i = 0, 1, 2$ from Example 2.19 are free because $(C_i(P))_r \subseteq (C_i(P))_{r'}$ whenever $r \leq r'$.*

Note that $X[r] \subseteq X[r']$ when $r \leq r'$, so there is an inclusion $V_K(X, \rho)_r \subseteq V_K(X, \rho)_{r'}$. We can deduce from this the following observation.

Observation 2.26 *A linear combination $\Sigma_x a_x x \in V_K(X)$ is in $V_K(X, \rho)_r$ if and only if $a_x = 0$ for all x satisfying $\rho(x) > r$.*

Definition 2.27 *For the persistence vector space from Observation 2.26, we write $V_K(X, \rho)$. Given any persistence vector space V, we say that it is **free** if there exists an isomorphism such that $V \cong V_K(X, \rho)$ for some pair (X, ρ). We say that it is **finite** if X can be chosen to be finite.*

2.2.2 Decomposition Theorem

We do not only consider persistence vector spaces, but also the direct sum of persistence vector spaces. It is defined taking the direct sum "pointwise" with respect to r.

Definition 2.28 *For linear maps $f : V_1 \to W_1$ and $g : V_2 \to W_2$, we define the **direct sum***

$$f \oplus g : V_1 \oplus V_2 \to W_1 \oplus W_2$$

by taking $f \oplus g(v, w) = (f(v), g(w))$. We then define the sum $M \oplus N$ to be the persistence vector space given by

$$(M \oplus N)_r = M_r \oplus N_r,\ (m \oplus n)_{r,s} = m_{r,s} \oplus n_{r,s}.$$

We define the direct sum of an arbitrary collection of persistence vector spaces in the same way.

As we see below, we can use this representation of persistence vector spaces to describe every persistence vector space up isomorphism. We can use this representation to compare different persistence vector spaces with each other. In this thesis, we prove the existence of a decomposition for so-called finitely presented persistence vector spaces, which are defined as follows.

Definition 2.29 *We say that a persistence vector space M is **finitely presented** if it is isomorphic to a persistence vector space of the form $N/\operatorname{im}(f)$ for some linear transformation $f : M \to N$ between finitely generated free vector spaces M and N.*

Recall that in the setting of linear algebra, the choice of basis for vector spaces V and W allows us to represent linear transformations between the two vector spaces by matrices. Our goal now is to transfer this representation to persistence vector spaces. To do this, we first define what we mean by an (X, Y)-matrix.

Definition 2.30 *For any pair (X, Y) of finite sets and a field K, an (X, Y)-**matrix** is an array $(a_{xy})_{x\in X, y\in Y}$ of elements $a_{xy} \in K$. We write $r(x)$ for the row corresponding to $x \in X$, and $c(y)$ for the column corresponding to $y \in Y$.*

For any finitely generated free persistence vector space $M = V_k(X, \rho)$, we observe that $V_k(X, \rho)_r = V_k(X)$ for r sufficiently large, since X is finite. Consequently, any linear transformation $f : V_K(Y, \sigma) \to V_K(X, \rho)$ of finitely generated persistence vector spaces gives a linear transformation $f_\infty : V_K(Y) \to V_K(X)$ between finite-dimensional vector spaces over K, and using the bases $\{\varphi_x\}_{x\in X}$ of $V_K(X)$ and $\{\varphi_y\}_{y\in Y}$ of $V_K(Y)$ determines (after imposing any orderings on X and Y) an (X, Y)-matrix $A(f) = (a_{xy})_{x,y}$ with entries in K.

Proposition 2.31 *The (X, Y)-matrix $A(f)$ has the property that $a_{xy} = 0$ whenever $\rho(x) > \sigma(y)$. Any (X, Y)-matrix A satisfying these conditions uniquely determines a linear transformation of persistence vector spaces*

$$f_A : V_K(Y, \sigma) \to V_K(X, \rho).$$

In particular, the correspondences $f \to A(f)$ and $A \to f_A$ are inverses to each other.

Proof Take any basis vector y, then it lies in $V_K(Y, \sigma)_{\sigma(y)}$. It holds that

$$f(\varphi_y) = \sum_{x\in X} a_{xy}\varphi_x.$$

From Observation 2.26, we know that $\Sigma_{x\in X} a_{xy}\varphi_x$ lies in $V_K(X, \rho)_{\sigma(y)}$ if and only if for $\rho(x) > \sigma(y)$, all coefficients a_{xy} are zero. So the (X, Y)-matrix $A(f)$ has the property that $a_{xy} = 0$ whenever $\rho(x) > \sigma(y)$. □

The matrices we work with are so-called (ρ, σ)-adapted (X, Y)-matrices. They do not only capture elements (simplices) of X and Y, but also the time at which they are added in the filtration.

Definition 2.32 *Given two $\mathbb{R}_+$-filtered finite sets (X, ρ) and (Y, σ), we say that the (X, Y)-matrix $A(f)$ from Proposition 2.31 satisfying $a_{xy} = 0$ whenever $\rho(x) > \sigma(y)$ is (ρ, σ)-**adapted**.*

Example 2.33 *Consider the filtration in Figure 2.4. The (ρ, σ)-adapted matrices $(\partial_1)_\infty$ and $(\partial_2)_\infty$ are the following:*

$$(\partial_1)_\infty = \begin{array}{c} \\ (a, t_0) \\ (b, t_0) \\ (c, t_0) \\ (p, t_0) \end{array} \begin{array}{c} \begin{array}{cccccc} (ab, t_1) & (bc, t_1) & (cp, t_2) & (ap, t_3) & (ac, t_4) & (bp, t_4) \end{array} \\ \begin{pmatrix} 1 & 0 & 0 & 1 & 1 & 0 \\ 1 & 1 & 0 & 0 & 0 & 1 \\ 0 & 1 & 1 & 0 & 1 & 0 \\ 0 & 0 & 1 & 1 & 0 & 1 \end{pmatrix} \end{array},$$

$$(\partial_2)_\infty = \begin{array}{c} \\ (ab,t_1) \\ (bc,t_1) \\ (cp,t_2) \\ (ap,t_3) \\ (ac,t_4) \\ (bp,t_4) \end{array} \begin{array}{c} \begin{array}{cccc} (abc,t_3) & (abp,t_4) & (bcp,t_4) & (acp,t_4) \end{array} \\ \begin{pmatrix} 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 0 & 0 & 1 & 1 \\ 0 & 1 & 0 & 1 \\ 1 & 0 & 0 & 1 \\ 0 & 1 & 1 & 0 \end{pmatrix} \end{array}.$$

Proposition 2.31 provides us with the following corollary which is crucial for the proof of the Decomposition Theorem 2.39 below.

Corollary 2.34 *Suppose that we are given (X,ρ) and (Y,σ), with ρ and σ both $[0,+\infty)$-valued functions on X and Y, respectively. Then any (ρ,σ)-adapted matrix $A=(a_{xy})_{x,y}$ determines a persistence vector space via the correspondence*

$$\theta : A \mapsto V_K(Y,\rho)/\operatorname{im}(f_A).$$

This correspondence between the matrix and the persistence vector space satisfies the following properties.

(i) For any A as described above, $\theta(A)$ is a finitely presented persistence vector space. Vice versa, any finitely presented persistence vector space is isomorphic to one of the form $\theta(A)$ for such a matrix A.

(ii) Let (X,ρ) be an $\mathbb{R}_+$-filtered set. Then, under the matrix/linear transformation correspondence, the automorphisms of $V_K(X,\rho)$ are identified with the group of all invertible (ρ,ρ)-adapted (X,X)-matrices.

We also emphasise the following proposition.

Proposition 2.35 *Let (X,ρ) and (Y,σ) be $\mathbb{R}_+$-filtered sets, and let A be a (ρ,σ)-adapted (X,Y)-matrix. Let B and C be (ρ,ρ)-, respectively (σ,σ)-adapted (X,X)-, respectively (Y,Y)-matrices. Then BAC is also (ρ,σ)-adapted, and the persistence vector space $\theta(A)$ is isomorphic to $\theta(BAC)$.*

With this knowledge, we can understand the Decomposition Theorem. We begin by defining a persistence vector space $P(a,b)$ for every pair (a,b) where $a\in\mathbb{R}_+, b\in\mathbb{R}_+\cup\{+\infty\}$, and $a<b$, with the obvious interpretation when $b=+\infty$.

Definition 2.36 *We define $P(a,b)$ as follows:*

$$M_r = P(a,b)_r = \begin{cases} K & \text{if } r\in[a,b), \\ \{0\} & \text{if } r\notin[a,b), \end{cases}$$

with linear maps

$$m_{r,r'} = \begin{cases} \mathrm{id}_K & \text{if } r,r'\in[a,b), \\ 0 & \text{otherwise.} \end{cases}$$

Example 2.37 *The persistence vector space $P(a,b)$ is finitely presented. Indeed, $P(a,b)_r$ is isomorphic to $\ker(\partial_n)_r / \operatorname{im}(\partial_{n+1})_r$ for every $r > 0$ and $n \in \mathbb{N}$.*

Remark 2.38 *In the case where b is finite, let (X,ρ) and (Y,σ) denote the $\mathbb{R}_+$-filtered sets (X,ρ) and (Y,σ), with the underlying sets consisting of single elements x and y, and with $\rho(x) = a$ and $\sigma(y) = b$. Then the (1×1) (X,Y)-matrix (1) is (ρ,σ)-adapted since $a \leq b$, and $P(a,b)$ is isomorphic to $\theta((1))$. When $b = +\infty$, $P(a,b)$ is isomorphic to the persistence vector space $V_K(X,\rho)$, and can therefore be written as $\theta(0)$, where 0 denotes the zero linear transformation from the persistence vector space 0.*

Theorem 2.39 (Decomposition Theorem) *Every finitely presented persistence vector space M over a field K is isomorphic to a finite direct sum of the form*

$$P(a_1,b_1) \oplus P(a_2,b_2) \oplus \ldots \oplus P(a_n,b_n)$$

for some choices $a_i \in [0,+\infty), b_i \in [0,+\infty]$, and $a_i < b_i$ for all i. This decomposition is unique up to permutation of summands.

There is a more general version of this statement for pointwise finite dimensional persistence vector spaces (persistence vector spaces M with $\dim(M_r) < \infty$ for all r), which was proven in [10] in 2014. Since we work with finite-dimensional vector spaces in the remainder of this work, the above version (Decomposition Theorem 2.39) is sufficient for us.

Proof
Existence: We start by proving the existence of the intended decomposition, but before going into the proof, we sketch how we proceed.

Idea of Proof: The idea of the proof is to note that any finitely presented persistence vector space is isomorphic to one of the form $\theta(A')$ for a (ρ,σ)-adapted (X,Y)-matrix A'. We then start by proving the theorem in the case where A' has at most one non-zero element and this is equal to one. Subsequently, for a general (ρ,σ)-adapted (X,Y)-matrix A, we construct matrices B and C like in Proposition 2.35 so that BAC satisfies the same property as A' and we conclude using Proposition 2.35. By Corollary 2.34 (i), any finitely presented persistence vector space is isomorphic to one of the form $\theta(A)$ for some (ρ,σ)-adapted (X,Y)-matrix A, i.e. an (X,Y)-matrix with $a_{xy} = 0$ whenever $\rho(x) > \sigma(y)$.

We start by taking a (ρ,σ)-adapted (X,Y)-matrix A' such that every row and column has at most one non-zero element and this element equals one.
Let $\{(x_1,y_1),(x_2,y_2),\ldots,(x_n,y_n)\}$ be all the pairs (x_i,y_i) so that $a_{x_iy_i} = 1$. Then

$$\begin{aligned}
\theta(A') &\cong V_K(X,\rho)/\operatorname{im}(f_{A'}) && (2.1)\\
&\cong \bigoplus_{x\in X} P(\rho(x),+\infty)/\operatorname{im}(f_{A'}) && (2.2)\\
&\cong \bigoplus_{i=1}^{n} P(\rho(x_i),\sigma(y_i)) \oplus \bigoplus_{x\in X\setminus\{x_1,\ldots,x_n\}} P(\rho(x),+\infty). && (2.3)
\end{aligned}$$

In 2.1 and 2.2, we just plugged in definitions. Note that 2.3 holds because the spaces are finite-dimensional. Thus, we have already shown the theorem for our special case.
In order to prove the general case, it is now sufficient to construct (X, X)-, respectively (Y, Y)-matrices B and C which are (ρ, ρ)-, respectively (σ, σ)-adapted so that BAC has the property that every row and every column has at most one non-zero element and that is the one-element.
Using row and column operations, we ultimately want to obtain our matrices B, C and BAC from Proposition 2.35. It is important that the matrices are adapted. We therefore adapt the row and column operations and allow the following operations in order to preserve the quotient space up to isomorphism.

- All possible multiplications of a row or a column by a non-zero element of K.
- All possible additions of a multiple of $r(x)$ to $r(x')$ when $\rho(x) \geq \rho(x')$.
- All possible additions of a multiple of $c(y)$ to $c(y')$ when $\sigma(y) \leq \sigma(y')$.

These adapted row operations make sense which we see by noting that left multiplication by the elementary matrix $E_{ij}(r)$ for an element r from a field (with $e_{ij} = r$, $e_{tt} = 1$, and $e_{uv} = 0$ for $u \neq v, u \neq i, v \neq j$) has the effect of adding r times the j^{th} row to the i^{th} row and right multiplication by this matrix has the effect of adding r times the i^{th} column to the j^{th} column.

We claim that by performing (ρ, σ)-adapted row and column operations, we can arrive at a matrix with at most one non-zero entry in each row and column and if such an entry exists, it is equal to the one-element in K.
Indeed, consider the algorithm described below.

1. Find y which minimizes $\sigma(y)$ over all y with $c(y) \neq 0$.
2. Find x which maximizes $\rho(x)$ over the set of all x for which $a_{xy} \neq 0$.
3. Note that
   - because of the way how x was chosen, we are free to add multiples of $r(x)$ to all other rows, and
   - because of the way how y was chosen, we can add multiples of $c(y)$ to all other columns without affecting $c(y)$.

   We now add multiples of $r(x)$ to the other rows so as to "zero out" $c(y)$, except in the xy-entry and we add multiples of $c(y)$ to the other columns so as to "zero out" $r(x)$ except in the xy-slot.
4. Make the xy-entry in the transformed matrix equal to one by multiplying $r(x)$ with $\frac{1}{a_{xy}}$. (This step is redundant when working over $\mathbb{F}_2$.)
5. Delete $r(x)$ and $c(y)$ in order to obtain a $(X\backslash\{x\}, Y\backslash\{y\})$-matrix which is $(\rho|_{X\backslash\{x\}}, \sigma|_{Y\backslash\{y\}})$-adapted.

6. Apply this process inductively to the matrices obtained in the end of step 5. Each of the row and column-operations required can be interpreted as row and column operations on the original matrix with no effect on $r(x)$ and $c(y)$. Consequently, by iterating this procedure, we eventually arrive at a matrix with only zero-entries. The transformed matrix has at most one non-zero element in each row and column and if there is such an element, it is one, by construction.

Finally, we apply Proposition 2.35, which yields that $\theta(A) \cong \theta(BAC)$ and therefore concludes the proof of the existence of a decomposition.

Uniqueness: Now we come to the proof of the uniqueness of the decomposition, i.e. we show that any two decompositions of the same persistence vector space are the same up to interchanging summands. Suppose that V is a finitely presented persistence vector space over K, and that we have two decompositions

$$M \cong \bigoplus_{i \in I} P(a_i, b_i) \text{ and } M \cong \bigoplus_{j \in J} P(c_j, d_j),$$

where I and J are finite sets.
Let $a_{\min}$ and $c_{\min}$ denote the smallest value of a_i and c_j, respectively and note that $a_{\min}$ is precisely $\min\{r | M_r \neq 0\}$. Thus, $a_{\min} = c_{\min}$ as otherwise, they would not both characterize V.
Now define $b_{\min}$ to be $\min\{b_i | a_i = a_{\min}\}$, and $d_{\min}$ to be $\min\{d_i | c_i = c_{\min}\}$. Similar to $a_{\min}$, $b_{\min}$ can be defined intrinsically, namely by $b_{\min} = \min\{r' | \ker m_{r,r'} \neq 0\}$. So $b_{\min} = d_{\min}$. This implies that $P(a_{\min}, b_{\min})$ appears in both decompositions.
The next step is to show that this summand occurs equally often in both decompositions. For each decomposition, the sums of occurences of $P(a_{\min}, b_{\min})$ are sub-persistence vector spaces of M and can in fact intrinsically be characterized as the sub-persistence vector space N, where $N = \ker(m_{r,b_{\min}}|_{\mathrm{im}(m|_{a_{\min},r})})$, because this sub-persistence vector space consists precisely of the elements born at $a_{\min}$ which are dead by the time $b_{\min}$. Consequently, both decompositions have the same number of $P(a_{\min}, b_{\min})$-summands.
Now define

$$I' = \text{subset of } I \text{ obtained by removing all } i \text{ such that } a_i = a_{\min}, \text{ and}$$
$$J' = \text{subset of } J \text{ obtained by removing all } j \text{ such that } b_j = b_{\min},$$

and form the quotient M/N. Then

- $M/N \cong \bigoplus_{i \in I'} P(a_i, b_i)$, and
- $M/N \cong \bigoplus_{j \in J'} P(c_j, d_j)$.

By induction over the number of summands in the decompositions, we obtain the uniqueness result. □

Remark 2.40 *From the algorithm described in the proof of Theorem 2.39 (existence), we can deduce an algorithm to compute persistent homology. Note that since the pairs of*

matrices (A, B) we consider are two consecutive boundary matrices and consequently, they satisfy $A \cdot B = 0$, the (ρ, σ)-adapted row and column operations translate to the following admissible operations on (A, B).

