Doctoral Dissertation
Doctoral Program in Pure and Applied Mathematics (33.th cycle)

Topological data analysis applied to metric structures and data sets

Alessandro De Gregorio
* * * * * *

Supervisor
Prof. Francesco Vaccarino

Doctoral Examination Committee:
Prof. Massimo Ferri, Referee - University of Bologna
Prof.ssa Stefania Bellavia – Università degli Studi di Firenze
Prof. Stefano Scialò – Politecnico di Torino
Doctor Andrea Guidolin, Referee - KTH Royal Institute of Technology

Politecnico di Torino
20-07-2021

I hereby declare that, the contents and organisation of this dissertation constitute my own original work and does not compromise in any way the rights of third parties, including those relating to the security of personal data.

. . . . . .

Alessandro De Gregorio

Turin, 20-07-2021

Summary

The following thesis concerns theoretical and practical aspects of Topological Data Analysis (TDA). This is a branch of Mathematics emerged in the last 30 years, devoted to the analysis of data using concepts of Geometry and Algebraic Topology.

After a brief introduction to the theoretical aspects of Persistent Homology, the thesis is divided in three main chapters, each of them focusing on a different aspect of TDA.

In the first one, we study the concept of weak similarity between semi-metric spaces. A weak similarity is a bijection between two semi-metric spaces such that the ordering given by the distance between pair of points is preserved. It can be used to compare spaces from a point of view looser than isometry. We study this concept and analyse some invariants that can be used to determine whether two spaces are weakly similar or not. We introduce a dissimilarity measure between semi-metric spaces, based on the Gromov-Hausdorff distance, to measure how far they are from being weakly similar. We see how persistent homology can be used to study weak similarity. We introduce a dissimilarity between persistence diagrams from the point of view of weak similarity and prove a stability theorem between the introduced dissimilarities.

In the second one, we focus on the problem of skeletonization based on homological features. Our aim is to improve the concept of homological scaffold, giving a more solid theoretical foundation. The idea is to consider minimal length generators of the first homology group of the simplicial complexes in a filtration. This process yields a weighted graph, called homological scaffold, where the weight of each edge is given by the number of times the edge appears in one of the selected generators. We study the computational aspects and theoretical limits of this approach, then we make a comparison of the new homological scaffold with its predecessor on synthetic and real data sets.

In the last chapter, we study a new perspective on data sets based on persistent homology. We see that each data set can be seen as a collection of measurements on a finite set. These measurements will induce different metrics and filtering function on the original space, and these can be analysed with persistent homology. We see set equivariant operators as a natural tool to compare data sets and we investigate

how persistent homology and set equivariant operators are related.

Acknowledgements

There is a great number of people I want to thank for the continuous support they gave me in these 4 years.

My first note of gratitude goes to my supervisor, Francesco Vaccarino, for the guidance and help he gave me in these years. He has supported my work, giving me a plethora of ideas and letting me free to pursue my own intuitions. I particularly appreciate the fact that he made me reconsider many areas of mathematics I neglected in the past.

Along with my supervisor, my work would have not been the same without my research group. My thanks goes to Francesco Della Santa, the first who shared with me the joys and sorrows of the PhD, Marco Guerra, my rescue when I was the only student interested in TDA, Ulderico Fugacci, for his sympathy and mentorship, Antonio Mastropietro, for trying to teach me how to be a better researcher and foosball player, the newly arrived, but already integral part of the group, Silvia Buccafusco for helping me when I was overwhelmed by several projects and Elia Migliore for his enthusiasm, and Sara Scaramuccia for having faith in me and the countless conversations about math and beyond.

I thank the numerous collaborators that made my research more interesting. Amongst all, Facundo Mémoli, for inviting me to the Ohio Stat University and for his precious teachings, Giovanni Petri, Francesca Tombari and Nicola Quercioli, who I can always count on.

My sincere gratitude goes to my family for the continuous support and love they give me, and to my friends. I want to thank particularly Serena, my conscience, for being so hard when I deserve it and so gentle when I need it, Matteo, the brother I got to choose, whose patience is endless, and Alberto who inspires me to always give my best.

My journey has been a lot easier thanks to the wonderful other mathematicians I met at DISMA. My thanks go to Leo and Iman, the laughs we had are priceless, Luca, one of the nicest people I have ever met, Giada, for her kind heart, Federica, for being so rock, Martina, because she is awesome and loca, and Giulia for being so empathetic and caring.

I wish to thank also all the people met at the Smartdata lab. I want to especially thank Michele for the great Piedmontese lessons, and Dena and Francesca, for the

fun we had and for giving me an escape from mathematics when I needed it.

Lastly, my gratitude goes to Andrea Guidolin and Massimo Ferri, whose corrections and suggestions improved this thesis.

To my cousin Clara

Contents

List of Figures

Chapter 1

Introduction

Topological Data Analysis (TDA) is a branch of mathematics devoted to gathering topological and geometric information from data sets [19]. It has been developed in the last 30 years, starting from the seminal works of Patrizio Frosini [45, 31], Herbert Edelsbrunner [43, 33], and Vanessa Robins [88]. The core idea of TDA is that data, gathered from measurements of a phenomenon, can be endowed with a notion of "shape" and that the shape of data can be used to characterize the phenomenon that has to be described. Data coming from similar experiments will yield similar shapes, whereas if the measurements are gathered from different phenomena, the shapes will be affected by these differences. Therefore, the experimenter can examine if the measurements come from the same phenomenon checking the similarity of the shapes of data.
With the informal word shape, we mean something related to manifolds. We can think of an experiment as an act of sampling of points from a manifold that implicitly describes the relations that the variables of the experiment have to satisfy. Different events are associated with different manifolds, and one goal of TDA is to identify this diversity.
In order to endow data sets with a shape, computational topology seems to be the most natural tool to use [41]. There are several ways to obtain a topological space from a set of measurements. The discrete nature of real world experiments makes simplicial complexes an efficient way to obtain a topological space from data. In order to compare different topological spaces, we need computable topological invariants to give us a description of the object under study. The easiest and most remarkable example is homology. Once a simplicial complex is gathered from data, homology groups give a description of the "holes" or "cavities" present in the topological space. The computation of homology groups boils down to the reduction of certain boundary matrices, it is therefore easily implementable.

The simplicial complexes we can obtain from data are dependent on a set of hyper-parameters, for which often there is not an a priori right choice. Instead of a

single simplicial complex it becomes more convient to consider a *filtration*, that is a sequence of nested simplicial complexes. Each simplicial complex of the filtration is obtained with different values of the hyper-parameters. Therefore, the methods of TDA try to take into account the differences between the simplicial complexes in a filtration. The main character in TDA is Persistent Homology. Its purpose is to relate the homology groups of different simplicial complexes, associated by inclusion relations. In its simpler version, it keeps track of the births and deaths of simplcial homology classes along the different steps of an increasing filtration of simplicial complexes. It produces an object called *persistence diagram* that identifies the number of homological classes that appear at a certain step of a filtration and that disappear a another.
These diagrams are well suited for data analysis since thanks to a stability theorem it is proved that small changes in data induce small changes in the persistence diagram.

The purpose of this thesis is threefold: to analyse certain aspects of applications of topological data analysis to finite metric structures, to study how to bring back the information given by TDA on the original object under study and to introduce a framework suitable to study data set from the point of view of persistent homology.

The second chapter is devoted to introduce all the tools of algebraic topology and persistent homology that will be used in the thesis. There will be a brief reminder of category theory, that will be used to formulate some of the results.

The third chapter analyses the problem of weak similarity of finite semi-metric spaces. In the classical TDA pipeline, the input is often a finite semi-metric space, or a space with a dissimilarity defined on it. The obtained filtration of simplicial complexes, and the associated persistent homology, depends on the values attained by the distance function. We notice, thanks to the stability theorem, that the persistent homologies of two isometric spaces will be equivalent. Informally, we can say that persistent homology doesn't distinguish congruent shapes. On the other hand, we may be interested in comparing shapes from a point of view weaker than congruence. For example, when classifying polygons on the plane, we may be more interested in considering equivalent two polygons that are just similar, instead of congruent. In its current formulation, persistent homology doesn't do that, since similar polygons will be associated to different persistence diagrams. This chapter aims at studying metric spaces from this perspective. Weak similarity is an extension of the concept of isometry. We will say that two finite spaces are weakly similar when we can find a strictly increasing function that *rescales* one of the metric spaces into the other. The aim of the chapter is to study certain invariants to determine if two spaces are coarsely isometric. We will introduce a dissimilarity between semi-metric spaces for this purpose, and a dissimilarity between persistence diagrams, to compare the persistent homology of two spaces from the point of view of weak similarity.

The fourth chapter focuses on the problem of skeletonization of simplicial complexes from the point of view of persistent homology. It revolves around the concept of homological scaffold, introduced as a tool to bring back the information of persistent homology on the starting data, [85]. The idea is to consider the persistent homology classes that appear across the filtration. For each of them, a representative cycle is taken and then it is assigned a weight to each simplex of the last simplicial complex of the filtration, that depends on how many times the simplex occurs in one of the selected representative cycles. The main issue with this definition is that the choice of the representative cycles is not bound to any additional properties of the cycle. Technically, any of the elements in the equivalence class given by homology can be taken as a representative cycle, therefore different scaffolds could be obtained from the same data, leading to different interpretations of the results. To solve this problem, we impose an additional condition on the *minimality* of cycles, to obtain a more rigorous definition. Given a *length* function that assigns to each 1-cycle of $C_1(K)$ a positive real number, a minimal homology basis is a set of 1-cycles such that their associated homology class are a basis of $H_1(K)$ and the sum of their lengths is minimal. We implement the recent algorithm by Dey et al. [35] to obtain a minimal homology basis. From the minimal homology bases associated to each simplicial complex of the filtration, we realize the so called *minimal homological scaffold*. Even if defined in a more rigorous way, the minimal scaffold still suffers the problem of the arbitrary choice of the representative cycles of an homology class: it is possible that in the same homology class there are several cycles with the same minimal length and it may be possible that there are more than one single minimal homology basis. We show how to define the minimal scaffold in case of several representatives of minimal length. We also propose a probabilistic result that ensures that under mild conditions the minimal scaffold is unique. In the end, we compare the new scaffold with the old one using real an synthetic data. We see that they exhibit similar characteristics, and that the old scaffold can be used as a fast approximation of the minimal scaffold.

The fifth and last chapter is dedicated to studying data sets a collections of functions on a common finite set. Each element of a data set can be seen as a measurement on a domain. Different measurements endow the domain with different structures, and the aim of the chapter is to introduce a framework suitable to study these structures. We show that a data set induces a pseudo-distance on the domain, then, through the use of multiparameter persistent homology, a persistence module is assigned to each measurement in the data set. Interestingly, this module does not depend only on the measurement, but on the whole data set. Persistent homology can then be used to compare the same function in different data sets. In certain application problems, data sets come also with additional structure on them. For example, in images classification problems the results should be invariant under translation or rotation up to a certain degree of the images. We define *incarnations* of data sets to take this structure into account. An incarnation is given by a data

set and a set of *endomorphisms* of its domain. These maps induce an action on the data sets, so that two measurements are related if there is a function that transforms one into the other. In order to link different incarnations we need functions that preserve the structure given by the set of endomorphisms of the incarnation. To do so we make use of *set equivariant operators*. These morphisms, together with incarnations form a category, and we show some examples in which persistent homology becomes or not a functor of this category. In the end we introduce Grothendieck graphs as a data structure to encode the objects that we defined.