- *An arbitrary adapted row operation on A.*
- *An arbitrary adapted column operation on B.*
- *Perform an adapted column operation on A and an adapted row operation on B simultaneously, with the operations related as follows. If the adapted column operation in A is a multiplication of the i^{th} column by a non-zero constant a, then the adapated row operation on B is the transposition of the i^{th} row by a^{-1}. If the adapted column operation on A is the transposition of two columns, then the adapted row operation on B is the transposition of the corresponding rows of B. Finally, if the adapted column operation on A is r times the i^{th} column to the j^{th} column, then the adapted row operation on B is the subtraction of r times the j^{th} row from the i^{th} row.*

We want to point out the following property of pairs of matrices A, B as described above.

Corollary 2.41 *Given a pair (A, B), with $A \cdot B = 0$, we can perform operations of the type described above to obtain a pair*

$$(A', B') = \left(\begin{pmatrix} I_n & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}, \begin{pmatrix} 0 & 0 & 0 \\ 0 & I_m & 0 \\ 0 & 0 & 0 \end{pmatrix} \right).$$

The tuple (A', B') is uniquely determined by ther pair (A, B).

Proof Applying adapted row and column operations as in the proof of the Decomposition Theorem 2.39, we obtain a pair

$$(A', B'') = \left(\begin{pmatrix} I_n & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}, \begin{pmatrix} B''_{11} & B''_{12} & B''_{13} \\ B''_{21} & B''_{22} & B''_{23} \\ B''_{31} & B''_{32} & B''_{33} \end{pmatrix} \right),$$

where we denote by B'' we the matrix obtained by performing the adapted row operations on B corresponding to the column operations performed on A. We force the submatrix $\begin{pmatrix} B''_{11} & B''_{12} & B''_{13} \end{pmatrix}$ to have n rows and we denote the number of rows of $\begin{pmatrix} B''_{21} & B''_{22} & B''_{23} \end{pmatrix}$ by r and the number of rows of $\begin{pmatrix} B''_{31} & B''_{32} & B''_{33} \end{pmatrix}$ by s. Since by assumption, $A \cdot B = 0$, the submatrix $\begin{pmatrix} B''_{11} & B''_{12} & B''_{13} \end{pmatrix}$ is the zero-matrix. We can now perform only adapted row operations involving the last $r + s$ rows. Each adapted row operation on B'' has an adapted column operation on A' affecting only the rightmost $r + s$ columns and thus having no effect on A'. So in this setting, performing adapted row and column operations on B'' is equivalent to performing arbitrary row and column operations on the submatrix of B'' consisting of the last $r + s$ rows of B''. By performing suitable operations, we obtain B' of the required form.

Uniqueness follows from the fact that n and m are the ranks of A and B, respectively. □

Remark 2.42 *By keeping track of row and column operations, we can find the generators of persistent homology groups.*

Example 2.43 *Consider the filtration in Figure 2.4. In this example, we compute the* 1*-dimensional persistent homology of the filtration. We have already determined the crucial persistence vector spaces in Example 2.19 and we have determined the linear transformations in the form of boundary matrices in Example 2.33. The next step is to apply adapted row and column operations to the pair of the first and second boundary matrix to transform the matrices into a pair as in Corollary 2.41, using the algorithm described in the existence part of the proof of the Decomposition Theorem 2.39. We start by considering columns* y *of* ∂_1 *by ascending* $\sigma(y)$ *and we sort the rows of* ∂_2 *accordingly, as presented in Example 2.33. In order to keep track of the generators, it is sufficient to only keep track of the headlines of the matrices. Furthermore, abusing notation, we sometimes write* $r(i)$ *and* $c(i)$ *for the* i^{th} *row and column of the matrix currently looked at, respectively, in this computation.*

$$\left(\begin{array}{c} \begin{array}{cccccc} (ab,t_1) & (bc,t_1) & (cp,t_2) & (ap,t_3) & (ac,t_4) & (bp,t_4) \end{array} \\ \left(\begin{array}{cccccc} 1 & 0 & 0 & 1 & 1 & 0 \\ 1 & 1 & 0 & 0 & 0 & 1 \\ 0 & 1 & 1 & 0 & 1 & 0 \\ 0 & 0 & 1 & 1 & 0 & 1 \end{array}\right) \end{array}, \begin{array}{c} \begin{array}{cccc} (abc,t_4) & (abp,t_4) & (bcp,t_4) & (acp,t_4) \end{array} \\ \left(\begin{array}{cccc} 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 0 & 0 & 1 & 1 \\ 0 & 1 & 0 & 1 \\ 1 & 0 & 0 & 1 \\ 0 & 1 & 1 & 0 \end{array}\right) \end{array}\right)$$

We want there to be a 1 *in the upper left entry of the matrix. Since this is already the case, we do not have to exchange any rows and can proceed directly by "zeroing out" all entries besides the first in the first column using elementary row operations. Note that since all time labels of the rows are zero, we can add them as we wish. We use the same notation as in the proof of Theorem 2.39. Adding* $r(1)$ *to* $r(2)$ *in* ∂_1 *yields*

$$\left(\begin{array}{c} \begin{array}{cccccc} (ab,t_1) & (bc,t_1) & (cp,t_2) & (ap,t_3) & (ac,t_4) & (bp,t_4) \end{array} \\ \left(\begin{array}{cccccc} 1 & 0 & 0 & 1 & 1 & 0 \\ 0 & 1 & 0 & 1 & 1 & 1 \\ 0 & 1 & 1 & 0 & 1 & 0 \\ 0 & 0 & 1 & 1 & 0 & 1 \end{array}\right) \end{array}, \begin{array}{c} \begin{array}{cccc} (abc,t_4) & (abp,t_4) & (bcp,t_4) & (acp,t_4) \end{array} \\ \left(\begin{array}{cccc} 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 0 & 0 & 1 & 1 \\ 0 & 1 & 0 & 1 \\ 1 & 0 & 0 & 1 \\ 0 & 1 & 1 & 0 \end{array}\right) \end{array}\right).$$

The next step is eliminating all ones besides the first in the first row of ∂_1*. Adding* $c(1)$ *to* $c(4)$ *and* $c(5)$ *and performing the corresponding row operations in* ∂_2*, i. e. subtracting* $r(4)$ *and* $r(5)$ *from* $r(1)$*, yields*

$$\left(\begin{array}{c} \begin{array}{cccccc} (ab,t_1) & (bc,t_1) & (cp,t_2) & (ap+ab,t_3) & (ac+ab,t_4) & (bp,t_4) \end{array} \\ \left(\begin{array}{cccccc} 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 1 & 1 & 1 \\ 0 & 1 & 1 & 0 & 1 & 0 \\ 0 & 0 & 1 & 1 & 0 & 1 \end{array}\right) \end{array}, \begin{array}{c} \begin{array}{cccc} (abc,t_4) & (abp,t_4) & (bcp,t_4) & (acp,t_4) \end{array} \\ \left(\begin{array}{cccc} 0 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 0 & 0 & 1 & 1 \\ 0 & 1 & 0 & 1 \\ 1 & 0 & 0 & 1 \\ 0 & 1 & 1 & 0 \end{array}\right) \end{array}\right).$$

Now the first row and the first column of the matrix obtained from ∂_1 *are in the correct form and the first row of the matrix obtained from* ∂_2 *is zero. Thus we can continue by*

applying the analogous steps to those described above to the lower 3×5 *submatrix of the matrix obtained from* ∂_1 *and so on. Afterwards,* ∂_1 *is in the form from Corollary 2.41 and as described in the proof of Corollary 2.41, we apply operations on* ∂_2 *to obtain the correct form for this matrix, as well. In the end, we get the following pair of matrices*

$$\partial_1' = \begin{array}{c} \begin{array}{cccccc} (ab,t_1) & (bc,t_1) & (cp,t_2) & (ap+ab+bc+cp,t_3) & (ac+ab+bc,t_4) & (bp+bc+cp,t_4) \end{array} \\ \begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix} \end{array},$$

$$\partial_2' = \begin{array}{c} \\ (ab,t_1) \\ (bc,t_1) \\ (cp,t_2) \\ (ap+ab+bc+cp,t_3) \\ (ac+ab+bc,t_4) \\ (bp+bc+cp,t_4) \end{array} \begin{array}{c} \begin{array}{cccc} (acp+abp+bcp+abc,t_4) & (abp+bcp,t_4) & (abc,t_4) & (bcp,t_4) \end{array} \\ \begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} \end{array}.$$

Observe that $ap+ab+bc+cp = \partial(abp+bcp)$, $ac+ab+bc = \partial(abc)$, *and* $bp+bc+cp = \partial(bcp)$. *From the matrix obtained from* ∂_1, *we can now read off that the persistence vector space* $\ker(\partial_1)$ *is isomorphic to* (X,ρ), *where* $X = \{\partial(abp+bcp), \partial(abc), \partial(bcp)\}$ *and*

$$\rho(x) = \begin{cases} t_3 & \text{if } x = \partial(abp+bcp), \\ t_4 & \text{otherwise.} \end{cases}$$

In order to compute the 1*-dimensional persistent homology group, we have to consider the linear map* $(\partial_2)_r$, *mapping from* $(C_2(P))_r$ *to* $\ker(\partial_1)_r$, *for* $r \in [0,+\infty)$. *Denote by* Y *the set* $Y = \{abp+bcp, abc, bcp\}$ *with*

$$\sigma(y) = t_4 \text{ for all } y \in Y.$$

∂_2 *can now be represented by the* (X,Y)*-matrix*

$$\begin{array}{c} \\ (\partial(abp+bcp),t_3) \\ (\partial(abc),t_4) \\ (\partial(bcp),t_4) \end{array} \begin{array}{c} \begin{array}{ccc} (abp+bcp,t_4) & (abc,t_4) & (bcp,t_4) \end{array} \\ \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \end{array}.$$

It follows directely from Proposition 2.39 that the 1*-dimensional persistent homology group of the filtration given in Figure 2.4 is isomorphic to*

$$P(t_3,t_4) \oplus P(t_4,t_4) \oplus P(t_4,t_4) \cong P(t_3,t_4).$$

The persistent 1*-dimensional homology group is generated by* $\partial(abp+bcp)$ *and persists over the interval* $[t_3,t_4)$.

2.3 Visualizing Persistent Homology: Persistence Barcodes

The isomorphism classes of finitely presented persistence vector spaces are in one-to-one correspondence with finite subsets (with multiplicity) of the set

$$\{(a,b)|a \in [0,+\infty), b \in [0,+\infty], a < b\}.$$

Notation 2.44 We denote the collection of intervals $\{(a_i,b_i)\}_i$ we obtain as a result of the Decompositon Theorem 2.39 by $\mathcal{B}_M$ and call $\mathcal{B}_M$ the **barcode** of M. For $\mathcal{F}$ a filtration, we write $\mathcal{B}_{\tilde{H}_i(\mathcal{F})}$ simply as $\mathcal{B}_i(\mathcal{F})$, where we denote by $\tilde{H}_i$ the i^{th} reduced homology. Similarly, for P a finite metric space, we write $\mathcal{B}_i(\mathbb{VR}(P))$ simply as $\mathcal{B}_i(P)$.

Remark 2.45 *As it is defined using reduced homology, $\mathcal{B}_0(\mathcal{F})$ differs from the 0^{th} barcode constructed using unreduced homology by the removal of an infinite length interval.*

We visualize persistence barcodes as families of intervals on the non-negative real lines.

Example 2.46 *Recall again the filtration given in Figure 2.4. In Example 2.43, we have computed its 1-dimensional persistent homology. When we capture it in a persistence barcode, we obtain Figure 2.5.*

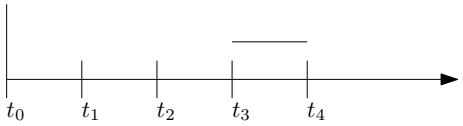

Figure 2.5: Barcode corresponding to the 1-dimensional persistent homology of the filtration in Figure 2.4.

2.3.1 Distances on Persistence Barcodes

In order to assess the similarity of point clouds, we work with metrics on the space of persistence barcodes. We pose the question to which degree a barcode changes when we have small (in suitable sense) modifications in the data. In order to even formulate an answer to such a question, we have to define what is meant by small changes in the barcode. For this purpose, we define the bottleneck distance between barcodes.

Definition 2.47 *For any pair of intervals $I = [x_1, y_1]$ and $J = [x_2, y_2]$, we define $\Delta(I,J)$ to be the l^∞-distance between the two, regarded as ordered pairs in $\mathbb{R}^2$, in other words, $\max(|x_2 - x_1|, |y_2 - y_1|)$. For a given interval $I = [x,y]$, we also define $\lambda(I)$ to be $\frac{y-x}{2}$. $\lambda(I)$ is the l^∞-distance between the closest interval of the form $[z,z]$ to I. Given two families $\mathfrak{I} = \{I_\alpha\}_{\alpha \in A}$ and $\mathfrak{J} = \{J_\beta\}_{\beta \in B}$ of intervals, for finite sets A and B, and any bijection θ from a subset $A' \subseteq A$ to $B' \subseteq B$, we define the **penalty** of θ, $P(\theta)$, to be*

$$P(\theta) = \max\left(\max_{a \in A'}(\Delta(I_a, J_{\theta(a)}), \max_{a \in A \setminus A'}(\lambda(I_a)), \max_{b \in B \setminus B'}(\lambda(J_b))\right).$$

We then define the ***bottleneck distance***

$$d_B(\mathfrak{I}, \mathfrak{J}) := \min_{\theta} P(\theta),$$

where the minimum is taken over all possible bijections from subsets of A to subsets of B.

Before giving an example, we want to make two remarks about the sets A' and B' in Definition 2.47.

Remark 2.48

(i) Note that since A and B are finite, A' and B' must be finite as well. Otherwise, there would not exist a bijection between these two sets.

(i) Consider the special case where $A' = B' = \emptyset$, then $\max_{a \in A'}(\Delta(I_a, J_{\theta(a)}) = -\infty$ for the bijection $\theta : \emptyset \to \emptyset$ since there are no intervals contained in the empty set. Moreover, the values $\max_{a \in A \setminus A'}(\lambda(I_a))$ and $\max_{b \in B \setminus B'}(\lambda(I_b))$ also appear in other choices of A' and B'. Consequently, when, in the process of computing the bottleneck distance, going through bijections between subsets A' and B' of A and B, we can leave out the case $A' = B' = \emptyset$.

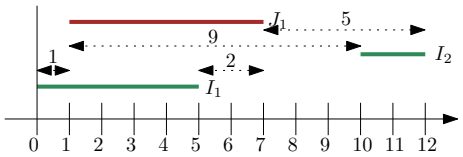

Figure 2.6: A visualization of two persistence barcodes $\mathcal{B}^1 = \{I_1, I_2\}$ (in green) and $\mathcal{B}^2 = \{J_2\}$ (in red).

Example 2.49 *Consider the persistence barcodes $\mathfrak{I} = \mathcal{B}^1 = \{I_1, I_2\} = \{I_\alpha\}_{\alpha \in A}$ and $\mathfrak{J} = \mathcal{B}^2 = \{J_1\} = \{J_\beta\}_{\beta \in B}$ as presented in Figure 2.6 and let us compute the bottleneck distance between the two barcodes. Note that $A = \{1, 2\}$ and $B = \{1\}$, so the only non-trivial bijections between a subset of $\mathcal{B}^1$ and a subset of $\mathcal{B}^2$ are the following: $\theta_1 : I_1 \to J_1$, $\theta_2 : J_1 \to I_1$, $\theta_3 : I_2 \to J_1$, and $\theta_4 : J_1 \to I_2$. We start by computing the penalties. Note that for symmetry reasons, it is sufficient to consider θ_1 and θ_3. We begin with θ_1. We can read from the picture that $\max_{a \in \{1\}}(\Delta(I_a, J_{\theta(a)}) = \Delta(I_1, J_1) = 2$. Moreover, if $A' = I_1$, then $A \setminus A' = I_2$, so $\max_{a \in A \setminus A'}(\lambda(I_a)) = (\lambda(I_2)) = \frac{12-10}{2} = 1$. Finally, we compute $\max_{b \in B \setminus B'}(\lambda(I_b))$, but this equals $-\infty$ since B' will always be equal to J_1, so $B \setminus B'$ is the empty set. Consequently, we obtain*

$$P(\theta_1) = \max(2, 1, -\infty) = 2.$$

Next, we do the same computations for θ_3. Again, we can read from the visualization of the barcodes that $\Delta(I_2, J_1) = 9$. Moreover, we note that if $A' = I_2$, then $A \setminus A' = I_1$, so $\max_{a \in A \setminus A'}(\lambda(I_a)) = (\lambda(I_1)) = \frac{5-0}{2} = \frac{5}{2}$ and $\max_{b \in B \setminus B'}(\lambda(I_b)) = -\infty$, just like in the computations for the bijection θ_1. Therefore, we obtain

$$P(\theta_3) = \max(9, \frac{5}{2}, -\infty) = 9.$$

The bottleneck is defined to be the minimum of P over the bijections $\{\theta_1, \theta_2, \theta_3, \theta_4\}$, which is the same as the minimum over θ_1 and θ_3, as defined above. As a result, we get

$$d_B(\mathcal{B}^1, \mathcal{B}^2) := \min_{\theta} P(\theta) = \min(P(\theta_1), P(\theta_3)) = \min(2, 9) = 2.$$

Now that we have formulated what it means for barcodes to be close to each other, let us find a notion of what it means for compact metric spaces to be close to each other. We introduce the Gromov-Hausdorff distance for this purpose.