The main contribution of this thesis are the following:

- **Weak similarity of finite semi-metric spaces**
  - we study the notion of weak similarity for finite semi-metric spaces and we show how curvature sets can be used as incomplete invariants for weak similarity;
  - we define a *weak Gromov-Hausdorff dissimilarity* to compare semi-metric space from the point of view of weak similarity;
  - we see how the Vietoris-Rips filtration and persistent homology can be used to study the problem of weak similarity;
  - we define a *weak bottleneck dissimilarity* and we prove a stability theorem for the dissimilarities introduced.
- **Minimal homological scaffold**
  - we define a *minimal homological scaffold* and provide an implemented algorithm to compute it;
  - we analyse the uniqueness issues related with the homological scaffold;
  - we see an application of the minimal scaffold on a real dataset coming from a neuroscience experiment;
  - we compare the minimal scaffold with its predecessor on synthetic datasets and we see which properties they have in common.
- **Landscapes of data sets**
  - we define a category where data sets are seen as measurements on a finite set;
  - we see how to compute the persistent homology of a measurement in a data set and we study when it is possible to compare different data sets with persistent homology;

– we introduce the concept of *incarnation* to study the action of a set of transformations on the data set and we analyse the properties of incarnations;

– we investigate how *set equivariant operators* can be used to compare incarnations and introduce some examples of set equivariant operators;

– we give a description of our results in terms of Grothendieck graphs.

Chapter 2

Background on Topological Data Analysis

2.1 Basic concepts of Category Theory

The aim of this Section is to give a quick introduction to Category Theory, especially to declare the notation that we will use in the rest of the thesis.

The following Definitions are taken from Spivak [92].

Definition 1 (category)**.** A category $\mathbf{C}$ consists of a collection $\mathrm{Ob}(\mathbf{C})$, whose elements are called objects, such that:

- for every $x, y \in \mathrm{Ob}(\mathbf{C})$ there is a (possibly empty) set $Hom_{\mathbf{C}}(x, y)$, whose elements are called morphisms from x to y;
- for every object $x \in \mathrm{Ob}(\mathbf{C})$ there is a morphism $\mathrm{id}_x \in Hom_{\mathbf{C}}(x, x)$, called the identity morphism on x;
- for every three objects $x, y, z \in \mathrm{Ob}(\mathbf{C})$ there exists a function

$$\circ : Hom(y, z)_{\mathbf{C}} \times Hom(x, y)_{\mathbf{C}} \to Hom(x, z)_{\mathbf{C}}$$

  called composition.

For the sake of simplicity, given $f \in Hom_{\mathbf{C}}(x, y)$ and $g \in Hom_{\mathbf{C}}(y, z)$, we will denote by $g \circ f$ the element $\circ(g, f) \in Hom_{\mathbf{C}}(x, z)$. The morphisms have to satisfy the following requirements:

1. for every $x, y \in \mathrm{Ob}(\mathbf{C})$ and every morphism $f \in Hom_{\mathbf{C}}(x, y)$ it has to be

$$f \circ \mathrm{id}_x = f \quad \text{and} \quad \mathrm{id}_y \circ f = f; \tag{2.1}$$

2. for any $w, x, y, z \in \mathrm{Ob}(\mathbf{C})$ and any $f \in Hom_{\mathbf{C}}(w, x)$, $g \in Hom_{\mathbf{C}}(x, y)$, $h \in Hom_{\mathbf{C}}(y, z)$ it holds
$$(h \circ g) \circ f = h \circ (g \circ f). \tag{2.2}$$

We notice that in this definition $\mathrm{Ob}(\mathbf{C})$ is a collection, and not only a set, in order to avoid problems like the Russel paradox. Some examples of categories are the following.

Example 1. The category **Set** whose objects are sets and for every $x, y \in \mathrm{Ob}(\mathbf{Set})$, $Hom(x, y)_{\mathbf{Set}}$ is the set of functions from x to y.

Example 2. Consider a field $\mathbb{F}$. The category $\mathbf{Vect}_F$ is the category whose objects are $\mathbb{F}$-vector spaces, and whose morphisms are linear maps between them.

Example 3. Given a partially ordered set $(P, \leq_P)$, we can define a category, denoted again by $(P, \leq_P)$. Its objects are the elements of P and the morphisms are given by the relation $a \leq_P b$ for $a, b \in P$. In particular, we will focus our attention on the category $(\mathbb{R}, \leq)$.

Example 4. The category **Top** of topological spaces, whose objects are topological spaces, and the morphisms are continuous maps. In the same way is defined $\mathbf{Top}_*$, the category of pointed topological spaces, with point preserving continuous maps as morphisms.

Example 5. The category **Grp** of groups, whose objects are groups and the morphisms are given by homomorphisms.

We introduce functors in order to link comparable categories. We can think of functors as functions between categories. They will map the objects of one category into the other, and will do the same with the morphisms, while satisfying certain conditions about the composition of morphisms.

Definition 2 (functor)**.** Given two categories $\mathbf{C}$ and $\mathbf{D}$, a (covariant) functor F from $\mathbf{C}$ to $\mathbf{D}$, denoted by $F : \mathbf{C} \to \mathbf{D}$ is constituted by

- a function that assigns to every object $x \in \mathrm{Ob}(\mathbf{C})$ an object $F(x) \in \mathrm{Ob}(\mathbf{D})$;
- for every $x, y \in \mathrm{Ob}(\mathbf{C})$, a function that assigns to every morphism $f \in Hom_{\mathbf{C}}(x, y)$ a morphism $F(f) \in Hom_{\mathbf{D}}(F(x), F(y))$,

such that the following requirements hold:

1. for any $x \in \mathrm{Ob}(\mathbf{C})$ it holds $F(\mathrm{id}_x) = \mathrm{id}_{F(x)}$;
2. for any objects $x, y, z \in \mathrm{Ob}(\mathbf{C})$ and any $f \in Hom_{\mathbf{C}}(x, y)$, $g \in Hom_{\mathbf{C}}(y, z)$, it is
$$F(g \circ f) = F(g) \circ F(f).$$

We also say that a covariant functor $F : \mathbf{A} \to \mathbf{B}$ is a *diagram* of type $\mathbf{A}$ in the category $\mathbf{B}$. We denote the collection of diagrams of type $\mathbf{A}$ in $\mathbf{B}$ by $\mathbf{B}^{\mathbf{A}}$.

Functors allow us to map a category to another. They have also some important properties, like preserving isomorphisms and commutative diagrams.

Example 6. Consider a category $\mathbf{C}$.Fixed an object a of $\mathbf{C}$, it is defined the *hom functor* $Hom(a, -)$ from the category $\mathbf{C}$ to **Set** in the following way. The image of an object x of $\mathbf{C}$ through the functor is the set $Hom(a, x)_{\mathbf{C}}$. Every map $f : x \to y$ is sent to the map $Hom(a, f) : Hom(a, x) \to Hom(a, y)$ that assigns to every function $g \in Hom(a, x)$ the composition $f \circ g \in Hom(a, y)$.

In order to compare different functors we need the concept of natural transformation.

Definition 3 (natural transformation)**.** Consider two categories $\mathbf{C}$ and $\mathbf{D}$ and two functors $F : \mathbf{C} \to \mathbf{D}$ and $G : \mathbf{C} \to \mathbf{D}$. A natural tranformation η from $\mathbf{C}$ to $\mathbf{D}$, is given by a family of morphisms $\{\eta_x\}$

$$\eta_x : F(x) \to G(x), \text{ for any } x \in \mathrm{Ob}(\mathbf{C})$$

such that for any $x, y \in \mathrm{Ob}(\mathbf{C})$ and any $f \in Hom_{\mathbf{C}}(x, y)$, the following diagram commutes

$$\begin{array}{ccc} F(x) & \xrightarrow{F(f)} & F(y) \\ \downarrow{\scriptstyle \eta_x} & & \downarrow{\scriptstyle \eta_y} \\ G(x) & \xrightarrow{G(f)} & G(y). \end{array} \tag{2.3}$$

The transformation is a *natural isomorphism* if for any object $x \in \mathrm{Ob}\,\mathbf{C}$ the map $\eta_x : F(x) \to G(x)$ is an isomorphism in $\mathbf{D}$.

Natural transformations are useful to transform one functor into another while preserving its structure.

2.2 Metric structures on finite sets

In several applications, the starting object is a point cloud, coming from certain experimental measurements. The foundational idea of Topological Data Analysis is that the relations among the points that come from a measurement carry valuable information of the investigated phenomenon. To explicit these relations between points, we will use certain measures of dissimilarity. Their purpose is to evaluate how different, or how similar, two objects are. We will denote by $\mathbb{R}^+$ the set of non-negative real numbers.

2.2.1 Dissimilarity measures

The most primitive kind of dissimilarity that we will use are dissimilarity measures.

Definition 4 (dissimilarity measure)**.** A *dissimilarity measure* d over a set S is a function $d : S \times S \longrightarrow \mathbb{R}^+$ such that:

1. $0 \leq d(s_1, s_2)$, for all $s_1, s_2 \in S$, and $d(s, s) = d_0$, for all $s \in S$,
2. $d(s_1, s_2) = d(s_2, s_1)$, for all $s_1, s_2 \in S$.

We are essentially requiring the function d to be at least symmetric and bounded from below. We notice that the concept of dissimilarity is related to that of undirected complete weighted graph.

Definition 5 (weighted graph)**.** An undirected weighted graph is given by a triplet (V, E, w), where V is a set of vertices, E is the set of edges, that is a subset of $\{(x, y) \in V^{(2)} \mid x \neq y\}$ and $w : E \to \mathbb{R}^+$ is the weight function.

Essentially, in a weighted graph we assign a real number, or *weight*, to each edge of the graph. Then, to each dissimilarity we can associate a weighted complete graph such that for all $x, y \in V$ with $x \neq y$, the edge (x, y) belongs to E and the weight of the edge is equal to the value of the dissimilarity.

2.2.2 Metric spaces

We can impose further restriction on a dissimilarity measure in order to have a stronger structure on the point cloud.

Definition 6. Given a set S and a function $d : S \times S \to \mathbb{R}^+$, if the following conditions hold

1. $d(x, x) = 0 \quad \forall x \in S$;
2. $d(x, y) = d(y, x) \quad \forall x, y \in S$;
3. $d(x, z) \leq d(x, y) + d(y, z) \quad \forall x, y, z \in S$,

then the function d is called *pseudo-metric*. Furthermore, if it holds

- $d(x, y) = 0$ if and only if $x = y$, $\forall x, y \in S$,

then the function d is called *metric* or *distance*.

Whenever a set S is finite and d is a distance on S, we say that (S, d) is a finite metric space.

In order to be able to compare finite metric spaces, we recall the definition of Gromov-Hausdorff distance.

Given two metric spaces (X, d_X) and (Y, d_Y), a *correspondence* between them is a set $C \subseteq X \times Y$ such that $\pi_X(C) = X$ and $\pi_Y(C) = Y$, where π_X and π_Y are the canonical projections of the product space. We denote by $\mathfrak{C}(X, Y)$ the set of all correspondences between X and Y. We define the *distortion* of a correspondence C, with respect to the metrics d_X and d_Y as

$$\operatorname{dis}(C, d_X, d_Y) = \sup_{(x,y),(x',y') \in C} |d_X(x, x') - d_Y(y, y')| . \tag{2.4}$$

Definition 7 (Gromov-Hausdorff distance)**.** The *Gromov-Hausdorff distance* between (X, d_X) and (Y, d_Y) is

$$d_{GH}((X, d_X), (Y, d_Y)) = \frac{1}{2} \inf_{C \in \mathfrak{C}(X,Y)} \operatorname{dis}(C, d_X, d_Y). \tag{2.5}$$

When (X, d_X) and (Y, d_Y) are finite metric spaces, the supremum in Eq. (2.4) is actually a maximum, and the infimum in Eq. (2.5) is a minimum.

Remark 1. This definition of the Gromov-Hausdorff distance coincides with the correspondence distortion distance introduced in [97], that can be extended to spaces that are not metric. A broader family of spaces, that includes finite metric spaces, is the collection of *networks*. A *network* is given by a pair (X, ω_X), where X is a first countable topological space, and $\omega_X : X \times X \to \mathbb{R}$ is a continuous function. The collection of networks includes all finite metric spaces, but also spaces for which the distance function does not satisfy symmetry or the triangular inequality. Chowdhury and Mémoli proved in [26, Theorem 12], that the Gromov-Husdorff distance, defined using the distortion of a correspondence, is a pseudo-metric on the collection of all networks. Moreover, if we restrict to dissimilarity networks, i.e. networks such that $\omega_X(x_1, x_2) = 0$ if and only if $x_1 = x_2$, then the pseudo-distance d_{GH} between two networks is zero if and only if there is a bijection between the two networks that preserve the distance functions, see [18, Theorem 11].

2.3 Simplicial Complexes

Simplicial complexes will be our building blocks to construct topological spaces from data. For a quick reference look at [42]. Let us recall some notions of affine geometry. Consider $k+1$ points $u_0, \ldots, u_k$ of a the vector space $\mathbb{R}^d$. An affine combination of these points is a linear combination $x = \sum_{i=0}^{k} \lambda_i u_i$ such that $\sum_{i=0}^{k} \lambda_i = 1$. We say that the points $u_0, \ldots, u_k$ are *affinely independent* if and only if for every

affine combinations $x = \sum_{i=0}^{k} \lambda_i u_i$, $y = \sum_{i=0}^{k} \mu_i u_i$ it holds

$$x = y \iff \lambda_i = \mu_i \ \forall i = 0, \dots, k. \tag{2.6}$$

This is equivalent to checking whether or not the vectors $u_1 - u_0, \dots, u_k - u_0$ are linearly independent. The *convex hull* of the points $u_0, \dots, u_k$ is defined as the set:

$$\left\{ \sum_{i=0}^{k} \lambda_i u_i \mid \sum_{i=0}^{n} \lambda_i = 1, \ \lambda_i \geq 0 \ \forall i = 0, \dots, k \right\}.$$

Definition 8 (simplex)**.** A k-simplex is the convex hull of $k+1$ affinely independent points. The number k is called the *dimension* of the simplex. We will write $\sigma = [u_0, \dots, u_k]$ to denote the convex hull of the points $u_0, \dots, u_k$.

In the general terminology 0-simplices are called *vertices*, 1-simplices *edges*, 2-simplices *triangles* and so on. It is possible to notice that given $k + 1$ affinely independent points $u_0, \dots, u_k$, for every $0 \leq m \leq k$, any subset with $m+1$ points is affinely independent. Given $\sigma = [u_0, \dots, u_k]$ every simplex $\tau = [u_{i_0}, \dots, u_{i_m}]$, with $\{i_0, \dots, i_m\} \subseteq \{0, \dots, k\}$ is called *face* of the simplex σ. We will use simplices to describe topological spaces. As with Lego blocks, we want simplices to be arranged in a proper way in order to be able to obtain well defined topological spaces.

In this thesis we will consider only finite simplicial complexes, since in the application problems we consider we will have a finite number of points to build the simplices.

Definition 9 (simplicial complex)**.** A finite simplicial complex is a set K of simplices such that

1. if $\sigma \in K$ and τ is a face of σ then $\tau \in K$;
2. if $\sigma_1, \sigma_2 \in K$ then either $\sigma_1 \cap \sigma_2 = \emptyset$ or $\sigma_1 \cap \sigma_2$ is a face of both σ_1 and σ_2.

The first condition ensures us that every face of a simplex is contained in the simplicial complex, the second one that simplices can intersect only on their faces.

With the definition given so far, we are able to build simplices from samples of points in an euclidean space. In general topological data analysis problems we may not have an embedding of the objects we are studying in an Euclidean space, but rather a distance or dissimilarity between objects. Nevertheless, we want to be able to build simplicial complexes also in these cases. To do so, we use the concept of abstract simplicial complex.

Definition 10 (abstract simplicial complex)**.** An abstract simplicial complex over a set V of vertices is a collection K of subsets of V, such that for every $\alpha \in K$ and every set β with $\beta \subseteq \alpha$ then $\beta \in K$.

The simplices in this case are just sets of points, called vertices. We define the dimension of a simplex σ as its cardinality minus 1. Every subset β of a simplex α is called face of α. We can notice that in the definition of abstract simplicial complex we only required that every face of a simplex is included in the complex, like in the first point of Definition 9. It is not needed to require anything about the intersection of simplices, since given two simplices α, β their intersection is always a face of both simplices, therefore it always belongs to K.

2.3.1 Simplicial complexes from data

We will now see some of the most used techniques to construct a simplicial complex from data. The starting object will be either a set of points, or a dissimilarity space, or simply a weighted graph. Consider a metric space (X, d), given a point $x \in X$ and a non-negative real r, we denote by $B_r(x)$ the closed ball $\{y \in X \mid d(x, y) \leq r\}$.

Definition 11 (Čech complex)**.** Given a set P of points in a metric space (X, d) and a positive real number r, the Čech complex $\check{C}_r(P)$ is the set

$$\check{C}_r(P) = \left\{ \sigma = [x_0, \ldots, x_k] \mid \bigcap_{i=\{0,\ldots,k\}} B_r(x_i) \neq \emptyset \right\}. \tag{2.7}$$

Computing the Čech complex of a set of points is computationally expensive. A simpler alternative is given by the Vietoris-Rips complex

Definition 12 (Vietoris-Rips complex)**.** Given a set of points P in a metric space (X, d) and a positive real number r the Vietoris-Rips complex of P of radius r is the set

$$VR_r(P) = \{\sigma = [x_0, \ldots, x_k] \mid \max_{i,j} d(x_i, x_j) \leq 2r\} \tag{2.8}$$

Notice that in the definition of Čech and Vietoris-Rips complex we did not exploit the triangle inequality of the distance, hence a dissimilarity is sufficient to define the two kind of complexes. If the points P are included an euclidean space $(\mathbb{R}^k, d)$, the following relation holds between the two types of complexes.

Theorem 1 (Vietoris-Rips lemma)**.** *For any $r > 0$ and any $P \subset (\mathbb{R}^k, d)$ it holds*

$$\check{C}_r(P) \subseteq VR_r(P) \subseteq \check{C}_{\sqrt{2}r}(P). \tag{2.9}$$

Notice also that the computational burden to obtain a Vietoris-Rips complex is much lower than that of computing the Čech counterpart. Another example of simplicial complexes that we will use comes from graph theory. We recall that a *clique* is a graph $G = (V, E)$ such that for every two vertices $x, y \in V$ the edge (x, y) belongs to E, i.e. it is fully connected.

Definition 13 (flag complex)**.** Given a graph $G = (V, E)$, its associated *flag* (or *clique*) *complex* is a simplicial complex K whose simplices are given by the cliques of G. That is, the k-simplex $[u_0, \dots, u_k]$ belongs to K if the subgraph of G induced by the points $u_0, \dots, u_k$ is a clique.

2.4 Simplicial Homology

Simplicial homology is a topological invariant, that is it doesn't change under homeomorphisms. It constitutes the main building block of Topological Data Analysis, since it is easy to build and to compute. We will introduce the main elements of homology theory that we will use. See [78] for a reference. For the purpose of applications, homology groups are often defined using coefficients in a field. We will do the same, with particular emphasis for the field $\mathbb{Z}_2$.
Consider a simplex σ given by a set of points $u_0, \dots, u_p$. We say that two orderings of the points are equivalent if there is an even permutation that brings one into the other. This is an equivalence relation that partition the set of permutations in two equivalence classes. These two are nothing but the quotient of the symmetric group S_{p+1} with the alternating group A_{p+1}. We call orientations of σ these two equivalence classes. An oriented simplex is a simplex σ together with an orientation of σ. In other words, when we write $\sigma = [u_0, \dots, u_p]$ we are considering the simplex spanned by the points $u_0, \dots, u_p$, with the equivalence class of the ordering $(u_0, \dots, u_p)$. Therefore, every oriented simplex σ can be seen as an equivalence class $\sigma = \{[u_{\alpha(0)}, \dots, u_{\alpha(p)}] \mid \alpha \in A_{p+1}\}$.

Definition 14 (p-chain)**.** Given a simplicial complex K and a field $\mathbb{F}$, we denote by $C_p(K)$ the free vector space over the set of oriented simplices of dimension p, quotiented by the relations $\sigma + \sigma' = 0$ if $\sigma = [u_0, \dots, u_p] = [u_{\alpha(0)}, \dots, u_{\alpha(p)}] = \sigma'$ with $\alpha \in S_{p+1} \setminus A_{p+1}$. In other words, $C_p(K)$ is the vector space given by finite formal sums of oriented simplices of K

$$c = \sum_i \lambda_i \sigma_i,$$

where the λ_i are elements of $\mathbb{F}$ and the σ_i are oriented p-simplices of K, such that if σ and σ' are opposite orientations of the same simplex, then $\sigma + \sigma' = 0$. The set $C_p(K)$ is usually called the *group of oriented p-chains*, and its elements are called p-chains.

To obtain a basis of $C_p(K)$ it is sufficient to assign an orientation to each simplex σ of K, then the set of elementary formal sums $\{\sigma \mid \sigma \in K\}$ is a basis.

Definition 15 (boundary operator)**.** Consider a simplicial complex K and for a given p the two groups $C_p(K)$ and $C_{p-1}(K)$. We define the *boundary operator* as

$$\partial_p : C_p(K) \longrightarrow C_{p-1}(K), \tag{2.10}$$

such that for every $\sigma = [u_0, \dots, u_p]$

$$\partial_p(\sigma) = \sum_{i=0}^{p} (-1)^p [u_0, \dots, \hat{u}_i, \dots, u_p], \tag{2.11}$$

where with $[u_0, \dots, \hat{u}_i, \dots, u_p]$ we mean the oriented simplex obtained removing u_i from the ordered tuple $u_0, \dots, u_p$. Once defined in this way on a basis of $C_p(K)$, then it is extendend linearly on all the p-chains.

Notice that for $p < 0$ the groups $C_p(K)$ are the trivial vector space $\{0_{\mathbb{F}}\}$, therefore the homomorphisms ∂_p are the null homomorphism for each $p \leq 0$.

Theorem 2 (Fundamental Lemma of Homology)**.** *For every p it holds*

$$\partial_{p-1} \circ \partial_p = 0$$

.

Proof. We will show that for any elementary chain $\sigma \in C_p(K)$ it holds $\partial_{p-1}(\partial_p(\sigma)) = 0$. Then, the result follows from the linearity of the boundary operators. Let σ be equal to $[u_0, \dots, u_p]$, then

$$\begin{aligned} \partial_{p-1}(\partial_p(\sigma)) &= \partial_{p-1}\left(\sum_{i=0}^{p} (-1)^i [u_0, \dots, \hat{u}_i, \dots, u_p]\right) = \\ &= \sum_{\substack{j=0 \\ j \neq i}}^{p} \sum_{i=0}^{p} (-1)^i (h(i,j)(-1)^j + h(j,i)(-1)^{j-1}) [u_0, \dots, \hat{u}_i, \dots, \hat{u}_j, \dots, u_p], \end{aligned} \tag{2.12}$$

where $h(x, y)$ is equal to 1 if $x > y$ and 0 otherwise. Therefore each simplex in the formal sum Eq. (2.12) appears two times, one with coefficient -1 and one with coefficient $+1$, therefore the sum is 0. □

We say that a p-chain in $C_p(K)$ is a *cycle* if $\partial_p(c) = 0$, and that it is a *boundary* if there exists a $(p+1)$-chain b such that $\partial_{p+1}(b) = c$. In other words, we define the group of p-cycles as

$$Z_p(K) = \ker \partial_p \tag{2.13}$$

and the group of p-boundaries as

$$B_p(K) = \operatorname{im} \partial_{p+1}. \tag{2.14}$$

Because of the Fundamental Lemma of Homology we always have that $B_p(K) \subseteq Z_p(K)$. They are both subgroups of $C_p(K)$ and their quotient space is well defined.