Definition 2.50 *Given two subspaces P, Q of a metric space Z, we define the **Hausdorff distance** between P and Q by*

$$d_H(P, Q) := \max \left\{ \sup_{p \in P} \inf_{q \in Q} d(p, q), \sup_{q \in Q} \inf_{p \in P} d(p, q) \right\}.$$

*For P and Q any compact metric spaces, define the **Gromomov-Hausdorff distance** between the two spaces to be*

$$d_{GH}(P, Q) := \inf_{\gamma, \kappa} d_H(\gamma(P), \kappa(Q)),$$

where $\gamma : P \to Z$ and $\kappa : Q \to Z$ are isometric embeddings into a metric space Z.

2.3.2 Stability Results

In this section, we state two stability results without proving them. We need the statements later in this thesis for proving a stability result for the topological novelty profile in Chapter 3.

Theorem 2.51 ([11], [12]) *For any finite metric spaces P, Q and $i \geq 0$,*

$$d_B(\mathcal{B}_i(P), \mathcal{B}_i(Q)) \leq d_{GH}(P, Q).$$

The following variant of Theorem 2.51, stated in slightly different language in [13], can be proven by a slight modification of the proof of Theorem 2.51.

Theorem 2.52 ([13], Proposition 5.6) *For finite metric spaces $P \subseteq Q$ and $i \geq 0$,*

$$d_B(\mathcal{B}_i(P), \mathcal{B}_i{}^S(Q)) \leq \frac{1}{2} d_H(P, Q),$$

where $\mathcal{B}_i{}^S(Q)$ is the barcode obtained by shifting each interval of $\mathcal{B}_i(Q)$ to the right by $\frac{1}{2} d_H(P, Q)$.

Chapter 3

Novelty Profiles

In this chapter, we introduce the novelty profile, a statistic of an evolving population which quantifies how recombination contributes to genetic diversity. We start by building up a mathematical framework representing genetic dynamics in an evolving population. We use a model called evolutionary history indexed by a phylogenetic graph. After that, we introduce the temporal and the topological novelty profile. We show that the topological novelty profile is bounded from above by the temporal novelty profile and we present a decisive advantage of the topological novelty profile, namely that, in contrast to the temporal novelty profile, it is stable with respect to time. In this section, we follow [1].

3.1 Phylogenetic Graphs and Evolutionary Histories

We want to develop a model that represents evolutionary relationships and allows for recombination events. We model evolutionary relationships with so-called phylogenetic graphs.

Definition 3.1 (Phylogenetic Graph) *A **phylogenetic graph** is a finite directed acyclic graph G such that*

1. *G has a unique vertex v_0, the root with in-degree 0, and*
2. *each vertex of G has in-degree at most 2.*

*We call a vertex in G of in-degree 1 a **clone**, and a vertex of in-degree 2 a **recombinant**. If (v, w) is a directed edge in G, we say that v is a **parent** of w. We define a **rooted tree** to be a phylogenetic graph with no recombinants.*

Example 3.2 *Consider the graphs in Figure 3.1. The graph in a) is a phylogenetic graph since it is finite, directed, acyclic, the vertex at the top is unique vertex v_0, the root has in-degree 0, and each vertex of G has in-degree at most 2. The graph in b) is not a phylogenetic graph since the vertex at the bottom has three incoming edges. The graph depicted in c) is not a phylogenetic graph either since it is cyclic, and the graph in d) has two vertices with in-degree 0 so it is also not a phylogenetic graph.*

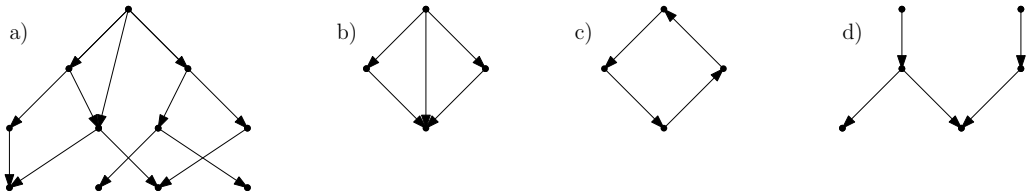

Figure 3.1: In a), we can see a phylogenetic graph. The figures in b), c), and d) do not represent phylogenetic graphs.

Definition 3.3 *For a rooted acyclic graph G with vertex set V and $S \subseteq V$, we say that $v \in S$ is the **minimum** or **minimal element** of S if for all $s \in S$, any directed path from r to s in G contains v.*

Note that G may not have a minimum, but if a minimal element exists, then it is unique.

We now define the evolutionary history, with the aim of mathematically formalizing biological ancestral relationships. Evolutionary histories are often referred to as ancestral recombination graphs (ARGs). However, we do not make use of this notion as we want to make a clear distinction between a history and its underlying phylogenetic graph.

Definition 3.4 (Evolutionary History) *Denote by* Set *the collection of all finite sets. For G a phylogenetic graph with vertex set V, an **evolutionary history indexed by** G is a map $\mathcal{E} : V \to$ Set satisfying the following three properties:*

1. *If w is a clone with parent v, then $\mathcal{E}_v \subseteq \mathcal{E}_w$.*
2. *For each $m \in \bigcup_{v \in V} \mathcal{E}_v$, the set $\{v \in V \mid m \in \mathcal{E}_v\}$ has a minimal element.*
3. *If w is a recombinant with parents u and v, then $\mathcal{E}_u \cap \mathcal{E}_v \subseteq \mathcal{E}_w \subseteq \mathcal{E}_u \cup \mathcal{E}_v$.*

*We call the elements of $\mathcal{E}_v$ **mutations**.*

An example for an evolutionary history can be found in Figure 3.2. For the moment, we can ignore the time function shown. It only becomes relevant as soon as we analyze novelty profiles.

We look at differences between the $\mathcal{E}_v$ for vertices v of the graph through which the evolutionary history $\mathcal{E}$ is indexed. For this, we introduce the following metric on finite sets.

Definition 3.5 *The **symmetric difference metric** d on any two finite sets $\mathcal{E}_v, \mathcal{E}_w$ from an evolutionary $\mathcal{E}$ indexed by a phylogenetic graph with vertex set V and $v, w \in V$ is*

$$d(\mathcal{E}_v, \mathcal{E}_w) := |\mathcal{E}_v \triangle \mathcal{E}_w| = |(\mathcal{E}_v \cup \mathcal{E}_w) \setminus (\mathcal{E}_v \cap \mathcal{E}_w)|.$$

*We call $d(\mathcal{E}_v, \mathcal{E}_w)$ the **symmetric distance** of $\mathcal{E}_v$ and $\mathcal{E}_w$. We denote the resulting metric space as* $\operatorname{met} \mathcal{E}$*, or, when no confusion is likely, simply as $\mathcal{E}$.*

When we write d, we always mean the symmetric distance from now on. For a history $\mathcal{E}$ indexed by a phylogenetic graph G with vertex set V, this is precisely the metric induced

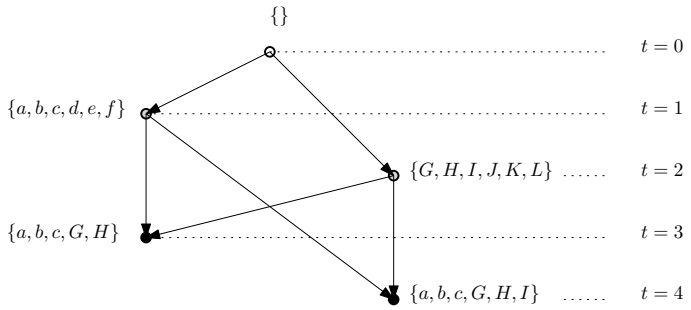

Figure 3.2: Phylogenetic graph indexed by an evolutionary history and time function. We denote the vertex on the top (the unique root) by v_0, the vertex at $t = 1$ by v_1, the vertex at $t = 2$ by v_2, the recombinant an $t = 3$ by r_1, and the recombinant at $t = 4$ by r_2.

by the cardinality of the symmetric distance on the collection of all finite sets, restricted to the set $\{\mathcal{E}_v \mid v \in V\}$. This means that we actually get a metric space.

Example 3.6 *We want to calculate the symmetric distance between two sets in the history from Figure 3.2. Exemplarily we calculate $d(\mathcal{E}_{v_1}, \mathcal{E}_{r_1})$, where v_1 is the left grey vertex and r_1 is the left black vertex.*

$$\begin{aligned} d(\mathcal{E}_{v_1}, \mathcal{E}_{r_1}) &= |(\{a,b,c,d,e,f\} \cup \{a,b,c,G,H\}) \setminus (\{a,b,c,d,e,f\} \cap \{a,b,c,G,H\})| \\ &= |\{a,b,c,d,e,f,G,H\} \setminus \{a,b,c\}| \\ &= |\{d,e,f,G,H\}| \\ &= 5 \end{aligned}$$

Let us now interpret the theory developed on a biological level. A phylogenetic graph describes the ancestral relationships between organisms in a history, and each set $\mathcal{E}_v$ specifies the genome (genetic information) of an organism v in terms of the difference between that genome and some fixed (unspecified) reference genome. Hereby, translated back into the biological language, item 1 makes sure that any clone inherits all mutations from its parent (and possible some new mutations). Item 2 can be understood as follows: for every mutation that occurs, there is a unique origin of the mutation in some organism. Finally, item 3 makes sure that if both parents of a recombinant inherit a mutation, then the recombinant inherits it. Moreover, it stipulates that any mutation carried by a recombinant is inherited from a parent.
The first two properties are standard in phylogenetics and are often referred to as the **infinite sites assumption**. In real-world evolving populations, the infinite sites assumption may not always hold. In other words, the same mutation may occur in different organisms despite being absent in their common ancestors. Such mutations, termed homoplasies, may be observed in sampled data either if the per-site mutation rate is high (which is typical for species with short genomes, such as RNA viruses) or if the mutations confer high fitness. Homoplasies are typically rare for species with long genomes, as the probability of mutating twice at the same exact genetic site is small. If they do occur, homoplasies usually involve

few sites, so that the metric space underlying the history only differs slightly from that of a history only differs slightly from that of a history satisfying the infinite sites assumption.

3.2 Novelty Profiles

To theoretically underpin the topological approach to the study of genetic recombination developed in [3], we introduce novelty profiles. They were first proposed in [1], and provide stable statistic of an evolutionary history that not only counts recombination events but also quantifies how recombination creates genetic diversity. The basic idea behind novelty profiles is to measure the genetic difference between recombinants and their parents.

3.2.1 Temporal Novelty Profile

For a phylogenetic graph G with vertex set V, define a **partial order** on V by taking

$$v \leq w :\Leftrightarrow \exists \text{ directed path from } v \text{ to } w.$$

We say that $t : V \to \mathbb{R}$ is a **time function** if $t(v) < t(w)$ whenever $v < w$. We interpret $t(v)$ as the birth time of the organism v.

Definition 3.7 (Temporal Novelty Profile) *Given a history $\mathcal{E}$ induced by G, a time function $t : V \to \mathbb{R}$, and a recombinant r of G, we define the **temporal novelty of** r to be*

$$\mathcal{N}_t(r) := \min\{d(\mathcal{E}_v, \mathcal{E}_r) \mid t(v) < t(r)\}.$$

*We define $\mathcal{N}_t(\mathcal{E})$, the **temporal novelty profile of $\mathcal{E}$ (with respect to t)**, to be the list of temporal novelties $\mathcal{N}_t(r)$ for all recombinants r of G, sorted in decreasing order.*

Example 3.8 *In the following, we show that the temporal novelty profile of the graph in Figure 3.2 is* $(5, 1)$.

To begin with, notice that the root v_0 satisfies $\mathcal{E}_{v_0} = \{\}$, so for any recombinant r, it holds that $d(\mathcal{E}_{v_0}, \mathcal{E}_r) = |(\{\} \cup \mathcal{E}_r) \setminus (\{\} \cap \mathcal{E}_r)| = |\mathcal{E}_r|$.

We start by determining the symmetric distances of $\mathcal{E}_{r_1}$ and the sets from the history corresponding to vertices with time value at most 3 (because $t(r_1) = 3$). These vertices are v_0, v_1 and v_2. Since $|\mathcal{E}_{r_1}| = 5$, the temporal novelty of r_1 is at most 5.

We already know from Example 3.6 that $d(\mathcal{E}_{v_1}, \mathcal{E}_{r_1}) = 5$. We still have to carry out the remaining calculations.

$$\begin{aligned} d(\mathcal{E}_{v_2}, \mathcal{E}_{r_1}) &= |(\{G, H, I, J, K, L\} \cup \{a, b, c, G, H\}) \setminus (\{G, H, I, J, K, L\} \cap \{a, b, c, G, H\})| \\ &= |\{a, b, c, G, H, I, J, K, L\} \setminus \{G, H\}| \\ &= |\{a, b, c, I, J, K, L\}| \\ &= 7 \end{aligned}$$

The above computations imply that $\mathcal{N}_t(r_1) = 5$.

Now we focus on r_2. Once again, we want to determine its temporal novelty. Since $|\mathcal{E}_{r_2}| = 6$, the temporal novelty of r_2 is at most 6.

$$\begin{aligned} d(\mathcal{E}_{v_1}, \mathcal{E}_{r_2}) &= |(\{a,b,c,d,e,f\} \cup \{a,b,c,G,H,I\}) \setminus (\{a,b,c,d,e,f\} \cap \{a,b,c,G,H,I\})| \\ &= |\{a,b,c,d,e,f,G,H,I\} \setminus \{a,b,c\}| \\ &= |\{d,e,f,G,H,I\}| \\ &= 6 \end{aligned}$$

$$\begin{aligned} d(\mathcal{E}_{v_2}, \mathcal{E}_{r_2}) &= |(\{G,H,I,J,K,L\} \cup \{a,b,c,G,H\}) \setminus (\{G,H,I,J,K,L\} \cap \{a,b,c,G,H\})| \\ &= |\{a,b,c,G,H,I,J,K,L\} \setminus \{G,H,I\}| \\ &= |\{a,b,c,J,K,L\}| \\ &= 6 \end{aligned}$$

In addition, since $t(r_1) = 3 < 4 = t(r_2)$, we also have to determine $d(\mathcal{E}_{r_1}, \mathcal{E}_{r_2})$.

$$\begin{aligned} d(\mathcal{E}_{r_1}, \mathcal{E}_{r_2}) &= |(\{a,b,c,G,H\} \cup \{a,b,c,G,H,I\}) \setminus (\{a,b,c,G,H\} \cap \{a,b,c,G,H,I\})| \\ &= |\{a,b,c,G,H,I\} \setminus \{a,b,c,G,H\}| \\ &= |\{I\}| \\ &= 1 \end{aligned}$$

Therefore, it holds that $\mathcal{N}_t(r_2) = 1$. This implies that the temporal novelty profile of $\mathcal{E}$ is indeed $\mathcal{N}_t(\mathcal{E}) = (5, 1)$.

Notation 3.9 For two finite vectors x and y of the same length we use the following notation for the maximum norm:

$$d_\infty(x, y) = \max_i |x_i - y_i|.$$

Proposition 3.10 *For any time function, the temporal novelty profile is stable with respect to genetic perturbations in the sense that for any two histories $\mathcal{E}$ and $\mathcal{E}'$ indexed by the same phylogenetic graph $G = (V, E)$ such that for all $v \in V$, it holds that $d(\mathcal{E}_v, \mathcal{E}'_v) \leq \frac{\delta}{2}$, it holds that the maximal distance between two elements in the same position of the topological novelty profile of the two histories,*

$$d_\infty(\mathcal{N}_t(\mathcal{E}), \mathcal{N}_t(\mathcal{E}')) := \max_i |\mathcal{N}_t(\mathcal{E})_i - \mathcal{N}_t(\mathcal{E}')_i)|,$$

is at most δ.

Proof Suppose that $\mathcal{E}$ and $\mathcal{E}'$ are histories indexed by the same phylogenetic graph such that for all $v \in V$, it holds that $d(\mathcal{E}_v, \mathcal{E}'_v) \leq \frac{\delta}{2}$. By applying the triangle inqeuality twice, we obtain

$$\begin{aligned} d(\mathcal{E}_v, \mathcal{E}_w) - d(\mathcal{E}'_v, \mathcal{E}'_w) &\leq d(\mathcal{E}_v, \mathcal{E}'_v) + d(\mathcal{E}'_v, \mathcal{E}_w) - d(\mathcal{E}'_v, \mathcal{E}'_w) \\ &\leq d(\mathcal{E}_v, \mathcal{E}'_v) + d(\mathcal{E}'_v, \mathcal{E}'_w) + d(\mathcal{E}'_w, \mathcal{E}_w) - d(\mathcal{E}'_v, \mathcal{E}'_w) \\ &\leq \frac{\delta}{2} + \frac{\delta}{2} \\ &= \delta. \end{aligned}$$

Similarly, we compute that

$$-d(\mathcal{E}_v, \mathcal{E}_w) + d(\mathcal{E}'_v, \mathcal{E}'_w) \leq \delta.$$

Therefore, $\mathcal{E}$ and $\mathcal{E}'$ satisfy $|d(\mathcal{E}_v, \mathcal{E}_w) - d(\mathcal{E}'_v, \mathcal{E}'_w)| \leq \delta$ for all vertices v, w of G. Now take any time function t on the vertices of G. Then the above computations imply that

$$\begin{aligned} d_\infty(\mathcal{N}_t(\mathcal{E}), \mathcal{N}_t(\mathcal{E}')) &= \max_i |\mathcal{N}_t(\mathcal{E})_i - \mathcal{N}_t(\mathcal{E}')_i)| \\ &\leq \delta. \end{aligned}$$

So in particular, the difference between any two entries at the same position of the temporal novelty profile is at most δ. This implies that temporal novelty profile is indeed stable with respect to genetic perturbations. □

Now the question arises as to whether we also have stability with regard to the time function. In the proof of the following proposition, we see why this is not the case.

Proposition 3.11 *The temporal novelty profile is unstable with respect to perturbations of the time function.*

Proof Consider the history $\mathcal{E}$ from Example 3.8. For $\delta \in (-1, \infty)$, let t_δ be the time function shown in Figure 3.8 by changing the time value of the vertex on the bottom right from 4 to $3 + \delta$. Then for all $\delta \in (0, 1)$, we have

$$\mathcal{N}_{t_\delta}(\mathcal{E}) = (5, 1) \text{ and } \mathcal{N}_{t_{-\delta}}(\mathcal{E}) = (6, 1).$$

This yields $d_\infty(\mathcal{N}_{t_\delta}(\mathcal{E}), \mathcal{N}_{t_{-\delta}}(\mathcal{E})) = 1$, so it does not approach zero as δ approaches zero, thus we do not have stability. □

3.2.2 Topological Novelty Profile

Although the temporal novelty profile is very intuitive and stable with respect to genetic perturbations, it not being stable with respect to the time function is a huge drawback. Therefore, we define the topological novelty profile, which turns out to be stable with respect to both, perturbations of the genome, and perturbation of birth times. For the definition of the topological novelty profile, we need the notions of relative minimal spanning trees and collapses.

Definition 3.12 *Given a weighted graph and a forest $F \subseteq G$ (i.e., a vertex-disjoint collection of subtrees), we define a **spanning tree of** G **rel** F to be a spanning tree T of G containing F and we refer to it as **relative spanning tree**. We say that T is a **minimal spanning tree** of G **rel** F if the sum of the edge weights of T is as small as possible among all spanning trees of G rel F. In this case, we speak of a **minimal relative spanning tree**.*

Definition 3.13 *The operation of removing faces γ that are a superset of some fixed face τ (including τ itself) is called a **collapse**.*

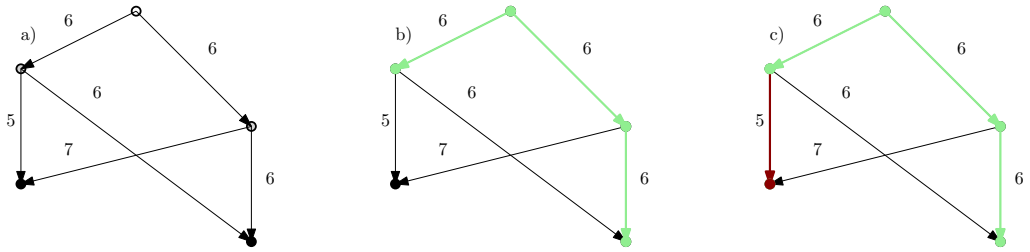

Figure 3.3: In a), we can see the graph G from Figure 3.2 with weights corresponding to the genetic distances between elements from the evolutionary history given in the figure. Sketched in green in b), we can see a forest $F \subseteq G$. Note that this forest only contains a single tree. In c) the dark red edge is the one we need to add in order to obtain a minimum spanning tree of G rel F.

Example 3.14 *The tree in Figure 3.4 a) can be interpreted as a simplicial complex. Then the three vertices on the bottom are all faces. Consider the face τ (colored red in b)) and note that the edge γ is a superset of τ. An example for a collapse is the deletion of τ and its superset γ (both colored in red in c)) from the graph. The graph we obtain by applying this operation is shown in d).*

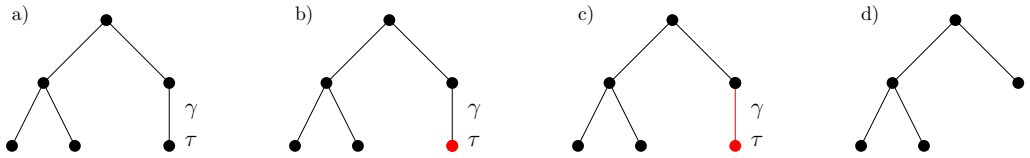

Figure 3.4: A collapse on a graph.

Remark 3.15 *Given a weighted graph and a forest $F \subseteq G$, by collapsing each tree in F to a point, the problem of finding a minimal spanning tree rel F reduces to the problem of finding an ordinary minimal spanning tree on a multigraph. The latter can be solved by first replacing all the given edges between vetices with the respective minimal one and then finding a minimal spanning tree of the reduced graph using Kruskal's Algorithm, described in [14]. The idea of Kruskal's Algorithm is to sort the edges of a graph by weight and then adding them one by one to the minimal spanning tree, whenever the edge added does not lead to a cycle.*

The following proposition helps us to identify minimal spanning trees.

Proposition 3.16 *A spanning tree T rel F is minimal if and only if for all i, the i^{th} smallest edge weight is less than or equal to the i^{th} smallest edge weight in any other spanning tree rel F.*

Proof Note that by Remark 3.15, it is sufficient to establish the result for ordinary spanning trees, i.e. in the case where F is an empty forest. We show the two implications separately.

"⇐": Assume that for all i, the i^{th} smallest edge weight in some spanning tree T is less than or equal to the i^{th} smallest edge weight in any other spanning tree rel F. Then the sum of the edge weights is as small as possible among all the spanning trees considered.

"⇒": Let T be a minimal spanning tree and let U be any other spanning tree. For the sake of deriving a contradiction, assume that for some i, the i^{th} smallest edge weight in T is greater than the i^{th} smallest edge weight in U. Denote this weight by ω. Now consider the subforests $T_\omega \subseteq T$ and $U_\omega \subseteq U$ consisting of all vertices and just those edge weights with weight at most ω.
By assumption, U_ω contains more edges than T_ω, so there exist two vertices u and v that lie in the same component of U_ω, but not in the same component of T_ω. In particular, there exists an edge $e = (x, y)$ for some vertices x and y along the path from u to v in U_ω such that x and y lie in different path-connected components of T_ω.
Now consider the path from x to y in T. Since x and y are in different path-connected components of T_ω, the path contains at least one edge e' with weight greater than ω. Consequently, replacing e with e' in T gives a new spanning tree with strictly smaller sum of edge-weights than in T. This contradicts the assertion that T is a minimal spanning tree and thus concludes the proof. □

We can deduce a corollary from this proposition that is useful once we define the topological novelty profile in Definition 3.18.

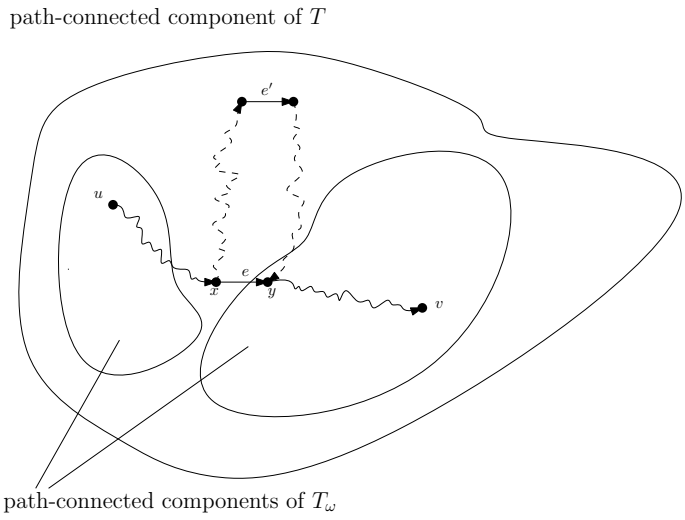

Figure 3.5: Construction from the proof of Proposition 3.16.

Corollary 3.17 *The collection of edge weights in a minimal spanning tree is independent of the choice of the tree.*

Proof It follows from Proposition 3.16 that the collection of edge weights in a minimal spanning tree rel F is determined by the property that the i^{th} smallest edge weight in it is at most the same as the i^{th} smallest edge weight in any other spanning tree rel F. □

We now finally define the topological novelty profile.

Definition 3.18 (Topological Novelty Profile) *For $\mathcal{E}$ a history indexed by a phylogenetic graph G, let F^G be the forest in G obtained by removing all edges pointing to recombinants. Let $\bar{G}$ denote the complete graph with same vertex set as G. Regard $\bar{G}$ as a weighted graph by taking the weight of edge (u, v) to be $d(\mathcal{E}_u, \mathcal{E}_v)$. Let T be a minimal spanning tree of $\bar{G}$ rel F^G. We define $\mathcal{T}(\mathcal{E})$, the* ***topological novelty profile*** *of $\mathcal{E}$, to be the list of distances*

$$\{d(\mathcal{E}_u, \mathcal{E}_v) | (u, v) \in T \setminus F^G\},$$

counted with multiplicity and sorted in descending order.

The following remark shows that the topological novelty profile is well-defined.

Remark 3.19 *Note that the elements in the topological novelty profile $\mathcal{T}(\mathcal{E})$ are determined by the collection of edge weights in a relative minimal spanning tree, so by Corollary 3.17, $\mathcal{T}(\mathcal{E})$ does not depend on the choice of a minimal spanning tree T.*

It is thus enough to find any minimal spanning tree of $\bar{G}$ rel F in order to determine the temporal novelty profile. Let us have a look at an example for computing the latter.

Example 3.20 *Consider the history indexed by the graph G in Figure 3.6. Let us determine its topological novelty profile.*

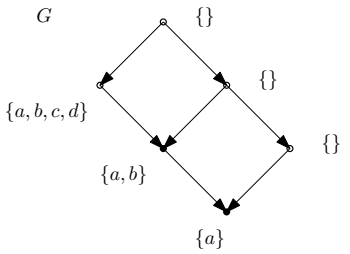

Figure 3.6: A history indexed by a graph G.

We start by determining F^G and $\bar{G}$. In our example, the forest F^G obtained from G by removing all edges pointing to recombinants consists of a single non-trivial tree and two vertices, and $\bar{G}$ is the complete graph with vertex set G, by definition. Both graphs are depicted in Figure 3.7.

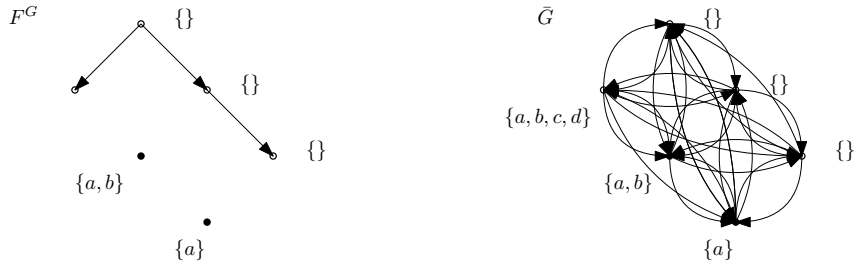

Figure 3.7: F^G (left) and $\bar{G}$ (right) corresponding to the history indexed by G in Figure 3.6, as defined in Definition 3.18.

We then determine a minimal spanning tree T of $\bar{G}$ rel F^G, i.e. a minimal spanning tree of $\bar{G}$ containing all edges of F^G using Proposition 3.16. Such a minimal spanning tree is shown in Figure 3.8.

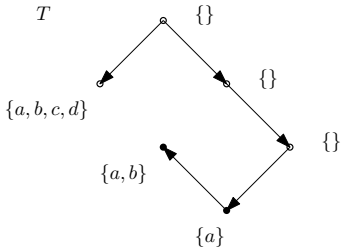

Figure 3.8: Minimal spanning tree T of $\bar{G}$ rel F^G ($\bar{G}$ and F^G from Figure 3.7).

The topological novelty profile is the list of distances $\{d(\mathcal{E}_u, \mathcal{E}_v) \mid (u, v) \in T \setminus F^G\}$, sorted in descending order. From the graphic representation of $T \setminus F^G$ in Figure 3.9 we can directly see that $\mathcal{T}(\mathcal{E}) = (1, 1)$.

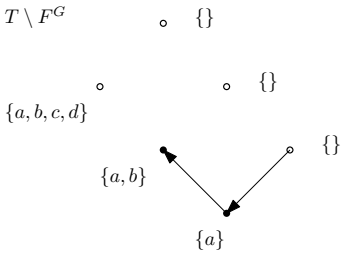

Figure 3.9: $T \setminus F^G$.

In the sequel, we use the following notation.

Notation 3.21 Given two lists A and B of numbers, each sorted in decreasing order and of the same length, we write $A \leq B$ if $|A| \leq |B|$ and for each $i \in \{1, \ldots, |A|\}$, $A_i \leq B_i$.

Proposition 3.22 *For any history $\mathcal{E}$ with time function t, the topological novelty profile is a lower bound for the temporal novelty profile, i.e.,*

$$\mathcal{T}(\mathcal{E}) \leq \mathcal{N}_t(\mathcal{E}).$$

Proof Suppose that $\mathcal{E}$ is a history indexed by some phylogenetic graph G. The idea of the proof is to construct a spanning tree T of $\bar{G}$ rel F^G such that the weights of the edges in $T \setminus F^G$ correspond to the temporal novelty profile. The result then follows from Proposition 3.16. To construct T, for each recombinant $r \in G$, we choose a vertex $v(r)$ in G with $t(v(r)) < t(r)$ such that $d(\mathcal{E}_{v(r)}, \mathcal{E}_r)$ is as small as possible among all such vertices. We take T to be the graph obtained from F^G by adding in the edge $(v(r), r)$ for each recombinant r. Then T is a tree and minimal by Proposition 3.16. □

The question arises as to what extent the temporal novelty profile and the topological novelty profile differ from each other. The next example shows that they are not in general the same, but in some cases, there exist time functions such that the temporal novelty profile equals the topological novelty profile.

Example 3.23 *The topological novelty profile for the history and time function in Figure 3.2 is the same as the temporal novelty profile computed in Example 3.8. However, we can define a different time function t' such that the topological novelty profile differs from the temporal novelty profile. Define, for example, t' to be the same time function as t but with $t = 1$ and $t = 2$ exchanged and $t = 3$ and $t = 4$ exchanged. Then the temporal novelty profile is given by $\mathcal{N}_{t'}(\mathcal{E}) = (6, 1) \neq (5, 1) = \mathcal{N}_t(\mathcal{E}) = \mathcal{T}(\mathcal{E})$.*

Now we come to the crucial stability result.

Proposition 3.24 (Stability of the Topological Novelty Profile) *Given histories $\mathcal{E}$ and $\mathcal{E}'$ indexed by the same phylogenetic graph G with $|d(\mathcal{E}_v, \mathcal{E}_w) - d(\mathcal{E}'_v, \mathcal{E}'_w)| \leq \varepsilon$ for all vertices v, w of G, we have*

$$d_\infty(\mathcal{T}(\mathcal{E}), \mathcal{T}(\mathcal{E}')) \leq \varepsilon.$$

Proof For $\mathcal{E}$ a history indexed by a graph G and F^G the forest defined in 3.18, define a filtration $\mathcal{F}$ by $\mathcal{F}_s := F^G \cup \mathbb{VR}(\mathcal{E})_s$, where we consider $\mathcal{E}$ to the metric space $\operatorname{met}\mathcal{E}$ with d the symmetric difference metric from Definition 3.5. We claim that the topological novelty profile $\mathcal{T}(\mathcal{E})$ of $\mathcal{E}$ is exactly the list of right endpoints of intervals in the barcode of the zeroth homology, possibly with some copies of 0 added in.
Indeed, when building up $\mathcal{F}_s$ for continuously increasing s, then a recombinant r persists until

$$\min_{\mathcal{E}_v \in \mathcal{E} \setminus \mathcal{E}_R} d(\mathcal{E}_r, \mathcal{E}_v) = 2s \iff s = \frac{1}{2} \min_{\mathcal{E}_v \in \mathcal{E} \setminus \mathcal{E}_R} d(\mathcal{E}_r, \mathcal{E}_v),$$

where we denote by $\mathcal{E}_R$ the set of elements from the history corresponding to $r \in R$ that previously died at $s' \leq s$. Note that the list of right endpoints of the persistence intervals of $\mathcal{F}_s$ may contain zeros. This is possible because it might be the case that $\mathcal{E}_v = \mathcal{E}_w$ although $v \neq w$ for vertices of the phylogenetic graph the evolutionary history is indexed by. The values computed above correspond precisely to the distances in $T \setminus F^G$ for a minimal spanning tree T of $\bar{G}$ rel F^G. □

Chapter 4

Novelty Profiles on Galled Trees

In this chapter, we define so-called galled trees, a form of phylogenetic graphs satisfying some extra properties that make sure that the graph structure does not become too complicated. Interpreted biologically, they represent evolution in a low combination regime, i.e., there are only few recombination events. We show that on galled trees, the temporal and the topological novelty profiles are equal. We introduce some notions from discrete Morse theory to prove the main theorem of this thesis. It roughly states that the topological novelty profile on galled trees is bounded from below by the list of life-times in the 1-dimensional persistent homology of the Vietoris-Rips complex of the history indexed galled tree, and the higher-dimensional persistent homology is trivial. We want to point out that in contrast to our notation in Chapter 2, we mainly use the usual notation for simplices in this chapter. Moreover, we work over $\mathbb{F}_2$, so we may omit orientations. We primarily follow [1]. For the parts on discrete Morse theory, we refer to [15] and [16].

4.1 Galled Trees

Our goal is to find bounds on novelty profiles. Our main bounds concern the special case that our phylogenetic graph is a galled tree.

We start by introducing two definitions that we need in order to define a galled tree.