Definition 16 (homology groups)**.** We say that the p-th *homology group* is the quotient vector space

$$H_p(K) = \frac{Z_p(K)}{B_p(K)}. \tag{2.15}$$

Its dimension is called the p-th *Betti number* of K.

From a very informal point of view, the homology groups represent a way to describe the p-dimensional holes of a shape, and the Betti numbers represent the numbers of such holes. Homology groups are topological invariants, if two simiplicial complexes are homeomorphic to the same topological space, there is an isomorphism between their homology groups. They are easy to compute and they will be our primary tool to extract topological features from a data set.

2.4.1 Functoriality of homology

So far we have seen that homology assigns to each simplicial complex a vector space. We will see that homology induces also maps between the homology groups, given by maps between the simplicial complexes. We will exploit this functorility to build persistent homology in the next Section.

Definition 17 (simplicial map)**.** A simplicial map between simplicial complexes is a function $f : K \to L$ such that:

- for every vertex $[u] \in K$, $f([u])$ is a vertex of L
- if a simplex σ is spanned by the vertices $u_0, \dots, u_p$ the image $f(\sigma)$ is the simplex of L spanned by the vertices $f([u_0]), \dots, f([u_p])$.

We can define a category using the collection of simplicial complexes, using simplicial maps as morphisms.

Definition 18 (category of simplicial complexes)**.** We define the category **Simp** of simplicial complexes whose objects are simplicial complexes and the morphisms are simplicial maps between them.

We can see that each simplicial map $f : K \to L$ induces for every p a homomorphisms $f_p^{\#}$ between the chain groups $C_p(K)$ and $C_p(L)$. We can define it on a basis of $C_p(K)$ given by the elementary chains $\{\sigma | \sigma \in K,\ dim(\sigma) = p\}$. The image of σ through $f^{\#}$ is the elementary chain $f(\sigma)$ of $C_p(L)$ if $f(\sigma)$ has dimension p and 0 otherwise.

Proposition 1. *For every $f : K \to L$ and every $p \in \mathbb{Z}$ it holds*

$$f_{p-1}^{\#} \circ \partial_p^K = \partial_p^L \circ f_p^{\#}, \tag{2.16}$$

where ∂_p^K is the boundary operator $C_p(K) \to C_{p-1}(K)$ and $\partial_p^L :_p (L) \to C_{p-1}(L)$.

Proof. It is sufficient to show that for every p-simplex $\sigma = [u_0, \dots, u_p]$ it holds $(f^{\#}_{p-1} \circ \partial^K_p)(\sigma) = (\partial^L_p \circ f^{\#}_p)(\sigma)$. In fact,

$$\begin{aligned}(\partial^L_p \circ f^{\#}_p)(\sigma) = \partial^L_p \left([f(u_0), \dots, f(u_p)]\xi(f(u_0), \dots, f(u_p))\right) = \\ = \sum_{i=0}^{p} (-1)^i [f(u_0), \dots, \hat{u}_i, \dots, f(u_p)]\xi(f(u_0), \dots, \hat{u}_i, \dots, f(u_p))\end{aligned} \tag{2.17}$$

where $\xi(f(u_0), \dots, \hat{u}_i, \dots, f(u_p)$ is equal to 1 if the points $\{f(u_j)\}_{j \neq i}$ are all different, and 0 otherwise. On the other hand, it holds

$$\begin{aligned}(f^{\#}_{p-1} \circ \partial^K_p)(\sigma) = f^{\#}_{p-1}\left(\sum_{i=0}^{p} (-1)^i [u_0, \dots, \hat{u}_i, \dots, u_p]\right) = \\ = \left(\sum_{i=0}^{p} (-1)^i [f(u_0), \dots, \hat{u}_i, \dots, f(u_p)]\right) \xi(f(u_0), \dots, f(u_p)).\end{aligned} \tag{2.18}$$

If $\xi(f(u_0), \dots, f(u_p)) = 1$, then for any $i = 0, \dots, p$ it is also $\xi(f(u_0), \dots, \hat{u}_i, \dots, f(u_p)) = 1$, and the equality holds. On the other hand, if $\xi(f(u_0), \dots, f(u_p)) = 0$ there are two cases: either three or more points of $f(u_0), \dots, f(u_p)$ are the same, or only two, namely $f(u_h)$ and $f(u_k)$ are the same. In the first case, all the numbers $\xi(f(u_0), \dots, \hat{u}_i, \dots, f(u_p)$ are 0, for any $i = 0, \dots, p$. In the second case, all the previous numbers are 0, except for $\xi(f(u_0), \dots, \hat{u}_h, \dots, f(u_p) = \xi(f(u_0), \dots, \hat{u}_k, \dots, f(u_p) = 1$. Then we must check that

$$(-1)^h [f(u_0), \dots, \hat{u}_h, \dots, f(u_p)] + (-1)^k [f(u_0), \dots, \hat{u}_k, \dots, f(u_p)] = 0. \tag{2.19}$$

But this is true, since we can notice that, assuming without loss of generality that $k > h$,

$$[f(u_0), \dots, \hat{u}_h, \dots, f(u_p)] = (-1)^{k-h+1} [f(u_0), \dots, \hat{u}_k, \dots, f(u_p)]. \tag{2.20}$$

□

In view of the last Proposition, for every $c \in Z_p(K)$ its image under $f^{\#}_p$ is an element of $Z_p(L)$, and in the same way for every boundary $b \in B_p(K)$ it holds $f^{\#}_p(b) \in B_p(L)$. Therefore, every simplicial map f induces a well-defined homomorphism

$$\begin{aligned}f^*_p : H_p(K) &\longrightarrow H_p(L) \\ [c]_K &\longmapsto [f^{\#}_p(c)]_L.\end{aligned} \tag{2.21}$$

Because of this last observation, it is possible to see that for any $p \in \mathbb{Z}$ we can define the *homology functor* H_p as the covariant functor from **Simp** $\to$ **Vect**$_{\mathbb{F}}$, that assigns to each simplicial complex K its p-th homology group $H_p(K)$, and to each simplicial map $f : K \to L$ the homomorphism $H_p(f) = f^*_p$, where $f^*_p : H_p(K) \to H_p(L)$ is the map induced by f as in Eq. (2.21).

2.5 Persistent Homology

At the moment we are able to assign to each metric space a simplicial complex, and therefore the associated homology groups. The main drawback is that we do not have a unique choice for the simplicial complex. For example, if we consider the Vietoris-Rips complex of a metric space, the resulting complex depends on the parameter r, that determines the maximal diameter of simplices in the complex. There is no a priori rule, to determine the correct value for this parameter. The idea behind persistent homology is to consider *all* the possible values for the parameter, and to exploit the functoriality of homology, to keep track of how simplicial homology changes as we let the parameter r change. At first, we will need the concept of filtration to take all the different complexes into account.

Definition 19 (filtration of simplicial complexes)**.** A filtration of simplicial complexes if a family $\mathcal{F} = \{F_i\}_{i\in\mathbb{R}}$ of simplicial complexes, indexed by a poset $(\mathbb{R}, \leq)$ such that whenever $i \leq j$ we have $F_i \subseteq F_j$.

Filtrations can be obtained in several ways, depending on how we build a simplicial complex from data. For example, given a finite metric space (X, d), we can check that the family $\{VR_r(X, d)\}_{r\in\mathbb{R}_+}$, is a set of nested simplicial complexes, since whenever we consider two real numbers $r \leq r'$, then

$$VR_r(X, d) \subseteq VR_{r'}(X, d).$$

Another way to obtain a filtration is considering a simplicial complex and a real valued function defined on it.

Definition 20 (filtering function)**.** Given a simplicial complex K, we say that a function $f : K \to \mathbb{R}$ is a *filtering function* if for every simplices $\tau, \sigma \in K$ it holds

$$\tau \subseteq \sigma \Rightarrow f(\tau) \leq f(\sigma). \tag{2.22}$$

A filtering function induces a filtration on its domain, in fact, if we define K_u as the set $\{\sigma \in K \mid f(\sigma) \leq u\}$, then the set $\{K_u\}_{u\in\mathbb{R}}$ is a filtration of simplicial complexes. Clearly, we can see a filtration as a diagram F of type $(\mathbb{R}, \leq)$ in the category **Simp** of simplicial complexes, such that for every $i \leq j$ the map $F(i \leq j) = \iota_{i,j} : F_i \to F_j$ is the inclusion of F_i into F_j. Further details on the categorification of persistent homology can be seen in [15]. We are now ready to define the persistent homology groups.

Definition 21 (persistent homology groups)**.** Given a filtration K and $u \leq v$ the p-th persistent homology group calculated at u, v is the vector space

$$PH_p(u, v) = \operatorname{im} \iota^*_{u,v} \subseteq H_p(K_v). \tag{2.23}$$

Then the p-th persistent Betti number at u, v is its dimension,

$$\beta_p(u, v) = \dim(PH_p(u, v)). \tag{2.24}$$

Persistent homology gives us information regarding how homology classes of a certain step of the filtration are preserved across the filtration. To take into account the persistent homology groups all together we will use the concept of persistence module.

Definition 22 (persistence module)**.** The p-th persistence module associated with a filtration F is the composition of the filtration with homology functor

$$PH_p(F) = H_pF = \{H_pF(i), H_pF(i \leq j)\}_{i,j\in\mathbb{R}}. \tag{2.25}$$

It is a diagram in $\mathbf{Vect}_{\mathbb{F}}$ indexed in $(\mathbb{R}, \leq)$.

In the following Section we will see some of the properties of the diagrams $\mathbf{Vect}_{\mathbb{F}}^{(\mathbb{R},\leq)}$ that will be extremely useful from the practical point of view.

2.5.1 Reminders of Algebra

We recall some of the notions of Algebra we will use to study persistence modules [62]. A *graded ring* is a ring $(R, +, \cdot)$ equipped with a direct sum decomposition $R = \bigoplus_i R_i$, where for every $i \in \mathbb{Z}$, R_i is an Abelian group of *homogeneous* elements of *degree* i, and such that the multiplication is defined so that the product between an element of degree i and one of degree j is an element of degree $i + j$. As an example, the ring of polynomials $\mathbb{F}[t]$ is a graded ring. We will make use of the concept of graded module. A *graded module* M over a graded ring R, is a module, equipped with a direct sum decomposition $M \simeq \bigoplus_i M_i$, such that the action of R on M is defined by bilinear pairings $R_i \otimes M_j \to M_{i+j}$. Notice that a graded ring R can be seen as a graded module over itself. Then, it makes sense to consider $\Sigma^\alpha R$, the upward α-shift of the module R, defined as

$$\Sigma^\alpha R = \oplus_i R_{i-\alpha}. \tag{2.26}$$

Theorem 3 (structure theorem graded modules)**.** *A graded module M over a graded principal ideal domain (PID) R decomposes uniquely into the form*

$$\left(\bigoplus_{i=1}^{n} \Sigma^{\alpha_i} R\right) \bigoplus \left(\bigoplus_{j=1}^{m} \Sigma^{a_i} R/(d_j)\right) \tag{2.27}$$

where $d_j \in R$ are homogeneous elements so that d_j divides d_{j+1}, $\alpha_i, b_i \in \mathbb{Z}$.

We will use this theorem to obtain a unique decomposition of persistence modules. Here we will focus our attention on tame persistence modules, i.e. persistent modules for which only a finite number of changes in homology can occur.

Definition 23. Consider a diagram $F \in \mathbf{Vect}_{\mathbb{F}}^{(\mathbb{R},\leq)}$. We say that F is constant on an interval $I \subseteq \mathbb{R}$ if for all intervals $[a,b] \subseteq I$ the map $F(a \leq b)$ is an isomorphism. We say that a is a *regular value* for F if there is an open interval I, containing a, such that F is constant on I. We say that a is a *critical value* if it doesn't exist any such interval. The diagram F is *tame* if it has only a finite number of critical values.

We will compare the notion of tame diagram with that of finite type diagram.