Definition 4.1 *An undirected graph is a **loop** if its geometric realization is homeomorphic to a circle. In the set-up of a directed graph we say that a vertex which has no incoming arcs is a **source** and a vertex which has no outgoing arcs is a **sink**. We call a directed graph a **source-sink loop** if the following two conditions hold.*

1. *The undirected graph underlying G is a loop.*
2. *G has a unique source and a unique sink.*

Definition 4.2 *For directed graphs G and H, with a source v in G and any vertex w in H, we define a directed graph $G \vee_{v,w} H$ by taking the disjoint union of G and H and then*

identifying v and w. We call $G \vee_{v,w} H$ a ***sum*** *of G and H. We sometimes write $G \vee_{v,w} H$ simply as $G \vee H$, suppressing v and w.*

We can think of the sum of two directed graphs as the graph we get by "gluing" the source from one graph to a vertex of the other graph.

Remark 4.3 *We do not define the sum $G \vee_{v,w} H$ in the case that neither of the vertices v and w is a source.*

The definitions from above enable us to understand the notion of a galled tree. Note that in contrast to the definition in the paper [1], we define galled trees to be finite in order for Proposition 4.5 to hold. However, this adjustment is minimal, as we work with finite data sets in the application anyway.

Definition 4.4 (Galled Tree) *Let $\mathcal{A}$ be the smallest collection of directed acyclic graphs such that the following conditions are satisfied.*

1. *Each rooted tree is in $\mathcal{A}$.*
2. *Each source-sink loop is in $\mathcal{A}$.*
3. *If G and H are in $\mathcal{A}$, then so is each sum $G \vee H$.*

We define a ***galled tree*** *to be a finite graph isomorphic to one in $\mathcal{A}$.*

Informally, a galled tree is a graph obtained by iteratively gluing rooted trees and source-sink loops along single vertices, using the sum operation from Definition 4.2. An example for a galled tree can be found in Figure 4.1.

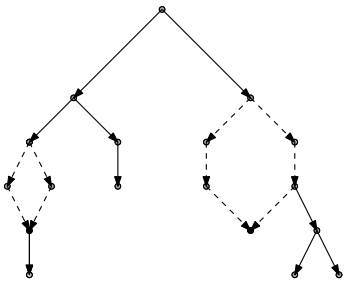

Figure 4.1: An example for a galled tree. Recombinants are printed in black. Note that the tree is obtained by alternatingly "gluing" dashed and non-dashed rooted trees / source-sink loops. In this case, the source-sink loops contained in the galled trees are dashed.

The following proposition is essential for our consideration of galled trees.

Proposition 4.5 *Any galled tree is a phylogenetic graph.*

Proof Let $\mathcal{A}$ be the smallest collection of directed acyclic graphs satisfying the conditions 1-3 from Definition 4.4. Let G be a graph isomorphic to one in $\mathcal{A}$. First note that G is a finite directed acyclic graph. We now check that the conditions 1 and 2 from Definition 3.1 are satisfied.

1. We claim that G has a unique root with in-degree 0. We prove the claim by induction over the minimum number n of summands that are source-sink loops or rooted trees since by the definition of galled trees, G can be written as a sum of source-sink loops and rooted trees.
   Base case ($n = 1$): If $n = 1$, then G is either a rooted tree or a source-sink loop, so it has a unique root.
   Induction step ($n \curvearrowright n + 1$): Assume that for any galled tree G consisting of n summands, i.e., a galled tree that can be written in the form

   $$G = (((G_1 \vee_{v_1,w_1} G_2) \vee_{v_2,w_2} G_3) \vee_{v_3,w_3} \ \ldots \vee_{v_{n-2},w_{n-2}} G_{n-1}) \vee_{v_{n-1},w_{n-1}} G_n,$$

   where the G_i's are either rooted trees or source-sink loops, it holds that G has a unique root. Consider a graph H which can be written as a sum of $n+1$ rooted trees and source-sink loops, i.e.,

   $$H = \underbrace{((H_1 \vee_{v_1,w_1} H_2) \vee_{v_2,w_2} \ldots \vee_{v_{n-2},w_{n-2}} H_{n-1}) \vee_{v_{n-1},w_{n-1}} H_n)}_{=H'} \vee_{v_n,w_n} H_{n+1}. \quad (4.1)$$

   Then H' has a unique has a unique root by the induction hypothesis and H_{n+1} has a unique root according to the base case. Gluing any vertex from H_{n+1} to the source of H' has the following consequence: if the vertex glued to the root is the root of H_{n+1}, it is still the only root in H. If another vertex gets glued to the root of H', then the root of H' has incoming vertices and thus no longer is a root. Consequently, the root of H_{n+1} is the unique root of H. The argument that the root exists and is unique when gluing H' to H_{n+1} works analogously.

2. It remains to show that each vertex of G has in-degree at most 2. Again, we prove the statement by induction on over the minimum number of summands n of G.
   Base case ($n = 1$): If $n = 1$, then G is either a rooted tree or a source-sink loop, so each vertex has in-degree at most 2.
   Induction step ($n \curvearrowright n + 1$): Assume that the assertion holds for any galled tree with minimum number of summands that are either rooted trees or source-sink loops being n. Now consider a galled tree H with this number being $n + 1$. Then $H = H' \vee H_{n+1}$ where H' is of the same form as in (4.1). The only vertex which may not have in-degree at most 2 therefore is the root at which H' and H_{n+1} are being added. If two roots get glued to each other, the in-degree is 0 and if a vertex which is not a root gets glued to a root, the in-degree of the vertices glued together is at most $0 + 2 = 2$. This yields the claim.

Consequently, a galled tree satisfies all the properties of a phylogenetic tree. □

We can now ask ourselves whether every phylogenetic graph is also a galled tree. But this is not true, as the following example shows.

Example 4.6 *Not every phylogenetic graph is a galled tree. Consider, for example, the phylogenetic graph indexed by a history $\mathcal{E}$ in Figure 3.2 and ignore the time function. It is*

not a galled tree since cycles formed by the reticulation events overlap in the sense that they share the same edges which contradicts the minimality condition in the definition of a galled tree.

The consideration of galled trees is of particular relevance, which is indicated by the validity of the following proposition.

Proposition 4.7 *For any history $\mathcal{E}$ indexed by a galled tree and time function t,*

$$\mathcal{N}_t(\mathcal{E}) = \mathcal{T}(\mathcal{E}).$$

Proof Suppose that $\mathcal{E}$ is a history indexed by a galled tree G. In order to determine the topological novelty profile, we have to find a minimum spanning tree of $\bar{G}$ rel F^G. Recall from Definition 4.4 that $\bar{G}$ is the complete graph with the same vertex set as G, and F^G is the forest in G obtained by removing all edges pointing to recombinants. A minimum spanning tree of $\bar{G}$ rel F^G can be constructed greedily (step-by-step and always choosing the "locally best solution") by starting with all edges contained in F^G and then considering the edges of $\bar{G} \setminus F^G$ in order of increasing weight (see Remark 3.15). In this construction, each edge in $\bar{G} \setminus F^G$ added to the relative minimum spanning tree can be chosen to connect a recombinant r to a vertex v of the source-sink loop in G that has r as its sink. This works since the weights corresponding to edges in the original graph G connecting vertices from a source-sink loop to its sink are smaller than any other edge weights of edges between between a vertex and the sink of the source-sink loop considered. Moreover, every vertex that is not a recombinant is connected via a walk to every vertex. In order not to obtain any cycles, we consider every recombinant only once (for obtaining a cycle, we would have to connect a recombinant to at least two other vertices). We then have that $t(v) < t(r)$ and $d(\mathcal{E}_v, \mathcal{E}_r) \leq d(\mathcal{E}_w, \mathcal{E}_r)$ for any other vertex w with $t(w) < t(r)$. This shows that indeed, it holds that $\mathcal{N}_t(\mathcal{E}) = \mathcal{T}(\mathcal{E})$. Since $\mathcal{E}$ was an arbitrary history, this concludes the proof. □

In order for the investigation of galled trees to be of biological relevance, the probability that a randomly constructed phylogenetic graph is a galled tree must be high. The following remark states that in scenarios characterized by low recombination rates, galled trees serve as effective models for studying evolution.

Remark 4.8 ([1], Remark 4.6, [17]) *Given a probabilistic model generating a phylogenetic graph, one may ask what the probability is of obtaining a galled tree. It can be shown analytically that the problem reduces to the study of a finite-state Markov chain. An analysis of this Markov chain yields, for fixed population size n, a system of linear equations $L(\rho)$ depending on a recombination parameter ρ, whose solution gives the probability $P(n, \rho)$ of obtaining a galled tree. Solving these linear systems numerically for various values of ρ and n, we observe that as ρ tends to 0, $P(n, \rho)$ tends to 1. This indicates that histories indexed by galled trees are reasonable models.*

4.2 Barcodes of Histories Indexed by Galled Trees

Now that we have clarified the relevance of galled trees in the study of reticulate evolution, let us analyze them using means from algebraic topology. We start with a result that reveals a connection between the topological novelty profile and the 0^{th} persistence barcode of an evolutionary history.

Notation 4.9 For $\mathcal{B}$ a barcode, we denote by lenghts($\mathcal{B}$) the list of intervals of $\mathcal{B}$, sorted in descending order.

Proposition 4.10 *Suppose we are given a history $\mathcal{E}$ and $\delta > 0$ such that $d(\mathcal{E}_v, \mathcal{E}_w) < \delta$ whenever w is a clone with parent v. Then the lists obtained from* lengths($\mathcal{B}_0(\mathcal{E})$) *and $\mathcal{T}(\mathcal{E})$ by removing all entries less than δ are equal.*

Proof Suppose we are given a history $\mathcal{E}$ indexed by a graph G and $\delta > 0$ such that $d(\mathcal{E}_v, \mathcal{E}_w) < \delta$ whenever w is a clone with parent v. We now compute the persistent homology of the Vietoris-Rips complex of $\mathcal{E}$. Note that since the distance between a clone and its parent is smaller than δ for all elements $\mathcal{E}_v \in \mathcal{E}$ in the history, for $\mathbb{VR}(\mathcal{E})_{\frac{\delta}{2}}$, it holds that, when deleting edges pointing to recombinants, all vertices that are not recombinants are parts of trees and there is precisely one more tree than recombinants in the graph. In particular, the trees form precisely the forest F^G from the definition of the topological novelty profile (possibly, there are additional edges connecting recombinants to other vertices, but those are exactly the ones where the distance from a recombinant to another vertex is smaller than δ). The list lengths($\mathcal{B}_0(\mathcal{E})$) now contains precisely the lifetimes of the recombinants that have lifetime at least δ. But these are precisely the entries of size at least δ in the topolological novelty profile which we get from the forest description above.□

The 0^{th} barcodes providing us with information about the topological novelty profile of a history, at least to some extent, suggests that the 0^{th} barcodes are useful in the study of recombination. However, if in applications δ is large, or when we are only given a subsample of a history, the 0^{th} barcodes might not give useful insights regarding the topological novelty profile. In combination with Theorem 4.13 about barcodes indexed by trees, stated below, this motivates studying the relationship between the topological novelty profile and higher barcodes of a history. Before stating a theorem describing the relationship, we define tree-like spaces and formulate a proposition used in the proof of Theorem 4.13 and also useful later in this thesis.

Definition 4.11 *An undirected tree with a non-negative weight function on its edges is a weighted tree. A metric space P is called* ***tree-like*** *if it is isometric to a subspace of a metric space arising from the shortest-path metric on a weighted tree.*

Proposition 4.12 ([3], Theorem 2.1, [1], Proposition 6.3) *If P is a tree-like metric space, then for all $r \in [0, +\infty)$, each component of $\mathbb{VR}(P)_r$ is contractible. Hence, $\mathcal{B}_i(P) = \emptyset$ for $i \geq 1$.*

Theorem 4.13 ([3]) *If G is a tree, $\mathcal{E}$ is a history indexed by G, and $S \subseteq \mathcal{E}$, then $\mathcal{B}_i(S) = \emptyset$ for $i \geq 1$.*

After adapting Proposition 4.12 to the right setting and noting that the metrization of a subset of a history indeed by a tree is tree-like, we can prove Theorem 4.13.

4.2.1 Metric Decomposition of an Evolutionary History

In the following, we reduce the problem of understanding the relation between the topological novelty profile of a history indexed by a galled tree and higher barcodes to a more accessible problem. More precisely, we split the problem into the study of galled trees and tree-like metric spaces. This requires some preparatory work.

Definition 4.14 *A **based metric space** is a metric space P, together with a choice of **basepoint** $p \in P$.*

Definition 4.15 *For based metric spaces P and Q with base points $p \in P$, $q \in Q$, we regard the wedge sum $P \vee Q$ (**sum of based metric spaces**) as a metric space, with the metric given by*

$$d_{P\vee Q}(x,y) = \begin{cases} d_P(x,y) & \text{if } x,y \in P, \\ d_Q(x,y) & \text{if } x,y \in Q, \\ d_P(x,p) + d_Q(q,y) & \text{if } x \in P, y \in Q. \end{cases}$$

For based metric spaces P and Q, let $\mathbb{VR}(P)_r \vee \mathbb{VR}(Q)_r$ denote the wedge sum filtration, given by

$$(\mathbb{VR}(P) \vee \mathbb{VR}(Q))_r := \mathbb{VR}(P)_r \vee \mathbb{VR}(Q)_r.$$

Using the above two definitions, we can find a decomposition of barcodes of the sum of P and Q when P and Q are both finite based metric spaces.

Proposition 4.16 *For based metric spaces P and Q, the inclusion*

$$\mathbb{VR}(P) \vee \mathbb{VR}(Q) \hookrightarrow \mathbb{VR}(P \vee Q)$$

is an objectwise homotopy equivalence. In particular, for any $i \geq 0$,

$$\mathcal{B}_i(\mathbb{VR}(P) \vee \mathbb{VR}(Q)) = \mathcal{B}_i(P) \cup \mathcal{B}_i(Q).$$

The proposition can be proved using discrete Morse theory. In the intermezzo following, we introduce the terminology we need for the purpose of proving the statements based on discrete Morse theory in this chapter.

Intermezzo: Discrete Morse Theory

Discrete Morse Theory (DMT) is a combinatorial theory concerning topology-preserving collapses of cell complexes. We give some basic definitions.

Definition 4.17 *Given a graph G with no self-edges, we define a **matching** X in G to be a subset of the edges such that no two edges are incident to the same vertex.*

Definition 4.18 *For a simplicial complex S, the **Hasse graph** G_S of S is the directed graph with vertices the simplices of S and and edge from s to s' if and only if s' is a codimension-1 face of s (a face of s which has dimension one lower than s, also often referred to as facet). When we say that a matching X matches two simplices in S, then this refers to matchings in the Hasse graph G_S.*

An example for a simplicial complex and the corresponding Hasse graph is presented in Figure 4.2.

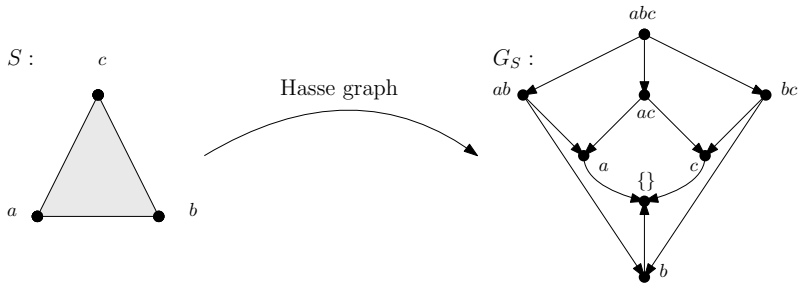

Figure 4.2: A simplicial complex S (on the left) and the corresponding Hasse graph G_S (on the right). As in Chapter 2, we denote the simplex $[a, b]$ by ab, and similarly for other simplices, to shorten the notation.

Definition 4.19 *A matching X in G_S is **acyclic** if when we modify the graph G_S by reversing the orientation of all edges in X, while leaving the orientation of all other edges unchanged, we obtain a directed acyclic graph.*

Example 4.20 *In Figure 4.3, we can see two examples for matchings on the Hasse graph from Figure 4.2. The matching in a) is not acyclic. Indeed, by reversing the arc directions of the arcs in the matching, we obtain a graph containing the cycle b, ab, a, ac, c, bc, b. The matching in b) is acyclic which we can see by trying out all possible walks starting at ab or bc and noticing that we always "get stuck" at some vertex.*

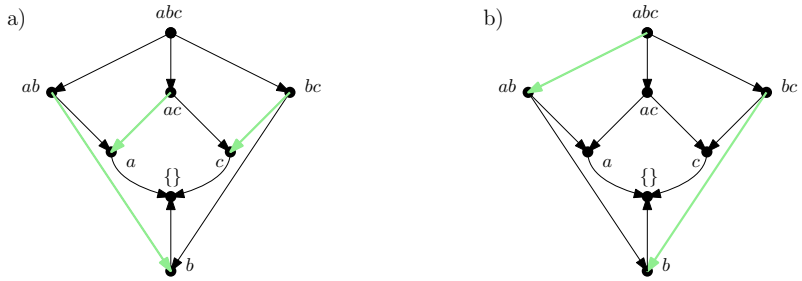

Figure 4.3: Two examples for matchings (arcs contained in the matching are colored in green). The matching in a) is not acyclic and the matching in b) is acyclic.

Definition 4.21 *A **discrete gradient vector field** (DGVF) X on S is an acyclic matching in G_S. A simplex $\sigma \in S$ is called **critical** in X if σ is not matched in X.*

Example 4.22 *Consider the matching $X = \{(abc, ab), (bc, b)\}$ from Figure 4.3. Since X is an acyclic matching in G_S, it defines a discrete gradient vector field. Since $a \in S$ is unmatched, $[a]$ is an example for a simplex which is critical in X.*

In the proof of Proposition 4.16, we also need the following result which we accept without giving a proof. It is often referred to as Main theorem of discrete Morse theory for CW complexes. A proof can be found in [18, Theorem 11.13].