Definition 24. Consider an interval $I \subseteq \mathbb{R}$. We define a diagram $\chi_I \in \mathbf{Vect}_{\mathbb{F}}^{(\mathbb{R},\leq)}$ as

$$\chi(a) = \begin{cases} \mathbb{F} & \text{if } a \in I, \\ 0 & \text{otherwise,} \end{cases} \qquad \chi(a \leq b) = \begin{cases} Id_{\mathbb{F}} & \text{if} a, b \in I, 0 \quad \text{otherwise.} \end{cases} \tag{2.28}$$

The diagram F has *finite type* if there exist N diagrams $\chi_{I_1}, \ldots, \chi_{I_N}$ such that $F \simeq \bigoplus_{j=1}^{N} \chi_{I_j}$.

In the case of diagrams indexed in $(\mathbb{R}, \leq)$ the two notions are equivalent.

Theorem 4 ([15])**.** *A diagram in* $\mathbf{Vect}_{\mathbb{F}}^{(\mathbb{R},\leq)}$ *is of finite type if and only if it is tame.*

It is possible to see that a finite type diagram $F \simeq \bigoplus_{j=0}^{N} \chi_{I_j}$ in $\mathbf{Vect}_{\mathbb{F}}^{(\mathbb{R},\leq)}$, can be seen as a sequence of vector spaces and linear maps. Suppose that the intervals $I_0, \ldots, I_N$ have endpoints $a_0, b_0, \ldots, a_N, b_N$. Here the endpoints can be taken from the extended line $\bar{\mathbb{R}}$. If $I_j = (-\infty, x]$ we set $a_j = -\infty$, $b_j = x$ and if $I_j = [x, +\infty)$ we set $a_j = x$, $b_j = +\infty$. We can reorder all these points into a sequence without repetition $c_0, \ldots, c_L$. Then we can define the vector space $M_i = F(c_i)$ for every $i = 0, \ldots, L$, and the linear maps $\phi_i : M_i \to M_{i+1}$ as $\phi_i = F(c_i \leq c_{i+1})$. Then the functor F is completely identified by the sequence $\{M_i, \phi_i\}_{i=0}^{L}$. As pointed out in [99], we can define a correspondence between persistence modules ad graded modules over $\mathbb{F}[t]$. Suppose to have a persistence module $\mathcal{M} = \{M_i, \phi_i\}_{i=0}^{L}$. Consider $\mathbb{F}[t]$ with the standard grading and define the graded module over $\mathbb{F}[t]$

$$\alpha(\mathcal{M}) = \bigoplus_{i=0}^{L} M_i \tag{2.29}$$

where the $\mathbb{F}$-module structure is given by the direct sum on the homogeneous components (the elements of each M_i have degree i), and the action of t is given by

$$t \cdot (m_0, m1, \ldots, m_L) = (0, \phi_0(m_0), \phi_1(m_1), \ldots, \phi_{L-1}(m_{L-1})). \tag{2.30}$$

There is an equivalence of categories between the category of persistence modules of finite type over $\mathbb{F}$ and the category of finitely generated non-negatively graded modules over $\mathbb{F}[t]$. For these modules it holds the structure theorem previously stated, and the following Krull-Schmidt theorem is satisfied:

Theorem 5. *If $F \simeq \bigoplus_{j=1}^{N} \chi_{I_j}$ and $F \simeq \bigoplus_{j=1}^{M} \chi_{I'_j}$, then $M = N$ and the intervals $I_1, \ldots, I_N$ and $I'_1, \ldots, I'_M$ are the same up to reordering.*

Because of this theorem we can identify each tame persistence modules with the multiset of intervals $\{I_j\}_{j=1}^{n}$ given by its unique decomposition. This multiset is called *barcode* of F. An equivalent notion to the barcode is that of persistence diagram, and it can be obtained directly from the persistent Betti numbers.

Definition 25 (persistence diagram)**.** Consider the extended plane $\bar{\mathbb{R}}^2$. For a given $\varepsilon > 0$ define the quantity

$$M^{\varepsilon}_{u,v} = \left[(\beta_p(u+\varepsilon, v-\varepsilon) - \beta_p(u+\varepsilon, v+\varepsilon)) - (\beta_p(u-\varepsilon, v-\varepsilon) - \beta_p(u-\varepsilon, v+\varepsilon))\right].$$

Then, the multiplicity of a point (u, v) with $u \leq v < \infty$ is

$$\mu_p(u, v) = \lim_{\varepsilon \to 0^+} M^{\varepsilon}_{u,v} \tag{2.31}$$

For a point of type (u, ∞), its multiplicity is defined as

$$\mu_p(u, v) = \lim_{\varepsilon \to 0^+} \left[\beta_p(u+\varepsilon, 1/\varepsilon) - \beta_p(u-\varepsilon, 1/\varepsilon)\right]. \tag{2.32}$$

The *persistence diagram* is the multiset of points of the extended plane with multiplicity greater than 0, each of them counted with its multiplicity.

2.5.2 Metrics between persistence diagrams and the Stability Theorem

We will devote this Section to introduce metrics between persistence diagrams, to make comparison between them. In this section we will focus our attention only on $(\mathbb{R}, \leq)$-indexed diagrams. Given a positive real number ε, it is possible to define a translation functor $T_\varepsilon : (\mathbb{R}, \leq) \longrightarrow (\mathbb{R}, \leq)$, with $T_\varepsilon(x) = x + \varepsilon$ and a natural transformation $\nu_\varepsilon : \mathrm{Id}_{(\mathbb{R},\leq)} \Longrightarrow T_\varepsilon$, with $\nu_\varepsilon(x) : x \longrightarrow x + \varepsilon$ defined as $x \leq x + \varepsilon$.

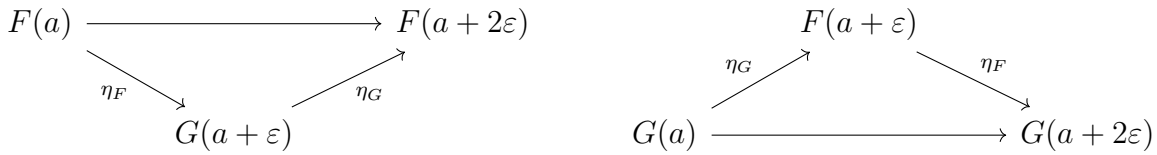
Definition 26 (ε-interleaving)**.** Two $(\mathbb{R}, \leq)$ indexed functors F, G are said to be ε-interleaved if there exist two natural transformation $\eta_F : F \Longrightarrow GT_\varepsilon$ and $\eta_G : G \Longrightarrow FT_\varepsilon$ such that

$$(\eta_G T_\varepsilon)\eta_F = F\nu_{2\varepsilon} \quad \text{and} \quad (\eta_F T_\varepsilon)\eta_G = G\nu_{2\varepsilon}. \tag{2.33}$$

The existence of the natural transformations $\eta_F,\ \eta_G$ implies the commutativity of the following diagrams

$$\begin{array}{ccc} F(a) & \longrightarrow & F(b) \\ & {\scriptstyle \eta_F(a)} \searrow \quad \swarrow {\scriptstyle \eta_F(b)} & \\ & G(a+\varepsilon) \longrightarrow G(b+\varepsilon) & \end{array} \qquad \begin{array}{ccc} F(a+\varepsilon) & \longrightarrow & F(b+\varepsilon) \\ & {\scriptstyle \eta_G(a)} \nearrow \quad \searrow {\scriptstyle \eta_G(b)} & \\ & G(a) \longrightarrow G(b) & \end{array}$$

and Eq. (2.33) can be summarised with the diagrams



Definition 27 (interleaving distance)**.** The *interleaving distance* d_I between two $(\mathbb{R}, \leq)$-indexed diagrams F and G is defined as

$$d_I(F, G) = \inf \{\varepsilon \geq 0 \mid F \text{ and } G \text{ are } \varepsilon\text{-interleaved}\}. \tag{2.34}$$

If there is no ε-interleaving for any $\varepsilon \geq 0$ we set $d_I(F, G) = \infty$.

Notice that $d_I(F, G) = 0$ doesn't always mean that F and G are 0-interleaved. Therefore, it can happen that $d_I(F, G) = 0$ even if $F \neq G$. It is possible to prove that d_I is an extended pseudo-metric on the collection of $(\mathbb{R}, \leq)$-indexed diagrams in $\mathbf{D}$. Let us denote this space by $\mathbf{D}^{\mathbb{R},\leq}$. We can define on it an equivalence relation: we say that F is equivalent to G ($F \sim G$) if and only if $d(F, G) = 0$. We can see that d_I is an extended metric on the quotient $\mathbf{D}^{(\mathbb{R},\leq)}/\sim$.

It is possible to see that the interleaving distance between persistence modules is in fact equivalent to the bottleneck distance between *finite* persistence diagrams. Let us call $\Delta^* = \{(x, y) \in \bar{\mathbb{R}}^2 \mid x \leq y\}$. We recall that a finite persistent diagram is a finite multiset of points $\{(a_i, b_i)\}_{i \in I}$, with $a_i < b_i \leq \infty$ augmented with the diagonal $\{(x, y) \in \mathbb{R}^2 | x = y\}$, counted with infinite multiplicity. Therefore, every persistent diagram is a multiset of elements of Δ^*. We start defining an extended pseudo-distance on this set. Given two points (x, y) and (x', y') in Δ^*, we define:

$$d_\infty((x, y), (x', y')) = \min\left(\max(|x - x'|, |y - y'|), \max\left(\frac{y - x}{2}, \frac{y' - x'}{2}\right)\right). \tag{2.35}$$

When computing the subtractions we use the convention that $\infty - \infty = 0$ and $\infty - r = \infty$, for every $r \in \mathbb{R}$.

Definition 28 (bottleneck (or matching) distance)**.** Given two finite persistent diagram D_1 and D_2, the *bottleneck* distance between them is

$$d_B(D_1, D_2) = \inf_{\gamma} \sup_{(x,y) \in D_1} d_\infty((x, y), \gamma(x, y)) \tag{2.36}$$

where γ ranges over all the bijections between D_1 and D_2.

This distance was first used (under the name of matching distance) with reduced size functions, equivalent to 0-th persistent Betti numbers, in [31] and than extended to general persistence diagrams in [28].

Theorem 6 ([15]). *Let $\mathcal{B}$ be the collection of finite barcodes and d_B the bottleneck distance between them. There is an isometric embedding of metric spaces:*

$$(\mathcal{B}, d_B) \hookrightarrow (\mathbf{Vect}_{\mathbb{F}}^{(\mathbb{R},\leq)}, d_I). \tag{2.37}$$

From the computational point of view, it is easier to consider the definition of bottleneck distance. On the other hand, once it is proved that on finite persistent diagrams the bottleneck distance and the interleaving distance coincide, it is useful to focus our attention on the interleaving distance in order to easily prove the so called *Stability Theorem*.

Theorem 7 ([15]). *Let $F, G \in \mathbf{D}^{()\mathbb{R},\leq}$ and consider a functor $J : \mathbf{D} \to \mathbf{E}$. If F and G are ε-interleaved, also JF and JG are ε-interleaved:*

$$d_I(JF, JG) \leq d_I(F, G). \tag{2.38}$$

This theorem allows us to prove easily a stability theorem to the foundations of Topological Data Analysis.

Theorem 8 (stability theorem - filtering functions). *Given two filtering functions $f, g : X \to \mathbb{R}$, for any $k \in \mathbb{Z}$ it holds*

$$d_I(H_k f, H_k g) \leq \|f - g\|_\infty. \tag{2.39}$$

Theorem 9 (stability theorem - metric structure). *Given two metric spaces (X, d_X) and (Y, d_Y), for any $k \in \mathbb{Z}$, it holds*

$$d_I(H_k VR(X), H_k VR(Y)) \leq d_{GH}((X, d_X), (Y, d_Y)). \tag{2.40}$$

The stability theorem ensures us that small changes in the data induce small changes in the associated perssitence diagrams, making persistent homology a reliable tool in presence of noise.

Chapter 3

A persistent point of view on weak similarity of finite semi-metric spaces

3.1 Introduction

This chapter is an extended version of a joint work with Ulderico Fugacci, Facundo Mémoli and Francesco Vaccarino, which resulted in the article [52] As we have seen in the previous Chapter one way to endow a data set with a notion of *shape* is by defining a metric structure on it. We have seen that some of the useful construction in topological data analysis can be obtained with points embedded in an euclidean space, like with the Čech filtration. Some other construction can arise with more general metric spaces, or even with spaces that do not satisfy the triangular inequality. In many application problems in fact the data is encoded in just an undirected weighted network. The methodologies introduced so far rely heavily on the actual values attained by the distance function, even though we can see that what really matters is the order in which the simplices appear in the filtration. This is why we would like to introduce an approach based on Topological Data Analysis that is invariant under "non-linear rescalings" of the metric space. In many problems, in fact, we may want results that are not changed by the scale at which we measure the objects of study.

Suppose that the object of study are physical quantities, then we may be interested in obtaining results that are not changed by different choices of the system used for measurements. It also frequent to transform data to have a better visualization, like with the log-scale transformation, but the phenomena under investigation is still the same.

Sometimes, it can happen that the observations undergo transformations that are not linear, but that preserve the structure given by the ordering of distance between pairs of points. For example, in [49], Giusti er al. such an approach is

utilized in order to study data from neural activity and connectivity, where the actual magnitude of the measurements conducted may drive to misleading results, since these quantities depend profoundly on physiological features that can be very different across the individuals in a study.

Moreover, in recent years, the analysis of ordinal data has emerged as a new field of data science. As pointed out by Kleindessner et al. [66], in certain applications the actual value of a dissimilarity between two objects is not reliable or actually informative for the problem. This is the case when the values of dissimilarity are for example estimated by humans or the measurements of the dissimilarities are just a proxy for the unknown phenomenon under study (e.g [49]).

In the last chapter of this thesis, regarding set equivariant operators [22], an interesting family of operators is that of the so called *change of units*, that are nothing but functions that transform a data set via a *rescaling* of the observations.

Moreover, this kind of ideas are relevant also from the point of view of persistent homology. For example in [21] the authors propose the so called *shift-invariant bottleneck distance* in order to measure the difference between persistence diagrams that are rescaled under a logarithmic function.

In this chapter, we study a definition of weak similarity already present in literature in [38, 65], that allows us to consider equivalent two finite metric spaces if it is possible to obtain one of the distance functions as a composition of the other distance with a strictly increasing real valued function.

This kind of problem is not new to the literature. In the beginning of the 90s, Ganyushkin and Tsvirkunov [46], investigated the problem of classifying finite metric spaces under a certain notion of isomorphism that takes into account only the ordering between pairs of points given by the distance function. In this Chapter, we use some of the concepts introduced by them in order to define the canonicalization of a finite metric space: a procedure which associates to each equivalence class of weak similarity a unique representative. We will show that two spaces are weakly similar if and only if their canonicalization are isometric, in the classical sense.

Even with such a simplification, the problem of determining the classical isometry between two spaces it is still computationally expensive. In order to simplify the problem of ascertaining whether two spaces are weakly similar, we introduce several complete and incomplete computable invariants for weak similarity.

We will focus our attention at first on curvature sets introduced by Gromov [53], that can be seen as sets of matrices obtained by considering only the subspaces of a given metric space of fixed cardinality. We will prove that, in the finite case, the only curvature sets that carry valuable information are the ones of order at most equal to the number of points of the space.

We focus on this construction since it is intuitive to see that it is related to the simplices of the Vietoris-Rips filtration associated with a metric space. We identify a certain categorification of the notion of weak similarity, and we see how it can be used as an invariant for weak similarity. Thanks to this property it will be possible

to use persistent homology as an incomplete invariant for weak similarity.

Furthermore, we introduce a *weak Gromov-Hausdorff dissimilarity* function (whose construction is based on the concept of Gromov-Hausdorff distance) which measures how much two spaces have to be modified in order to be weakly similar. We prove that this dissimilarity vanishes if and only if there is a weak similarity of finite semi-metric spaces. In the same fashion of the weak Gromov-Hausdorff dissimilarity, we define a dissimilarity between persistence modules from the point of view of weak similarity. As a main result, we prove that these two dissimilarities satisfy a stability theorem equivalent to the one between the interleaving distance of persistence modules and the Gromov-Hausdorff distance [24]. The latter dissimilarity is based on the bottleneck distance between persistence diagrams. This has the advantage of being easier to compute than the Gromov-Hausdorff distance and we propose a way to approximate this dissimilarity using the framework introduced in [72] to differentiate barcode-valued functions.

3.2 Weakly similar finite semi-metric spaces

We recall that the symbol $\mathbb{R}^+$ denotes for us the set $\{x \in \mathbb{R} \mid x \geq 0\}$.

Definition 29 (semi-metric and metric spaces)**.** A *semi-metric space* is a pair (X, d_X), where X is a set and $d_X : X \times X \to \mathbb{R}^+$ is a function such that

- $d_X(x, y) = 0 \iff x = y$;
- $d_X(x, y) = d_X(y, x)$.

(X, d_X) is called a *finite* semi-metric space if the set X is a finite. We will denote by FSS the collection of finite semi-metric spaces. If a semi-metric space (X, d_X) satisfies the condition

- $d_X(x, y) \leq d_X(x, z) + d_X(z, y) \quad \forall x, y, z \in X$,

it is called a *metric* space. The collection of finite metric spaces will be denoted with FMS.

Along this chapter except otherwise stated the semi-metric spaces are supposed to be finite. For the sake of simplicity, we will drop the word finite, calling them semi-metric spaces. Furthermore, when there is no confusion on the semi-metric d_X defined on a space X, we will simply denote the semi-metric space as X. The main concept investigated in this chapter is that of weak similarity, defined in [38].

Definition 30 (weak similarity)**.** Let us consider two semi-metric spaces (X, d_X) and (Y, d_Y). A bijection $\varphi : X \longrightarrow Y$ is a *weak similarity* if there exists a strictly increasing function $\psi : \mathbb{R}^+ \longrightarrow \mathbb{R}^+$ such that, for all $x_1, x_2 \in X$,

$$\psi(d_X(x_1, x_2)) = d_Y(\varphi(x_1), \varphi(x_2)). \tag{3.1}$$

If there is a weak similarity between (X, d_X) and (Y, d_Y), we say that the spaces are *weakly similar* and we will write

$$(X, d_X) \cong^w (Y, d_Y).$$

Theorem 10. *The relation of weak similarity is an equivalence relation.*

Proof. We check the three properties of equivalence relations.

- Reflexivity: using the identity, $X \cong^w X$ for all finite semi-metric spaces X.
- Symmetry: if $X \cong^w Y$ we have a bijection $\varphi : X \longrightarrow Y$ and a strictly increasing function $\psi : \mathbb{R}^+ \longrightarrow \mathbb{R}^+$ with $\psi(d_X(x_1, x_2)) = d_Y(\varphi(x_1), \varphi(x_2))$ for all $x_1, x_2 \in X$. We have that ψ is an invertible function since it is strictly monotone. For each $y_1, y_2 \in Y$ consider $x_1, x_2 \in X$ with $y_i = \varphi(x_i)$, where $i = 1,2$. Then,

  $$d_X(\varphi^{-1}(y_1), \varphi^{-1}(y_2)) = \psi^{-1}(d_Y(y_1, y_2)).$$

  and so, by definition, $Y \cong^w X$.
- Transitivity: if $X \cong^w Y$ and $Y \cong^w Z$ consider the functions $\varphi_1, \varphi_2, \psi_1, \psi_2$ such that

  $$\begin{aligned} \psi_1(d_X(x_1, x_2)) &= d_Y(\varphi_1(x_1), \varphi_1(x_2)) \quad \forall x_1, x_2 \in X, \\ \psi_2(d_Y(y_1, y_2)) &= d_Z(\varphi_2(y_1), \varphi_2(y_2)) \quad \forall y_1, y_2 \in Y. \end{aligned}$$

  Then,

  $$\psi_1(d_X(x_1, x_2)) = d_Y(\varphi_1(x_1), \varphi_1(x_2)) = \psi_2^{-1}(d_Z(\varphi_2(\varphi_1(x_1)), \varphi_2(\varphi_1(x_2))))$$

  hence, considering the functions $\varphi_2 \circ \varphi_1$ and $\psi_2 \circ \psi_1$, we have $X \cong^w Z$.

□

Ganyushkin and Tsvirkunov [46] introduced a notion of isomorphism for finite semi-metric spaces which we will compare to weak similarity below.

Definition 31 (isomorphism - [46])**.** We say that two semi-metric spaces (X, d_X) and (Y, d_Y) are *isomorphic* if there is a bijection $\varphi : X \longrightarrow Y$ such that for all $x_1, x_2, x_1', x_2' \in X$ we have

$$d_X(x_1, x_2) = d_X(x_1', x_2') \Rightarrow d_Y(\varphi(x_1), \varphi(x_2)) = d_Y(\varphi(x_1'), \varphi(x_2')) \tag{3.2}$$

$$d_X(x_1, x_2) < d_X(x_1', x_2') \Rightarrow d_Y(\varphi(x_1), \varphi(x_2)) < d_Y(\varphi(x_1'), \varphi(x_2')). \tag{3.3}$$

If (X, d_X) and (Y, d_Y) are isomorphic we will write $(X, d_X) \simeq (Y, d_Y)$.

It turns out than the notions of weak similarity and isomorphism of semi-metric spaces are related.

Theorem 11. *Two finite semi-metric spaces (X, d_X) and (Y, d_Y) are weakly similar if and only if they are isomorphic.*

Proof. Assume that $X \cong^w Y$. Then, we have two functions $\varphi : X \to Y$ and $\psi : \mathbb{R}_+ \to \mathbb{R}_+$ such that, for all $x_1, x_2, x_1', x_2' \in X$,

$$\psi(d_X(x_1, x_2)) = d_Y(\varphi(x_1), \varphi(x_2)), \tag{3.4}$$
$$\psi(d_X(x_1', x_2')) = d_Y(\varphi(x_1'), \varphi(x_2')). \tag{3.5}$$

Then, if $d_X(x_1, x_2) = d_X(x_1', x_2')$, we have

$$d_Y(\varphi(x_1), \varphi(x_2)) = \psi(d_X(x_1, x_2)) = \psi(d_X(x_1', x_2')) = d_Y(\varphi(x_1'), \varphi(x_2')). \tag{3.6}$$

If $d_X(x_1, x_2) < d_X(x_1', x_2')$, since ψ is a strictly increasing function,

$$d_Y(\varphi(x_1), \varphi(x_2)) = \psi(d_X(x_1, x_2)) < \psi(d_X(x_1', x_2')) = d_Y(\varphi(x_1'), \varphi(x_2')). \tag{3.7}$$

Therefore, $X \cong^w Y \Rightarrow X \simeq Y$.
On the other hand, assume $X \simeq Y$ and consider the bijection φ given by Definition 31. It is possible to order all the pairs $(x_i, x_j) \in X \times X$ so that

$$d_X(x_{i_1}, x_{j_1}) \leq d_X(x_{i_2}, x_{j_2}) \leq \cdots \leq d_X(x_{i_{n^2}}, x_{j_{n^2}}), \tag{3.8}$$

where n is the number of points of X and Y. Because of implications (3.2) and (3.3), we have that

$$d_Y(\varphi(x_{i_1}), \varphi(x_{j_1})) \leq d_Y(\varphi(x_{i_2}), \varphi(x_{j_2})) \leq \cdots \leq d_Y(\varphi(x_{i_k}), \varphi(x_{j_k})), \tag{3.9}$$

hence, we can define an increasing function $\psi : \mathbb{R}^+ \longrightarrow \mathbb{R}^+$ with

$$\psi(d_X(x_{i_r}, x_{j_r})) = d_Y(\varphi(x_{i_r}), \varphi(x_{j_r})) \quad \forall (x_{i_r}, x_{j_r}) \in X \times X \tag{3.10}$$

and then $X \simeq Y \Rightarrow X \cong^w Y$. □

We have seen that the two concepts are the same, but weak similarity explicitly shows the non-linear rescaling that has to be performed to obtain one space from the other.