Proposition 4.23 ([18], Theorem 11.13)

(i) Suppose that X is a DGFV on a finite simplicial complex S. Then S is homotopy equivalent to a CW-complex with exactly one cell of dimension i for each critical i-simplex of X.

(ii) If the critical simplices of X form a subcomplex $S' \subseteq S$, then in fact S deformation retracts onto S'.

This concludes our excursion into discrete Morse Theory.

Back to: Metric Decomposition of an Evolutionary History

We can now prove Proposition 4.16. The main idea is to define a DGVF on the Vietoris-Rips complex $\mathbb{VR}(P \vee Q)_r$ such that the critical simplices form a subcomplex of $\mathbb{VR}(P \vee Q)_r$ and then use Proposition 4.23 to infer the first part of the proposition. For the second part, we then recall a lemma from the chapter about persistence.

Proof of Proposition 4.16 Take any two finite metric spaces P and Q, and fix a radius $r \in [0, +\infty)$. Assume that a simplex $\sigma \in \mathbb{VR}(P \vee Q)_r$ in the Vietoris-Rips complex of the sum of P and Q contains vertices in both P and Q, but not the common vertex $p = q$. Then $\{p = q\} \cup \sigma$ is also in $\mathbb{VR}(P \vee Q)_r$ since it is the disjoint union of two simplices in the complex. With this in mind, define a DGVF X on $\mathbb{VR}(P \vee Q)_r$ by matching each simplex σ to $\{p = q\} \cup \sigma$ (acyclic matching by construction).
The critical simplices in X are, precisely those contained in $(\mathbb{VR}(P) \vee \mathbb{VR}(Q))_r$, again by construction, and consequently, they form a subcomplex of $\mathbb{VR}(P \vee Q)_r$. So by Proposition 4.23 (ii), $\mathbb{VR}(P \vee Q)_r$ deformation retracts onto $(\mathbb{VR}(P) \vee \mathbb{VR}(Q))_r$ and therefore, the inclusion

$$(\mathbb{VR}(P) \vee \mathbb{VR}(Q))_r \hookrightarrow \mathbb{VR}(P \vee Q)_r$$

is a homotopy equivalence.
Finally, we verify the result for persistence barcodes stated in the proposition. By Lemma 2.17, we already have that

$$\tilde{H}_i((\mathbb{VR}(P) \vee \mathbb{VR}(Q))_r) \cong \tilde{H}_i(\mathbb{VR}(P \vee Q)_r),$$

and thus

$$\mathcal{B}_i(P \vee Q) = \mathcal{B}_i(\mathbb{VR}(P) \vee \mathbb{VR}(Q)).$$

Consequently, it is sufficient to prove that

$$\mathcal{B}_i(\mathbb{VR}(P) \vee \mathbb{VR}(Q)) = \mathcal{B}_i(P) \cup \mathcal{B}_i(Q).$$

In order to prove this, we refer to [9, Corollary 2.25], according to which

$$\tilde{H}_i((\mathbb{VR}(P) \vee \mathbb{VR}(Q))_r) \to \tilde{H}_i(\mathbb{VR}(P)_r) \oplus \tilde{H}_i(\mathbb{VR}(Q)_r),$$

for each $r \in [0, +\infty)$ via a natural isomorphism. Since we have natural isomorphisms for every r, we also get an isomorphism of persistence modules. This yields

$$\mathcal{B}_i(P \vee Q) = \mathcal{B}_i(P) \cup \mathcal{B}_i(Q),$$

which is the desired result. □

Using Proposition 4.16, we can prove the following theorem.

Theorem 4.24 *Suppose G is a galled tree with $G = G^1 \vee ... \vee G^l$ for source-sink loops $G^1, \dots, G^k \subseteq G$ and rooted trees $G^{k+1}, \dots, G^l \subseteq G$ and that $\mathcal{E}$ is a history indexed by G. Let $\mathcal{E}^j$ denote the restriction of $\mathcal{E}$ to G^j for $j \in \{1, \dots, l\}$.*

(i) There is an objectwise homotopy equivalence from an iterated wedge sum of the filtrations $\mathbb{VR}(\mathcal{E}^j)$ to $\mathbb{VR}(\mathcal{E})$.

(ii) For $i \geq 1$,

$$\mathcal{B}_i(\mathcal{E}) = \bigcup_{j=1}^{k} \mathcal{B}_i(\mathcal{E}^j).$$

Proof

(i) We prove this statement by induction. The case $l = 1$ holds tautologically. Let now $l = 2$. Observe that when G is a phylogenetic graph with $G = G^1 \vee G^2$ for subgraphs $G^1, G^2 \subseteq G$, $\mathcal{E}$ is a history indexed by G, and $\mathcal{E}^1$ and $\mathcal{E}^2$ are the respective restrictions of $\mathcal{E}$ to G^1 and G^2, then

$$\operatorname{met} \mathcal{E} \cong \operatorname{met} \mathcal{E}^1 \vee \operatorname{met} \mathcal{E}^2.$$

Assuming that the statement holds for $l = n$, it holds for $l = n + 1$ by the inductive hypothesis:

$$\operatorname{met} \mathcal{E}^1 \vee ... \vee \operatorname{met} \mathcal{E}^n \vee \operatorname{met} \mathcal{E}^{n+1} = (\operatorname{met} \mathcal{E}^1 \vee ... \vee \operatorname{met} \mathcal{E}^n) \vee \operatorname{met} \mathcal{E}^{n+1}.$$

The result then follows from Proposition 4.16.

(ii) According to (i), it holds that $\bigvee_{j=1}^{l} \mathbb{VR}(\mathcal{E}^j)$ is homotopy equivalent to $\mathbb{VR}(\mathcal{E})$, so in particular, just like in the proof of Proposition 4.16, we inductively obtain that

$$\mathcal{B}_i(\mathcal{E}) = \bigcup_{j=1}^{l} \mathcal{B}_i(\mathcal{E}^j).$$

By Proposition 4.12, $\mathcal{B}_i(S) = \emptyset$ for every subset S of a history indexed by a tree and for every $i \geq 1$. So we get that $\mathcal{B}_i(\mathcal{E}^j) = \emptyset$ for $j \in \{k+1, \ldots, l\}$. This implies that

$$\mathcal{B}_i(\mathcal{E}) = \bigcup_{j=1}^{k} \mathcal{B}_i(\mathcal{E}^j).$$

□

4.2.2 Vietoris-Rips Filtrations of Almost Linear Metric Spaces

In this section, we introduce almost linear metric spaces and show some results about them. This is useful as the metrization of histories indexed by a source-sink loop is almost linear (Proposition 4.27).

Definition 4.25 *We say that a non-empty finite metric space P is **almost linear** if there is a point $p \in P$ such that $P \setminus \{p\}$ is isometric to a finite subset of $\mathbb{R}$. We call any such point p a **distinguished point**.*

Example 4.26 *An example for an almost linear metric space can be found in Figure 4.4. Note that not every almost linear metric space can be embedded isometrically into the two-dimensional Euclidean space.*

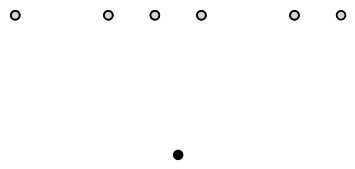

Figure 4.4: An almost linear metric space embedded in $\mathbb{R}^2$. The distinguished point is drawn in black.

Proposition 4.27 *If $\mathcal{E}$ is a history indexed by a source-sink loop, then* $\operatorname{met}\mathcal{E}$ *is almost linear.*

Proof Let $\mathcal{E}$ be a history indexed by a source-sink loop. Then there exists precisely one sink which is the unique recombinant of the phylogenetic graph indexed by $\mathcal{E}$. So when we delete this recombinant, the remaining graph is just a tree with a root and at most two branches going out of this root and there are no other edges in the three. Thus, $\operatorname{met}\mathcal{E} \setminus \{\mathcal{E}_p\}$ is isometric to a subset of $\mathbb{R}$. A visualization of this proof is shown in Figure 4.5. □

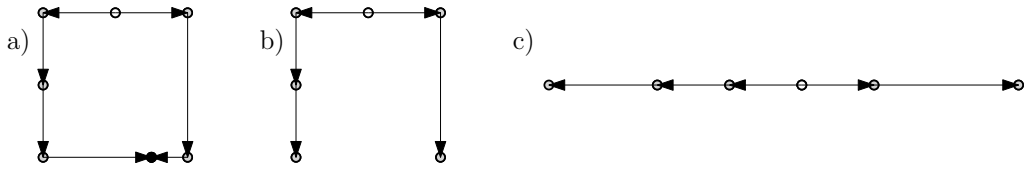

Figure 4.5: In a), we can see a source-sink loop. The vertex representing the source is printed in white and the vertex representing the sink is printed in black. In b), we can see the same graph, but without the sink and the edges going into the sink. In c), we can see an isometric embedding of the tree in b) into $\mathbb{R}^2$.

In view of Theorem 4.24 and Proposition 4.27, the problem of understanding the topology of the Vietoris-Rips filtrations of histories indexed by galled trees reduces to understanding the homology of almost linear metric spaces. Our goal is to prove the following statement about the topology of the Vietoris–Rips filtration of an almost linear metric space.

Theorem 4.28 *Let P be an almost linear metric space with distinguished point p.*

(i) For each $r \in [0, +\infty)$, the connected component C_r of $\mathbb{VR}(P)_r$ containing p is either contractible or homotopy equivalent to a circle, and each other component of $\mathbb{VR}(P)_r$ is contractible. In particular, $\mathcal{B}_i(P) = \emptyset$ for $i \geq 2$.

(ii) If C_r and $C_{r'}$ are both homotopy equivalent to circles and $r \leq r'$, then the inclusion $C_r \hookrightarrow C_{r'}$ is a homotopy equivalence. Thus, $\mathcal{B}_1(P)$ has at most one interval.

(iii) The unique interval of $\mathcal{B}_1(P)$, when it exists, has length at most $d(p, P \setminus \{p\})$ and is contained in the interval

$$[\frac{1}{2}d(p, P \setminus \{p\}), \frac{1}{2}\operatorname{diam}(P \setminus \{p\})),$$

where $\operatorname{diam}(P \setminus \{p\})$ denotes the diameter of the set $P \setminus \{p\}$.

Remark 4.29 *As a consequence of Theorem 4.24 (i), Proposition 4.27, and Theorem 4.28 (i), we obtain that for a history $\mathcal{E}$ indexed by a galled tree G and $r \in [0, +\infty)$, each component of $\mathbb{VR}(\mathcal{E})_r$ is homotopy equivalent to a bouquet of circles.*

In what follows, we provide several definitions and lemmas leading up to the proof of Theorem 4.28. We work in the following set-up.

Set-up: Let P be an almost linear metric space with distinguished point p, and let $r \in [0, +\infty)$ be such that $\mathbb{VR}(P)_r$ is connected. By choosing an isometric embedding $P \setminus \{p\} \hookrightarrow \mathbb{R}$, we may regard $P \setminus \{p\}$ as a subset of $\mathbb{R}$.

Definition 4.30 *Let $P_{left} \subseteq P \setminus \{p\}$ denote the set of points y such that*

1. *$[p, y] \notin \mathbb{VR}(P)_r$, and*
2. *there is no $w \in P \setminus \{p\}$ satisfying each of the following conditions:*
   - *$w < y$,*
   - *w and y lie in the same connected component of $\mathbb{VR}(P \setminus \{p\})$, and*
   - *$[p, w] \in \mathbb{VR}(P)_r$.*

In order to get a better understanding of Definition 4.30, we give an example.

Example 4.31 *Consider the point set P with Vietoris-Rips complex for some parameter r in Figure 4.6.*

In the following, we demonstrate how to determine P_{left}. In order to do so, we go through the points in $P \setminus \{p\}$ point by point and consider every point as a possible $y \in P_{left}$.

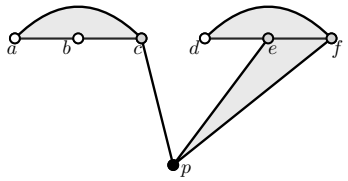

Figure 4.6: An example for a Vietoris-Rips complex on an almost linear space P and some choice of parameter r. The metric on P is not assumed to be the one given by the embedding in $\mathbb{R}^2$. The distinguished point is colored in black, points in P_{left} are colored in white, and the remaining points are grey.

a : *1.* $[a,p] \notin \mathbb{VR}(P)_r$*, so condition 1 from Definition 4.30 is satisfied.*

2. Since there exists no $w < a$ *in* P*, the second condition is satisfied.*

Thus, $a \in P_{left}$*.*

b : *1.* $[b,p] \notin \mathbb{VR}(P)_r$*, so condition 1 from Definition 4.30 is satisfied.*

2. The only $P \ni w < b$ *is* a*. Since* $[p,a] \notin \mathbb{VR}(P)_r$*, the second condition is satisfied.*

Consequently, $b \in P_{left}$*.*

c : *1. It is sufficient to note that* $[c,p] \in \mathbb{VR}(P)_r$*, so the first condition is not satisfied by* c*.*

The same applies to e *and* f*, so* $c, e, f \notin P_{left}$*.*

d : *Finally, consider* d*.*

1. $[d,p] \notin \mathbb{VR}(P)_r$*, so condition 1 from Definition 4.30 is satisfied.*

2. The points a, b *and* c *are smaller than* d*. However, none of them is in the same connected component of* $\mathbb{VR}(P \setminus \{p\})_r$ *as* d*, so the second condition is also satisfied for* d*.*

This implies that $d \in P_{left}$*.*

The above yields that $P_{left} = \{a, b, d\}$*.*

Lemma 4.32 *Within our set-up (*P *an almost linear metric space with distinguished point* p*, and* $r \in [0, +\infty)$ *such that* $\mathbb{VR}(P)_r$ *is connected),* $\mathbb{VR}(P)_r$ *deformation retracts onto* $\mathbb{VR}(P \setminus P_{left})_r$*.*

It is often convenient to use an alternative formulation of the acyclity condition which uses the following definition of an X-path.

Definition 4.33 *For a matching* X *in* G_S*, we define an* X***-path*** *to be a sequence of simplices in* S

$$\sigma_0, \tau_0, \sigma_1, \tau_1, \ldots, \sigma_m, \tau_m, \sigma_{m+1}$$

such that for each $j \in \{0, \ldots, m\}$*, the following are true:*

- σ_j *is a face of* τ_j *and* X *matches* σ_j *to* τ_j*,*

- σ_{j+1} *is a codimension-1 face of* τ_j,
- $\sigma_j \neq \sigma_{j+1}$.

We say that the X*-path* ***non-trivial*** *if* $m \geq 0$*, and* ***closed*** *if* $\sigma_0 = \sigma_{m+1}$.

Example 4.34 *For the matching* X *in Figure 4.3 b), the a non-trivial* X*-path is given by* $\sigma_0 = ab, \tau_0 = abc, \sigma_1 = bc$*. It is not closed since* $\sigma_0 \neq \sigma_1$.

We provide the following proposition without giving a proof. A proof can be found in [16, Chapter 2.2.1].

Proposition 4.35 ([15], Theorem 6.2, [16]) *A matching* X *in* G_S *is a DGVF (i.e. acyclic) if and only if there exists no non-trivial closed* X*-path.*

Assuming that this proposition holds, we prove Lemma 4.32. The idea is to define a DGVF on $\mathbb{VR}(P)_r$ that, in the Hasse graph $G_{\mathbb{VR}(P)_r}$, matches every simplex containing a point in P_{left} and no others. Then the critical simplices form the subcomplex $\mathbb{VR}(P \setminus P_{\text{left}})_r$ of $\mathbb{VR}(P)_r$ and we can deduce the result using Proposition 4.23 (ii).

Proof of Lemma 4.32 Define a DGVF W on $\mathbb{VR}(P)_r$ as follows: for $j \geq 2$ and a simplex

$$\sigma := [a_1, a_2, \ldots, a_j], \text{ where } a_1 < a_2 < \ldots < a_j \text{ (we regard } P \setminus \{p\} \text{ as a subset of } \mathbb{R}),$$

in $\mathbb{VR}(P)_r$ such that $a_1 \in P_{\text{left}}$ and a_2 is the point in P immediately to the right of a_1, W matches σ to its face $[a_1, a_3, \ldots, a_j]$. Note that W is acyclic since for any W-path

$$\sigma_0, \tau_0, \ldots, \sigma_m, \tau_m, \sigma_{m+1},$$

the τ_j's are strictly increasing with respect to the lexicographical order induced by the vertex ordering. If $m \geq 0$ and $\sigma_0 = \sigma_{m+1}$, then

$$\sigma_0, \tau_0, \ldots, \sigma_m, \tau_m, \sigma_0, \tau_0, \sigma_1$$

is a W-path with $\tau_m < \tau_0$, so there cannot exist a non-trivial closed W-path. Moreover, in the Hasse graph $G_{\mathbb{VR}(P)_r}$, W matches every simplex containing a point in P_{left} and no others, so the critical simplices of W form the subcomplex $\mathbb{VR}(P \setminus P_{\text{left}})_r$. Now we use Proposition 4.23 (ii) to conclude that $\mathbb{VR}(P)_r$ deformation retracts onto $\mathbb{VR}(P \setminus P_{\text{left}})_r$ since the critical simplices form a subcomplex of the simplicial complex $\mathbb{VR}(P)_r$. □

Note that this lemma allows us to assume, without loss of generality, that $P = P \setminus P_{\text{left}}$. So from now on, we assume that $P_{\text{left}} = \emptyset$.