Another similar concept is that of *ordinal spaces*, introduced by Keller and Petrov [65] as a mean to study ordinal data. Consider the set of relations {‘ < ’,‘ = ’,‘ > ’} on the real numbers. We denote by −‘ < ’ the opposite of the relation ‘ < ’ and we consider equivalent the relations −‘ < ’ and ‘ > ’. In the same way we write −‘ > ’ = ‘ < ’ and −‘ = ’ = ‘ = ’.

Definition 32. An *ordinal space* (X, δ_X) is composed by a nonempty set X and a map $\delta_X : X \times X \times X \times X \to \{\text{`} < \text{'}, \text{`} = \text{'}, \text{`} > \text{'}\}$ such that for any $x, y, u, v, z, w \in X$ the following conditions are satisfied:

1. $\delta_X(x, y, x, y) = \text{`} = \text{'}$;
2. $\delta_X(x, y, z, w) = \delta_X(y, x, z, w) = \delta_X(x, y, w, z)$;
3. $\delta_X(x, y, z, w) = -\delta_X(z, w, x, y)$;
4. $\delta_X(x, y, u, v) = \delta_X(u, v, z, w) = \text{`} = \text{'} \Rightarrow \delta_X(x, y, z, w) = \text{`} = \text{'}$;
5. $\delta_X(x, y, u, v) = \text{`} < \text{'}$ and $\delta_X(u, v, z, w) \in \{\text{`} < \text{'}, \text{`} = \text{'}\} \Rightarrow \delta_X(x, y, z, w) = \text{`} < \text{'}$;
6. $\delta_X(x, y, u, v) \in \{\text{`} < \text{'}, \text{`} = \text{'}\}$ and $\delta_X(u, v, z, w) = \text{`} < \text{'} \Rightarrow$ $\delta_X(x, y, z, w) = \text{`} < \text{'}$;
7. $\delta_X(x, x, z, w) = \text{`} < \text{'}$ if $z \neq w$ and $\delta_X(x, x, z, w) = \text{`} = \text{'}$ if $z = w$.

Definition 33. Two ordinal spaces (X, δ_X) and (Y, δ_Y) are said to be *isomorphic* if there exists a bijection $\phi : X \to Y$ such that

$$\delta_X(x, y, w, z) = \delta_Y(\phi(x), \phi(y), \phi(w), \phi(z)), \tag{3.11}$$

for all $x, y, w, z \in X$.

One can associate an ordinal space to each semi-metric space.

Example 7 ([65])**.** Given a semi-metric space (X, d_X), it can be seen as an ordinal space (X, δ_X). Define δ_X for all $x, y, w, z \in X$ as

$$\delta_X(x, y, w, z) = \begin{cases} \text{`} < \text{'} & \text{if } d_X(x, y) < d_X(w, z), \\ \text{`} = \text{'} & \text{if } d_X(x, y) = d_X(w, z), \\ \text{`} > \text{'} & \text{if } d_X(x, y) > d_X(w, z). \end{cases} \tag{3.12}$$

This ordinal space will be called the *ordinal type* of (X, d_X)

Ordinal spaces and their isomorphisms are related to semi-metric spaces and weak similarity thanks to the following proposition:

Proposition 2 ([65])**.** *Let (X, d_X) and (Y, d_Y) be semi-metric spaces. The ordinal types (X, δ_X) and (Y, δ_Y) are isomorphic if and only if (X, d_X) and (Y, d_Y) are weakly similar.*

To simplify the study of weak similarities, we want to associate to each equivalence class of weak similarity a good representative.

Definition 34 (distance set)**.** Given a semi-metric space (X, d_X), we define its *distance set* $\mathrm{D}(X, d_X)$ as the set of all pairwise distances between points of X.

$$\mathrm{D}(X, d_X) := \{d_X(x_1, x_2) \mid x_1, x_2 \in X\}\,. \tag{3.13}$$

When there is no ambiguity for the metric d_X defined on the space X, we will simply write $\mathrm{D}(X)$.

Lemma 1. *Two weakly similar semi-metric spaces have distance sets of the same cardinality.*

Proof. If $X \cong^w Y$, there is a strictly increasing function ψ such that

$$\psi(\mathrm{D}(X)) := \{\psi(l) \mid l \in \mathrm{D}(X)\} = \mathrm{D}(Y). \tag{3.14}$$

So, since $\psi|_{\mathrm{D}(X)}$ is injective by definition and surjective on $\mathrm{D}(Y)$, then $|\mathrm{D}(X)| = |\mathrm{D}(Y)|$, where $|A|$ denotes the cardinality of the space A. □

Remark 2. The reciprocal statement does not hold. Two spaces can have the same distance set but not be weakly similar. For example, Boutin and Kemper in [13] study the problem of recontruction of a semi-metric space given the distribution of distances between points.

Example 8. The two semi-metric spaces in Fig. 3.1, $X = \{a, b, c\}$ with $d_X(a, c) = d_X(b, c) = 6$, $d_X(a, b) = 5$ and $Y = \{d, e, f\}$ with $d_Y(d, f) = d_X(e, f) = 5$, $d_Y(d, e) = 6$ have the same distance set, but they are not weakly similar.

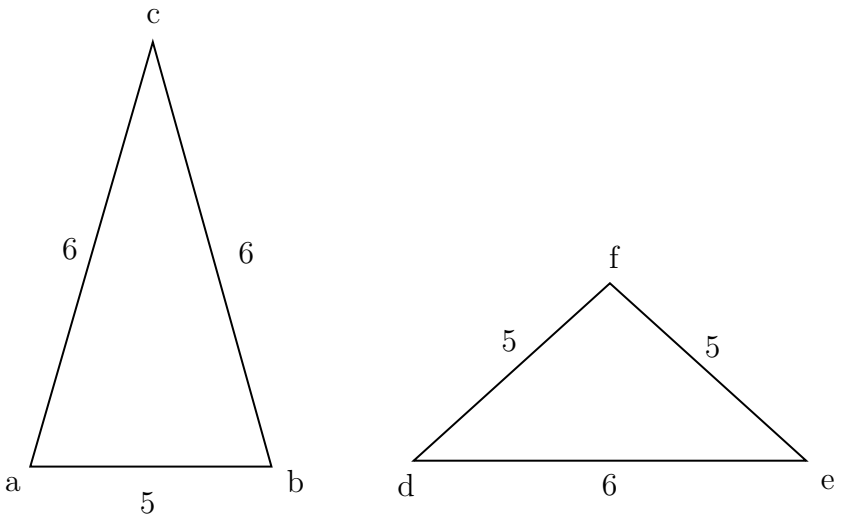


Figure 3.1: Semi-metric spaces that are not weakly similar.

Definition 35 (natural-valued semi-metric space)**.** We say that the metric d defined on the space X is *natural-valued* if $\mathrm{D}(X) \subset \mathbb{N}$.

Definition 36 (dense distance set)**.** Given a semi-metric space (X, d_X) with natural-valued metric d_X we say that its distance set is *dense* if, a part from zero, it is a list of consecutive natural numbers,

$$\mathrm{D}(X) = \{0, a+1, a+2, \ldots, a+|\,\mathrm{D}(X)|\}\,. \tag{3.15}$$

Definition 37 (canonical, [46])**.** A semi-metric space of cardinality n is called *k-canonical* if it has a natural valued and dense distance set of the form

$$D_{n,k} := \left\{2\binom{n}{2}, 2\binom{n}{2}-1, \ldots, 2\binom{n}{2}-k+1, 0\right\}. \tag{3.16}$$

It is said to be *canonical* if it is k-canonical for some k. We will say that a semi-metric space $\mathcal{C}$ is *canonical* for the semi-metric space X if they are weakly similar and $\mathcal{C}$ is canonical.

Remark 3. Every canonical space is metric. In fact, given a canonical space (X, d_X), for every $x, y, z \in X$ it holds

$$d_X(x,y) \leq \max_{u,v\in X} d_X(u,v) = 2\binom{n}{2} \leq d_X(x,z) + d_X(z,y),$$

since for every $w, z \in X$ it must be $d_X(w,z) \geq \binom{n}{2}$ by definition of canonical space.

Canonical semi-metric spaces are interesting because for them the notion of weak similarity is equivalent to that of isometry, as we show in the next lemma.

Lemma 2. *Let $\mathcal{C}_1$ and $\mathcal{C}_2$ be canonical semi-metric spaces. Then, $\mathcal{C}_1$ is weakly similar to $\mathcal{C}_2$ if and only if $\mathcal{C}_1$ is isometric to $\mathcal{C}_2$.*

Proof. If $\mathcal{C}_1$ is isometric to $\mathcal{C}_2$, then $\mathcal{C}_1$ is weakly similar to $\mathcal{C}_2$. On the other hand assume $\mathcal{C}_1 \cong^w \mathcal{C}_2$. By Lemma 1, we know that the two spaces have distance sets of the same cardinality. On the other hand, the distance set of a canonical space is defined by its cardinality, so

$$\begin{aligned}\mathrm{D}(\mathcal{C}_1) &= \left\{2\binom{n}{2}, 2\binom{n}{2}-1, \ldots, 2\binom{n}{2}-|\,\mathrm{D}(\mathcal{C}_1)|+1, 0\right\} = \\ &= \left\{2\binom{n}{2}, 2\binom{n}{2}-1, \ldots, 2\binom{n}{2}-|\,\mathrm{D}(\mathcal{C}_2)|+1, 0\right\} = \mathrm{D}(\mathcal{C}_2),\end{aligned} \tag{3.17}$$

where n is the cardinality of the spaces $\mathcal{C}_1$ and $\mathcal{C}_2$. By hypothesis, we have a bijection $\varphi : \mathcal{C}_1 \longrightarrow \mathcal{C}_2$ and a strictly increasing function $\psi : \mathbb{R}^+ \longrightarrow \mathbb{R}^+$ such that

$\psi(\mathrm{D}(\mathcal{C}_1)) = \mathrm{D}(\mathcal{C}_2)$. Since the two distance sets are equal, such a function can only be the identity and, therefore,

$$\psi(d_{\mathcal{C}_1}(x_1, x_2)) = d_{\mathcal{C}_1}(x_1, x_2) = d_{\mathcal{C}_2}(\varphi(x_1), \varphi(x_2)) \tag{3.18}$$

and the two spaces are isometric. □

In Proposition 2 of [46], the authors proved the following useful result.

Theorem 12. *Each semi-metric space is isomorphic to a canonical semi-metric space.*

We can now define a map called *canonicalization* which assigns to each finite semi-metric space X a canonical semi-metric space $\mathcal{C}_X$.

Definition 38 (canonicalization)**.** Let X be a semi-metric space of cardinality n and distance set $\mathrm{D}(X) = \{0, a_1, a_2, \ldots, a_k \mid a_i < a_j \text{ if } i < j\}$. Define $\psi : \mathbb{R}^+ \longrightarrow \mathbb{N}$ as

$$\psi(a_i) = 2\binom{n}{2} - k + i. \tag{3.19}$$

The *canonicalization* of X is the space $(\mathcal{C}_X, d_{\mathcal{C}_X})$ with:

- $\mathcal{C}_X = X$
- $d_{\mathcal{C}_X}(x_1, x_2) = \begin{cases} \psi(d_X(x_1, x_2)) & \text{if } x_1 \neq x_2 \\ 0 & \text{if } x_1 = x_2. \end{cases}$

Canonical semi-metric spaces are important because, as we will see with the following corollary, they allow us to associate a unique representative to each weak similarity equivalence class.