Still with the goal of proving Theorem 4.28, we show some lemmas leading up to Lemma 4.41 stating that for an almost linear metric space P and any radius, each component of the Vietoris-Rips complex is contractible or deformation retracts to a wedge sum of finitely many circles. In order to show this, we define a DGVF Y on the Vietoris-Rips complex in a way that the critical simplices are i-simplices for $i \leq 1$. The exact set of critical simplices is given in Lemma 4.40 and it turns out in the proof of Theorem 4.28 why we want the

matching to be of exactly that nature. The construction allows, among other benefits, that with Proposition 4.23 (i), Lemma 4.41 follows directly.

We construct a DGVF Y on $\mathbb{VR}(P)_r$ in the following steps:

- **Step 1:** We define and order on the vertices in P.
- **Step 2:** We define a DGVF X on $\mathbb{VR}(P)_r$.
- **Step 3:** We prove a lemma suggesting a way to extend X to a DGVF Y with the properties desired for the proof of Lemma 4.41 by matching more simplices.

Step 1: We order the vertices in P by taking $\{p\}$ to be the minimum, and ordering $P \setminus \{p\}$ from left to right, via the chosen embedding of $P \setminus \{p\}$ into $\mathbb{R}$. Henceforth, it is our convention that the vertices of a simplex in $\mathbb{VR}(P)_r$ are always written in increasing order.

Step 2: We define X as follows.

Definition of X: If a simplex $\sigma = [a_1, a_2, \ldots, a_j]$ in $\mathbb{VR}(P)_r$ has a coface (a simplex τ is a coface of $\tilde{\tau}$ when $\tilde{\tau}$ is a face of τ) $a_0 \cup \sigma := [a_0, a_1, a_2, \ldots, a_j]$ with $a_0 < a_1$, then, in the Hasse graph $G_{\mathbb{VR}(P)_r}$, X matches σ to $a_0 \cup \sigma$ with a_0 as small as possible. X matches no other simplices.

A visualization of this definition can be found in Figure 4.7.

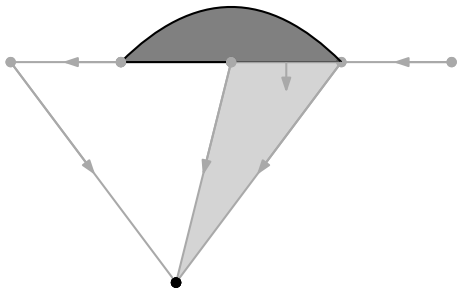

Figure 4.7: Illustration of the DGVF X on a Vietoris-Rips complex of an almost linear metric space with distinguished point the bottom vertex ordered first, and the other vertices ordered from left to right. Matched simplices are grey and critical simplices are black. The arrows indicate which simplices are matched with which (simplex of lower dimension points to simplex of higher dimension). Thus, X has one critical 0-simplex, two critical 1-simlices and one critical 2-simplex.

Step 3: Let us prove the following lemma which provides us with a characterization of critical simplices in X.

Lemma 4.36 *For $j \geq 2$, a simplex $[a_1, \ldots, a_j]$ is critical in X if and only if the following conditions are satisfied:*

1. $a_1 \neq p$,
2. $[q, a_1, \ldots, a_j] \notin \mathbb{VR}(P)_r$ *for any* $q < a_1$,
3. $[p, a_2, a_3, \ldots, a_j] \in \mathbb{VR}(P)_r$.

Proof

"⇒": Denote by σ a critical simplex. We show by contradiction that the three properties are satisfied.

1. Assume that condition 1 does not hold. Then σ matches to $[a_2, a_3, \ldots, a_j]$ in $G_{\mathbb{VR}(P)_r}$, and is therefore not critical, so we get a contradiction.
2. Assume that condition 2 does not hold. Then σ matches to a simplex of higher dimension in $G_{\mathbb{VR}(P)_r}$, which contradicts the fact that σ is a critical simplex.
3. Finally, assume that condition 3 does not hold. Since $\sigma = [a_1, a_2, \ldots, a_j]$ is a simplex in $\mathbb{VR}(P)_r$, it holds for any $k, l \in \{1, \ldots, j\}$ that $d(a_k, a_l) \leq 2r$. Furthermore, under the assumption that $[p, a_2, \ldots, a_j] \notin \mathbb{VR}(P)_r$ and because $[a_1, a_2, \ldots, a_j]$ is critical, it holds that $[a_2, \ldots, a_j]$ matches to some simplex $[q, a_2, \ldots, a_j]$ with $p < q < a_1$. Consequently, it also holds that $d(a_k, q) \leq 2r$ for all $k \in \{2, \ldots, j\}$. Moreover, since $p < q < a_1 < a_2$, the following inequalities are satisfied: $d(q, a_1) \leq d(q, a_2) \leq 2r$. This yields that $[p, a_1, a_2, \ldots, a_j] \in \mathbb{VR}(P)_r$, which contradicts the criticality of σ.

"⇐": Suppose that σ satisfies all three conditions. By condition 2, $\sigma = [a_1, \ldots, a_j]$ cannot be the simplex of lower dimension in a pair matched by X since by definition, if σ is the lower dimension simplex, X would have to match it to a simplex of the form $[q, a_1, \ldots, a_j] \notin \mathbb{VR}(P)_r$ for $q < a_1$, which contradicts the condition. By condition 3 and the definition of X, X matches the simplex $[a_2, \ldots, a_j]$ to $[p, a_2, \ldots, a_j]$ in the Hasse graph $G_{\mathbb{VR}(P)_r}$. With condition 1, it follows that σ cannot be the simplex of higher dimension in a matching since the only possible lower-dimensional simplex σ could be matched to is $[a_2, \ldots, a_j]$, and because $a_1 \neq p$, $[a_2, \ldots, a_j]$ is matched to a simplex different from σ. □

The two following corollaries follow directly from Lemma 4.36.

Corollary 4.37 *If $[a_1, \ldots, a_j]$ is critical in X, then $[p, a_1] \notin \mathbb{VR}(P)_r$.*

Proof Assume for the sake of deriving a contradiction that $[a_1, \ldots, a_j]$ is critical and a_1 is incident to p. Since $[a_1, \ldots, a_j]$ is a simplex in the Vietoris-Rips complex $\mathbb{VR}(P)_r$, it holds that $d(a_k, a_l) \leq 2r$. From condition 3 and from Lemma 4.36, it follows that $[p, a_2, \ldots, a_j] \in \mathbb{VR}(P)_r$ and therefore $d(p, a_k) \leq 2r$ for all $k \in \{2, \ldots, j\}$. This yields a violation of condition 2 from Lemma 4.36 since $p < a_1$ and $[p, a_1, \ldots, a_j] \notin \mathbb{VR}(P)_r$ implies that there exist points $x, y \in \{p, a_1, \ldots, a_j\}$ such that $d(x, y) > 2r$. However, the only pair not examined before is the pair of points (a_1, p) and by assumption, a_1 and p are incident in the Vietoris-Rips complex. □

Corollary 4.38 *For $j \geq 3$, denote by $[a_1, a_2, a_3, \ldots, a_j]$ a critical simplex in X. If there does not exist a vertex b which satisfies $[p, b] \in \mathbb{VR}(P)_r$ and $a_1 < b < a_2$, then the simplex $[a_1, a_3, \ldots, a_j]$ is critical in X.*

Proof We verify the three conditions from Lemma 4.36.

1. It holds that $a_1 \neq p$ because the simplex $[a_1, a_2, a_3, \ldots, a_j]$ is critical, so condition 1 from Lemma 4.36 applies.
2. For the sake of deriving a contradiction, assume that there exists a $q < a_1$ such that $\sigma = [q, a_1, a_3, \ldots, a_j] \in \mathbb{VR}(P)_r$. Then $q \neq p$ since by Corollary 4.37, $d(p, a_1) > 2r$. Thus, the vertices $q, a_1, a_3, \ldots, a_j$ can be isometrically embedded into the real line. This means that we can add a_2 to the simplex in order to obtain the simplex $[q, a_1, a_2, a_3, \ldots, a_j]$ and since $d(q, a_2) \leq d(q, a_3) \leq 2r$ and because $[a_1, a_2, a_3, \ldots, a_j]$ is a simplex in $\mathbb{VR}(P)_r$, we get that $[q, a_1, a_2, a_3, \ldots, a_j] \in \mathbb{VR}(P)_r$. This contradicts condition 2.
3. Since $[a_1, a_2, a_3, \ldots, a_j]$ is critical in X, by condition 3 from Lemma 4.36, it holds that $[p, a_2, a_3, \ldots, a_j] \in \mathbb{VR}(P)_r$, so in particular, since it is a face of this simplex, $[p, a_3, \ldots, a_j] \in \mathbb{VR}(P)_r$.

To conclude, $[a_1, a_3, \ldots, a_j]$ satisfies all conditions from Lemma 4.36 and is thus a critical simplex in X. □

Note that we did not use the fact that there does not exist a vertex b such that the simplex $[p, b]$ is in $\mathbb{VR}(P)_r$ and $a_1 < b < a_2$. However, this condition is necessary by the assumption that P_{left} is empty.

We continue by extending our definition of X to a DGVF Y. Lemma 4.36 suggests the following definition.

Definition of Y: For $[a_1, a_2, a_3, \ldots, a_j]$ a critical simplex in X such that there does not exist a vertex b such that $[p, b] \in \mathbb{VR}(P)_r$ and $a_1 < b < a_2$, we match the simplex $[a_1, a_2, a_3, \ldots, a_j] \in \mathbb{VR}(P)_r$ to $[a_1, a_3, \ldots, a_j]$ (critical in X by Corollary 4.38). We then take all matched pairs in $Y \setminus X$ to be of this form.

A visualization of Y is shown in Figure 4.8.

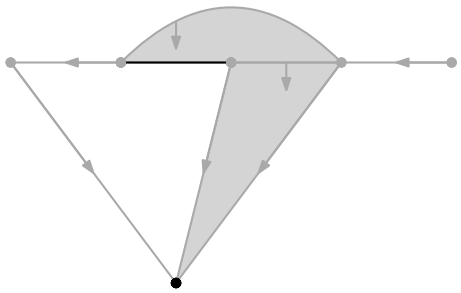

Figure 4.8: Illustration of the extension of the DGVF X of Figure 4.7 to the DGVF Y. Y contains one pair of matched simplices not in X: the curved 1-simplex in the top of the figure now is matched to its coface. Thus, Y has two critical simplices: a critical 0-simplex and a critical 1-simplex.

Remark 4.39 *[1, Lemma 6.22] assures that Y is acyclic, hence a DGVF.*

This finishes our construction of Y. In the next step we describe the critical simplices of Y.

Lemma 4.40 *The critical simplices of Y are $[p]$ and the 1-simplices $[a_1, a_2]$ such that*

1. *$[a_1, a_2]$ satisfies the conditions of Lemma 4.36, and*
2. *$[p, b] \notin \mathbb{VR}(P)_r$ for all $a_1 < b < a_2$.*

In particular, Y has a single critical 0-simplex and no critical simplices of dimension greater than one.

Proof First, note that all simplices matched by X are also matched by Y. Furthermore, Y matches all critical simplices in X of the form $[a_1, \ldots, a_j]$ for $j \geq 3$ to $[a_1, a_3, \ldots, a_j]$. So the only simplices remaining, besides the 0-simplex $[p]$ are the 1-simplices $[a_1, a_2]$ for which there does not exist a b such that $a_1 < b < a_2$ and $[a_1, b, a_2]$ is critical. We therefore examine under which condition the latter simplex is critical. For this, we use Lemma 4.36. Condition 1 is always satisfied because the simplex $[a_1, a_2]$ is critical by assumption and condition 2 is always satisfied as well since $[q, a_1, a_2] \notin \mathbb{VR}(P)_r$ implies in particular that $[q, a_1, b, a_2] \notin \mathbb{VR}(P)_r$. Thus, we have to check when exactly $[p, b, a_2] \in \mathbb{VR}(P)_r$. We know by assumption that $[p, a_2] \in \mathbb{VR}(P)_r$ and $[a_1, b, a_2] \in \mathbb{VR}(P)_r$ since $[a_1, a_2]$ is a critical simplex in X and $a_1 < b < a_2$. Thus, $[a_1, b, a_2]$ is critical in X if and only if $[p, b] \in \mathbb{VR}(P)_r$. This implies that $[a_1, a_2]$ is critical in Y if and only if $[a_1, a_2]$ satisfies the conditions of Lemma 4.36, and $[p, b] \notin \mathbb{VR}(P)_r$ for all $a_1 < b < a_2$.

In order to be able to prove Theorem 4.28, we present one more lemma. From here on, we no longer work with the set-up presented after 4.29.

Lemma 4.41 *For an almost linear metric space P and $r \in [0, +\infty)$, each component of $\mathbb{VR}(P)_r$ is contractible or deformation retracts onto a wedge sum of finitely many circles.*

Proof For P an almost linear metric space and $r \in [0, +\infty)$, consider $\mathbb{VR}(P)_r$. Since P is an almost linear metric space, every component of $\mathbb{VR}(P)_r$ not containing p is isometric to a subspace of a metric space arising from the shortest path metric on a weighted tree and is thus tree-like. So by Proposition 4.12, these components of $\mathbb{VR}(P)_r$ are contractible. Therefore we may assume without loss of generality that $\mathbb{VR}(P)_r$ is connected. Consequently, we may also assume that $P_{\text{left}} = \emptyset$ by Lemma 4.32. We conclude that the DGVF Y is defined on $\mathbb{VR}(P)_r$. The critical simplices of Y are described by Lemma 4.40: there is one critical 0-simplex, $[p]$, and there are possibly some critical 1-simplices, but there are no critical simplices of higher dimension. Now we can use Proposition 4.23 (i) to deduce the lemma, since according to this proposition, it follows from Y being a DGVF on a finite simplicial complex that the Vietoris-Rips complex is equivalent to a CW-complex with exactly one cell of dimension i for each critical i-simplex of Y. To conclude, recall that taking a 0-cell and attaching 1-cells yields a wedge sum of circles. Since we are given a finite simplicial complex, this wedge sum is finite, as well. □

Example 4.42 *Recall the filtration presented in Figure 2.4. Note that we can interpret the space given at time t_0 as an almost metric space with the upper vertices embedded in $\mathbb{R}$ and the upper vertices and the lower vertex not embedded in $\mathbb{R}^2$. At time t_3, the Vietoris-Rips complex is connected and $P_{left} = \emptyset$. So we can consider Y on the complex. Y matches $[a]$*

with $[p, a]$, $[b]$ with $[a, b]$, and $[c]$ with $[p, c]$. The critical simplices are $[p]$ and $[b, c]$ ($[b, c]$ satisfies all properties from Lemma 4.40). This information is sufficient to note that the homology of the complex at time t_3 has one 1-dimensional hole by the previous lemma.

After this preliminary work, we can now finally prove Theorem 4.28.

Proof of Theorem 4.28

(i) We prove Theorem 4.28 (i) as follows. We start by showing that the fundamental group we consider is free. We deduce that it is sufficient to show that it is either cyclic or trivial. We then find a basis of the fundamental group and prove that its cardinality is at most one.

First, note that as in the proof of Lemma 4.41, we may assume without loss of generality that $\mathbb{VR}(P)_r$ is connected and that $P_{\text{left}} = \emptyset$.

Claim *The fundamental group $\pi_1(\mathbb{VR}(P)_r, p)$ is free.*

Proof of Claim By Lemma 4.41, each component of $\mathbb{VR}(P)_r$ is contractible or deformation retracts onto a wedge sum of finitely many circles. Since the fundamental group of the wedge sum of circles is free by [9, Example 1.21], this yields that $\pi_1(\mathbb{VR}(P)_r, p)$ is free. □

It is thus sufficient to show that $\pi_1(\mathbb{VR}(P)_r, p)$ is trivial or cyclic. To show this, we find a basis of the fundamental group. In the following, we show that the DGVF Y provides us with a basis as for the fundamental group.

Let Γ denote the set of critical 1-simplices of Y as described by Lemma 4.40. For $\sigma = [b, c] \in \Gamma$ with $b < c$, let $a \in P \setminus \{p\}$ denote the maximum vertex such that $a < b$ and $[p, a] \in \mathbb{VR}(P)_r$. Such a always exist by our assumption that $P_{\text{left}} = \emptyset$. Indeed, $P_{\text{left}} = \emptyset$ implies that every $y \in P$ satisfies $[p, y] \in \mathbb{VR}(P)_r$ or there exists a $w < y$ in the same connected component of $\mathbb{VR}(P \setminus \{p\})_r$ as y with $[p, w] \in \mathbb{VR}(P)_r$. Since $[b, c]$ is critical by assumption, $[p, b] \notin \mathbb{VR}(P)_r$, so there must be a w as described above and we deonte this w by a.
Let us regard S^1 as a based topological space, with the base point denoted as 1, and let $\gamma_\sigma : S^1 \to [p, a] \cup [a, c] \cup [p, c]$ be a homeomorphism sending 1 to p.

Claim $G := \{\gamma_\sigma | \sigma \in \Gamma\}$ *is a basis for $\pi_1(\mathbb{VR}(P)_r, p)$.*

Proof of Claim For $\sigma \in \Gamma$, let S^1_σ denote a copy of S^1. The proof of Proposition 4.23 (i) presented in [18] gives a (not necessarily unique) homotopy equivalence

$$h : \mathbb{VR}(P)_r \to \bigvee_{\sigma \in \Gamma} S^1_\sigma,$$

mapping the interior of σ homeomorphically to $S^1_\sigma \setminus \{1\}$, so that $h \circ \gamma_\sigma$ is homotopic either to the inclusion $i_\sigma : S^1_\sigma \hookrightarrow \bigvee_{\sigma \in \Gamma} S^1_\sigma$ or to its inverse in $\pi_1(\bigvee_{\sigma \in \Gamma} S^1_\sigma, 1)$. Since h is a homotopy equivalence and $\{i_\sigma | \sigma \in \Gamma\}$ is a basis for $\pi_1(\bigvee_{\sigma \in \Gamma} S^1_\sigma, 1)$, we see that G is a basis for $\pi_1(\mathbb{VR}(P)_r, p)$. □

Now that we have found a description for a basis of the fundamental group, in order to show that the group is trivial or cyclic, it suffices to prove that the basis we have found contains at most one element.

Claim $|G| \leq 1$.

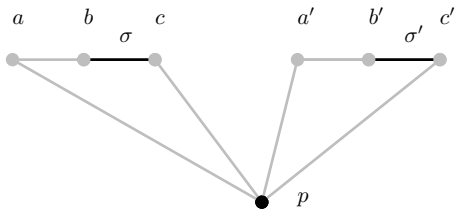

Figure 4.9: Illustration of the argument by contradiction that $|G| \leq 1$ in the proof of Theorem 4.28 (i). Critical simplices are black and matched simplices are gray.

Proof of Claim We prove the claim by contradiction, assuming that there are at least two different elements in the basis as follows. Consider the sketch in Figure 4.9. For $[b, c] = \sigma \in \Gamma$ with $b < c$, let $a < b$ be as above, and for $[b', c'] = \sigma' \in \Gamma$ with $b' < c'$, define $a' < b'$ in the same way. For the sake of deriving a contradiction, suppose $\sigma \neq \sigma'$. Then either $c \leq a'$ or $c' \leq a$. Switching the labels of σ and σ' if necessary, we may assume without loss of generality that $c \leq a'$. Note that since $[p, a], [p, c'] \in \mathbb{VR}(P)_r$, and thus $d(a, p) \leq 2r$ and $d(p, c') \leq 2r$, the triangle inequality yields

$$d(a, c') \leq d(a, p) + d(p, c') \leq 4r.$$

Thus, since $P \setminus \{p\}$ is isometric to a subset of $\mathbb{R}$, we have

$$d(a, c) + d(a', c') \leq d(a, c') \leq 4r.$$

In particular, this implies that $d(a, c) \leq 2r$ or $d(a', c') \leq 2r$, so either $[a, c] \in \mathbb{VR}(P)_r$ or $[a', c'] \in \mathbb{VR}(P)_r$. But then either γ_σ or $\gamma_{\sigma'}$ is nullhomotopic contradicting G being a basis for $\pi_1(\mathbb{VR}(P)_r, p)$. □

All in all, this implies that the fundamental group of the connected component C_r of $\mathbb{VR}(P)_r$ containing p has a basis and this basis consists of at most one element. Consequently, the fundamental group is either trivial or cyclic, so C_r must either be contractible or homotopy equivalent to a circle. If it is contractible, than all reduced homology groups are zero by [9, p. 111] and if it is homotopy equivalent to a circle, then all reduced homology groups starting from the second are zero by [9, Corollary 2.14]. This concludes the proof of (i).

(ii) As in the statement of the theorem, denote by C_r the component of $\mathbb{VR}(P)_r$ containing p. We show that for $r \leq r' \in [0, +\infty)$, if $C_r \simeq S^1 \simeq C_{r'}$, then the inclusion $C_r \hookrightarrow C_{r'}$ is a homotopy equivalence. Consider the generators $\gamma_\sigma : S^1 \to C_r$ and $\gamma_{\sigma'} : S^1 \to C_{r'}$ of the fundamental groups $\pi_1(C_r, p)$, respectively $\pi_1(C_{r'}, p)$. Note that since they

are generators of homotopy groups of spaces that are homotopic to a circle, they are non-trivial. Furthermore, recall that by definition it holds that $a < b < c$, and $a' < b' < c'$. Moreover, a is defined to be the maximum vertex such that $a < b$ and $[p, a] \in \mathbb{VR}(P)_r$ and a' is defined to be the maximum vertex such that $a' < b'$ and $[p, a'] \in \mathbb{VR}(P)_{r'}$. This implies that exactly one of the following must be true:

1. $c \leq a'$,
2. $c' \leq a$,
3. $a \leq a' < c' \leq c$.

Indeed, if item 1 is true, i.e., $c \leq a'$, then $a < b < c \leq a' < b' < c'$, so in particular $c' > a$, which contradicts 2 and $c' > c$, which contradicts 3. If item 2 is true, i.e., $c' \leq a$, then $a' < b' < c' < a < b < c$, thus $a' < c$, which contradicts 1 and $a > a'$, which contradicts 3. Finally, if items 1 and 2 both do not hold, then item 3 holds by the choice of a and a' specified above.

In the sequel, we show that neither the first, nor the second item can be true.

Claim *It is not possible that $c \leq a'$ (1.).*

Proof of Claim We prove the statement by contradiction using the triangle inequality, similar to way we did it in our proof of Theorem 4.28 (i). Suppose that $c \leq a'$. Then $d(a, p) \leq 2r$ and $d(c', p) \leq 2r'$ since $[a, p] \in \mathbb{VR}(P)_r$ and $[c', p] \in \mathbb{VR}(P)_{r'}$. Since $P \setminus \{p\}$ is embedded into the real numbers and by the triangle inequality, it holds that

$$d(a, c) + d(a', c') \leq d(a, c') \leq d(a, p) + d(c', p) \leq 2(r + r').$$

Consequently, $d(a, c) \leq 2r$ or $d(a', c') \leq 2r'$. This implies that $[a, c] \in \mathbb{VR}(P)_r$ or $[a', c'] \in \mathbb{VR}(P)_{r'}$. This yields a contradiction since this would imply that γ_σ or $\gamma_{\sigma'}$, which is not possible since they are both generators of non-trivial groups. □

Claim *It is not possible that $c' \leq a$ (2.).*

Proof of Claim The proof works analogously to the proof of the previous claim. □

Consequently, item 3, i.e. $a \leq a' < c' \leq c$, is always true. The next part of the proof is showing that if $a \neq a'$, then $[a, a'] \in C_{r'}$ and symmetrically if $c \neq c'$, then $[c, c'] \in C_{r'}$.

Claim *If $a \neq a'$, then $[a, a'] \in C_{r'}$.*

Proof of Claim Let $a \neq a'$. Then it must hold that $a < a'$. Since $[p, a] \in \mathbb{VR}(P)_r$ and $[p, c'] \in \mathbb{VR}(P)_{r'}$, we have that $d(a, p) \leq 2r$ and $d(p, c') \leq 2r'$, so by the triangle inequality and because the points a, b, c, a', b', c' are embedded in the real line,

$$d(a, a') + d(a', c') = d(a, c') \leq d(a, p) + d(p, c') \leq 2(r + r').$$

Therefore, $d(a, a') \leq 2r'$ or $d(a', c') \leq 2r'$, but since $\gamma_{\sigma'}$ is not nullhomotopic by assumption, we must have $d(a', c') > 2r' \geq 2r$, so $d(a, a') \leq 2r'$. Thus, $[a, a'] \in C_{r'}$, as desired. □

It follows that $[p, a, a'] \in C_{r'}$ since $d(p, a) \leq 2r \leq 2r'$ and $d(p, a') \leq 2r'$.

Claim *If $c \neq c'$, then $[c, c'] \in C_{r'}$.*

Proof of Claim This can be proved using a symmetric argument to the one above.□

So we also obtain that $[p, c, c'] \in C_{r'}$.

Let

$$j : C_r \hookrightarrow C_{r'}$$

denote the inclusion. We thus have that $j \circ \gamma_\sigma \simeq \gamma_{\sigma'}$. Since γ_σ and $\gamma_{\sigma'}$ are homotopy equivalences, j must be a homotopy equivalence as well, which concludes the proof of (ii).

(iii) Consider the set G of generators of the fundamental group $\pi_1(\mathbb{VR}(P)_r, p)$ given in the proof of Theorem 4.28 (i). Note that if

$$r \notin [\frac{1}{2}d(p, P \setminus \{p\}), \frac{1}{2}\mathrm{diam}(P \setminus \{p\})),$$

then $\pi_1(\mathbb{VR}(P)_r, p)$ is trivial. By (i), this implies that each component of the Vietoris-Rips complex is contractible. Hence, the unique interval of $\mathcal{B}_1(P)$, if it exists, is contained in

$$[\frac{1}{2}d(p, P \setminus \{p\}), \frac{1}{2}\mathrm{diam}(P \setminus \{p\})).$$

To finish the proof of (iii), we need to show that the unique bar of $\mathcal{B}_1(P)$ is of length at most $d(p, P \setminus \{p\})$. For this, we use the stability of persistent homology. Note that since $P \setminus \{p\}$ is isometric to a subset of $\mathbb{R}$, it is tree-like, so by Proposition 4.12, $\mathcal{B}_1(P \setminus \{p\}) = \emptyset$. Therefore, by Theorem 2.52,

$$2d_B(\mathcal{B}_1(P), \emptyset) = 2d_B(\mathcal{B}_1(P), \mathcal{B}_1(P \setminus \{p\}) \leq d_H(P, P \setminus \{p\}) = d(p, P \setminus \{p\}),$$

where the last equality follows from the definition of d_H. The bottleneck distance of any barcode $\mathcal{B}$ to the empty barcode is half the length of the longest interval of $\mathcal{B}$, so the result follows.

This completes the proof of Theorem 4.28. □

4.2.3 Interference about Recombination from Barcodes

We proceed with the formulation of the main theorem which states that the topological novelty profile is bounded from below by the list of life-times in the 1-dimensional persistent homology of the Vietoris-Rips complex of the set of vertices from the galled tree, and the higher-dimensional persistent homology is trivial. Recall from Proposition 4.7 that the topological novelty profile equals the temporal novelty profile for histories indexed by galled trees. This means particularly that the temporal novelty profile does not depend on the time function. Thus, we use the notation $\mathcal{N}(r)$ for the temporal novelty of a recombinant r for whatever time function is chosen.

Notation 4.43 For a phylogenetic graph G, let $\mathcal{R}^G$ denote the set of recombinants of G.

Theorem 4.44 (Main Theorem) *Let $\mathcal{E}$ be a history indexed by a galled tree G.*

(i) Theorems 4.24 (ii) and 4.28 (ii) yield a canonical injection

$$\varphi : \mathcal{B}_1(\mathcal{E}) \hookrightarrow \mathcal{R}^G$$

such that $\text{length}(I) \leq \mathcal{N}(\varphi(I))$ *for all* $I \in \mathcal{B}_1(\mathcal{E})$*. In particular,*

$$\text{lengths}(\mathcal{B}_1(\mathcal{E})) \leq \mathcal{T}(\mathcal{E}).$$

(ii) $\mathcal{B}_i(\mathcal{E}) = \emptyset$ *for* $i \geq 2$.

Proof For a galled tree G, each $r \in \mathcal{R}^G$ corresponds to an entry of $\mathcal{T}(\mathcal{E})$ via the correspondence

$$r \longleftrightarrow d(\mathcal{E}^{L_r \setminus r}, \mathcal{E}_r),$$

where we denote by L_r the source-sink loop corresponding to r and by $\mathcal{E}^{L_r \setminus r}$ the restriction of $\mathcal{E}$ to vertices of $L_r \setminus \{r\}$. With this setup, we prove the two statements from the theorem.

(i) Since the history of any source-sink loop is an almost linear space by Proposition 4.27, Theorem 4.28 (iii) yields

$$\text{length}(I) \leq d(\mathcal{E}^{L_r \setminus r}, \mathcal{E}_r) = \mathcal{N}(\varphi(I)).$$

(ii) Since $\mathcal{E}$ is a history indexed by a galled tree G, it can be interpreted as a history indexed by an iterated sum of source-sink loops and rooted trees, i.e. $G = G^1 \vee ... \vee G^l$ for source-sink loops $G^1, \ldots, G^k \subseteq G$ and rooted trees $G^{k+1}, \ldots, G^l \subseteq G$. So we know from Proposition 4.24 (ii) that

$$\mathcal{B}_i(\mathcal{E}) = \bigcup_{j=1}^{k} \mathcal{E}^j.$$

Since in this union, all histories are indexed by source-sink loops, met $\mathcal{E}^j$ is almost linear for $j \in \{1, \ldots, k\}$ by Proposition 4.27. Finally, Theorem 4.28 (i) yields the desired result. □

Remark 4.45 *Theorem 4.44 can be relaxed in terms of the complete sampling assumption and in terms of the galled tree assumption. The corresponding results can be found in [1, Chapter 7].*

Remark 4.46 *The results presented in this thesis are purely deterministic. However, one can observe that in a wide class of probabilistic models of genetic sequence evolution on galled trees, the intervals of the first persistence barcode are independent random variables. To understand the statistical properties of these barcodes, it suffices to understand the special case that the galled tree is a source-sink loop.*

Bibliography

[1] M. Lesnick, R. Rabadán, and D. I. S. Rosenbloom, "Quantifying Genetic Innovation: Mathematical Foundations for the Topological Study of Reticulate Evolution," *SIAM Journal on Applied Algebra and Geometry*, vol. 4, no. 1, pp. 141–184, 2020.

[2] C. Darwin, *On the Origin of Species by Means of Natural Selection*. London: John Murray, 1859.

[3] J. Chan, G. E. Carlsson, and R. Rabadán, "Topology of viral evolution," *Proceedings of the National Academy of Sciences*, vol. 110, pp. 18 566–18 571, 2013.

[4] G. E. Carlsson, "Topological pattern recognition for point cloud data.," *Acta Numerica*, 23:289–368, 2014.

[5] T. K. Dey and Y. Wang, *Computational Topology for Data Analysis*. Cambridge University Press, 2022.

[6] J. Matoušek, *Lectures on Discrete Geometry*. Springer New York, 2002.

[7] U. Bauer, M. Kerber, F. Roll, and A. Rolle, "A unified view on the functorial nerve theorem and its variations," *Expositiones Mathematicae*, vol. 41, no. 4, p. 125 503, 2023.

[8] E. H. Spanier, *Algebraic Topology*. Springer New York, 1981.

[9] A. Hatcher, *Algebraic Topology*. Cambridge Univ. Press, 2000.

[10] W. Crawley-Boevey, "Decomposition of pointwise finite-dimensional persistence modules," *Journal of Algebra and Its Applications*, vol. 14, no. 05, p. 1 550 066, 2015.

[11] F. Chazal, D. Cohen-Steiner, L. Guibas, F. Mémoli, and S. Oudot, "Gromov-Hausdorff stable signatures for shapes using persistence," in *Proceedings of the Symposium on Geometry Processing*, Eurographics Association, 2009, pp. 1393–1403.

[12] F. Chazal, V. De Silva, and S. Oudot, "Persistence stability for geometric complexes," *Geometriae Dedicata*, vol. 173, pp. 193–214, 2014.

[13] S. Harker, M. Kramárand, R. Levanger, and K. Mischaikow, "A comparison framework for interleaved persistence modules," *J. Appl. Comput. Topol.*, vol. 3, pp. 85–118, 2019.

[14] J. B. Kruskal, "On the Shortest Spanning Subtree of a Graph and the Traveling Salesman Problem," *Proceedings of the American Mathematical Society*, vol. 7, no. 1, pp. 48–50, 1956.

[15] R. Forman, "A User's Guide To Discrete Morse Theory," *Séminaire Lotharingien de Combinatoire*, vol. 48, 2001.

[16] N. A. Scoville, *Discrete Morse Theory* (Student Mathematical Library). American Mathematical Society, 2019, vol. 92.

[17] M. Arenas, G. Valiente, and D. Posada, "Characterization of reticulate networks based on the coalescent with recombination," *Molecular Biology and Evolution*, vol. 25, pp. 2517–2520, 2008.

[18] D. Kozlov, *Combinatorial Algebraic Topology*. Springer Berlin Heidelberg, 2008.

Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

Declaration of originality

The signed declaration of originality is a component of every written paper or thesis authored during the course of studies. In consultation with the supervisor, one of the following three options must be selected:

- (●) I confirm that I authored the work in question independently and in my own words, i.e. that no one helped me to author it. Suggestions from the supervisor regarding language and content are excepted. I used no generative artificial intelligence technologies[1].
- ( ) I confirm that I authored the work in question independently and in my own words, i.e. that no one helped me to author it. Suggestions from the supervisor regarding language and content are excepted. I used and cited generative artificial intelligence technologies[2].
- ( ) I confirm that I authored the work in question independently and in my own words, i.e. that no one helped me to author it. Suggestions from the supervisor regarding language and content are excepted. I used generative artificial intelligence technologies[3]. In consultation with the supervisor, I did not cite them.

Title of paper or thesis:

Application of Persistent Homology to the Study of Reticulate Evolution

Authored by:
If the work was compiled in a group, the names of all authors are required.

Last name(s):	**First name(s):**
Rosenberg	Naomi

With my signature I confirm the following:
- I have adhered to the rules set out in the Citation Guide.
- I have documented all methods, data and processes truthfully and fully.
- I have mentioned all persons who were significant facilitators of the work.

I am aware that the work may be screened electronically for originality.

Place, date	**Signature(s)**
Zurich, 05/07/2024	N. Rosenberg

If the work was compiled in a group, the names of all authors are required. Through their signatures they vouch jointly for the entire content of the written work.

[1] E.g. ChatGPT, DALL E 2, Google Bard
[2] E.g. ChatGPT, DALL E 2, Google Bard
[3] E.g. ChatGPT, DALL E 2, Google Bard