Corollary 1 (uniqueness of canonical representations)**.** *Let X be a semi-metric space with $\mathcal{C}$ and $\mathcal{C}'$ canonical for X. Then, $\mathcal{C}$ and $\mathcal{C}'$ are isometric, that is, there is a unique semi-metric space canonical for X up to isometry.*

Proof. By the transitivity of weak similarity, we have that $\mathcal{C} \simeq X \simeq \mathcal{C}'$ and, by Lemma 2, they are isometric. □

For a given semi-metric space X, Corollary 1 allows us to identify a canonical representative of a X in the form of its canonicalization, as introduced in Definition 38.

In this way, we can now reformulate the problem of weak similarity to that of classical isometry between canonical spaces.

Theorem 13 (weak similarity is equivalent to isometry of the canonicalizations)**.** *Two semi-metric spaces are weakly similar if and only if their canonicalizations are isometric.*

Proof. If $\mathcal{C}$ is canonical for X and Y, then $X \cong^w \mathcal{C} \cong^w Y$ and, hence, X and Y are weakly similar. On the other hand, assume that X is weakly similar to Y. By the transitivity of weak similarity, we have

$$\mathcal{C}_X \cong^w X \cong^w Y \cong^w \mathcal{C}_Y,$$

therefore, $\mathcal{C}_X$ and $\mathcal{C}_Y$ are weakly similar. By Lemma 2, we have that they have to be isometric. □

Remark 4. Note that in order for the construction of the canonicalization to be possible we do not require the finiteness of the semi-metric space, but rather the finiteness of its distance set. It is therefore possible to extend the concepts introduced so far to non-finite semi-metric spaces whenever the distance function assumes only a finite number of values.

3.3 Curvature sets of finite metric spaces

Establishing whether two spaces are isometric or not is in general a computationally intensive problem. Hence, we would like to have a set of complete or incomplete invariants to study this problem. We will focus our attention firstly on the concept of curvature set introduced by Gromov in [53]. They have already been used by Mémoli in [77], to obtain a lower bound for the Modified Gromov-Hausdorff distance between two metric spaces.

Definition 39 (curvature set - [53])**.** Given a, non-necessarily finite, metric space (X, d_X) we can consider the function

$$\begin{aligned} \Psi_X^m : X^m &\longrightarrow \mathbb{R}^{m\times m} \\ (x_1, \dots, x_m) &\longmapsto M \text{ s.t. } M_{i,j} = d_X(x_i, x_j). \end{aligned} \tag{3.20}$$

We call *m-th curvature set* of (X, d_X) the set

$$\mathrm{K}_m(X) := \mathrm{im}\Psi_X^m. \tag{3.21}$$

Let us see an example of some curvature sets for a finite metric space.

Example 9. Consider the set $X = \{x_1, x_2, x_3\}$ and endow it with the metric d such that $d(x_1, x_2) = 3$, $d(x_1, x_3) = 5$, $d(x_2, x_3) = 4$. Then, $\mathrm{K}_2(X)$ and $\mathrm{K}_3(x)$ are

$$\mathrm{K}_2(X) = \left\{ \begin{bmatrix} 0 & 3 \\ 3 & 0 \end{bmatrix}, \begin{bmatrix} 0 & 4 \\ 4 & 0 \end{bmatrix}, \begin{bmatrix} 0 & 5 \\ 5 & 0 \end{bmatrix}, \begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix} \right\}.$$

$$\mathrm{K}_3(X) = \left\{ P^T \begin{bmatrix} 0 & 3 & 5 \\ 3 & 0 & 4 \\ 5 & 4 & 0 \end{bmatrix} P,\ P^T \begin{bmatrix} 0 & 0 & 3 \\ 0 & 0 & 3 \\ 3 & 3 & 0 \end{bmatrix} P,\ P^T \begin{bmatrix} 0 & 0 & 4 \\ 0 & 0 & 4 \\ 4 & 4 & 0 \end{bmatrix} P, \right.$$

$$\left. P^T \begin{bmatrix} 0 & 0 & 5 \\ 0 & 0 & 5 \\ 5 & 5 & 0 \end{bmatrix} P,\ \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} \;\middle|\; P \text{ runs in } 3\times 3 \text{ permutation matrices} \right\}.$$

Curvature sets encode information about finite metric subspaces of a given metric space. A similar, but coarser, invariant is the isometric sequence of a space, introduced by Hirasaka and Shinohara [56]. Curvature sets are important in virtue of the next theorem.

Theorem 14 (isometry of compact metric spaces - [53])**.** *Two compact metric spaces X and Y are isometric if and only if $\mathrm{K}_m(X) = \mathrm{K}_m(Y)$ for all $m \in \mathbb{N}$.*

Remark 5. Notice that every finite metric space is compact and then satisfies the hypothesis of the above theorem.

Therefore, checking the equality of curvature sets is a way to find out whether two metric spaces are isometric or not. In the general case, in which a metric space is not finite, we have to prove the equality of all curvature sets to ensure the isometry between metric spaces, but for the finite case the problem becomes easier. We can see that the r-th curvature set carries all the information included in all the l-th curvature sets for $l < r$.

Lemma 3. *For any two, possibly infinite, metric spaces (X, d_X), (Y, d_Y), if $\mathrm{K}_r(X) = \mathrm{K}_r(Y)$ for a certain $r \in \mathbb{N}$ then $\mathrm{K}_l(X) = \mathrm{K}_l(Y)$ for all $l \leq r$.*

Proof. Each matrix of $\mathrm{K}_l(X)$ can be obtained from a matrix of $\mathrm{K}_r(X)$ removing $r - l$ rows and columns. Given a set of indices $I := \{i_1, \dots, i_k\} \subseteq \{1, \dots, n\}$ and a matrix $M \in \mathbb{R}^{n \times n}$ we can define M_I as the matrix obtained by M removing the columns and the rows whose indices are in I. Then

$$\begin{aligned} \mathrm{K}_l(X) &= \{M_I \mid M \in \mathrm{K}_r(X),\ I \subseteq \{1, \dots, n\},\ |I| = r - l\} = \\ &= \{M_I \mid M \in \mathrm{K}_r(Y),\ I \subseteq \{1, \dots, n\},\ |I| = r - l\} = \mathrm{K}_l(Y). \end{aligned} \tag{3.22}$$

□

For finite metric spaces we can further improve this result. In the following corollary, we show that given a finite metric space X of cardinality n, all the curvature sets are determined by $\mathrm{K}_n(X)$.

Corollary 2. *Let (X, d_X) and (Y, d_Y) be two finite metric space of cardinality n. Then*

$$\mathrm{K}_m(X) = \mathrm{K}_m(Y) \quad \forall m \in \mathbb{N} \iff \mathrm{K}_n(X) = \mathrm{K}_n(Y). \tag{3.23}$$

Proof. The forward implication is given by the hypothesis. We have to prove the other direction $\Leftarrow$. We already know, by Lemma 3, that $\mathrm{K}_m(X) = \mathrm{K}_m(Y)$ for all $m \leq n$. We have only to prove the case in which $m > n$. For every matrix $M \in \mathrm{K}_m(X)$, we can find m points $x_1, \dots, x_m$ of X such that $\Psi_X^m(x_1, \dots, x_m) = M$. Of these m points at most $k \leq n$ of them can be different, let them be $x_{i_1}, \dots, x_{i_k}$.

Then, the matrix $\Psi_X^k(x_{i_1}, \dots, x_{i_k})$ is in $\mathrm{K}_k(X) = \mathrm{K}_k(Y)$ by Lemma 3. Then, we have $y_1, \dots, y_k \in Y$ such that $\Psi_X^k(x_{i_1}, \dots, x_{i_k}) = \Psi_Y^k(y_1, \dots, y_k)$, and it means that

$$d_X(x_{i_a}, x_{i_b}) = d_Y(y_a, y_b). \tag{3.24}$$

Consider the bijection $\phi : \{x_{i_1}, \dots, x_{i_k}\} \longrightarrow \{y_1, \dots, y_k\}$ with $\phi(x_{i_j}) = y_j$, $j = 1, \dots, k$. Thanks to Eq. (3.24) we have that

$$\Psi_X^m(x_1, \dots, x_m) = \Psi_Y^m(\phi(x_1), \dots, \phi(x_m)) \in \mathrm{K}_m(Y).$$

Therefore, $\mathrm{K}_m(X) \subseteq \mathrm{K}_m(Y)$. Analogously, we can see that $\mathrm{K}_m(Y) \subseteq \mathrm{K}_m(X)$, hence they must be equal. □

We have seen in Example 9 that when we compute the m-th curvature set of a metric spaces we take m-tuples of points of X with repetitions. This means computing n^m matrices. We would like to reduce such a computational cost, and we try to do so introducing the concept of reduced curvature set.

Definition 40 (reduced curvature set)**.** Consider a metric space (X, d_X) and the associated function $\Psi_X^m : X^m \longrightarrow \mathbb{R}^{m \times m}$ defined in Eq. (3.20). We call *m-th reduced curvature set* of (X, d_X) the set

$$\tilde{\mathrm{K}}_m(X) = \{\Psi_X^m(x_1, \dots, x_m) \mid x_1, \dots, x_m \in X \text{ and } x_i \neq x_j \text{ if } i \neq j\}. \tag{3.25}$$

As we can see from the definition, to obtain the m-th reduced curvature set we need to compute $\frac{n!}{(n-m)!}$ matrices. We want to show that we do not lose any information with this reduction.

Lemma 4. *For any two metric spaces (X, d_X), (Y, d_Y), if $\tilde{\mathrm{K}}_r(X) = \tilde{\mathrm{K}}_r(Y)$ for a certain $r \in \mathbb{N}$, then $\tilde{\mathrm{K}}_l(X) = \tilde{\mathrm{K}}_l(Y)$ for all $l \leq r$.*

Proof. The proof is analogous to that of Lemma 3. □

Corollary 3. *Let (X, d_X) and (Y, d_Y) be two finite metric spaces of cardinality n. For any $m \leq n$, we have*

$$\mathrm{K}_m(X) = \mathrm{K}_m(Y) \iff \tilde{\mathrm{K}}_m(X) = \tilde{\mathrm{K}}_m(Y). \tag{3.26}$$

Proof. Assume $\mathrm{K}_m(X) = \mathrm{K}_m(Y)$. Then a matrix $M \in \mathrm{K}_m(X)$ is also an element of $\tilde{\mathrm{K}}_m(X)$ if and only if $(M_{i,j} = 0 \iff i = j)$. The same argument holds also for $\mathrm{K}_m(Y)$, therefore, given $M \in \tilde{\mathrm{K}}_m(X)$, we have $M \in \mathrm{K}_m(X) = \mathrm{K}_m(Y)$ and $M \in \mathrm{K}_m(Y)$. Since M has null entries only in its diagonal, $M \in \tilde{\mathrm{K}}_m(Y)$. In this way, we can see that $\tilde{\mathrm{K}}_m(X) \subseteq \tilde{\mathrm{K}}_m(Y)$ and $\tilde{\mathrm{K}}_m(Y) \subseteq \tilde{\mathrm{K}}_m(X)$, hence they are equal. Assume now that $\tilde{\mathrm{K}}_m(X) = \tilde{\mathrm{K}}_m(Y)$. We want to prove that, for any $M \in \mathrm{K}_m(X) \setminus \tilde{\mathrm{K}}_m(X)$, we have $M \in \mathrm{K}_m(Y)$. We know there are $x_1, \dots, x_m \in X$ such that $M = \Psi_X^m(x_1, \dots, x_m)$, where at most $k < m$ of the points are different.

Suppose these points are $x_{i_1}, \dots, x_{i_k}$. For Lemma 4, we have that $\tilde{\mathrm{K}}_{m-k}(X) = \tilde{\mathrm{K}}_{m-k}(Y)$, then we have $y_1, \dots . y_k$ points of Y such that

$$\Psi_X^{m-k}(x_{i_1}, \dots, x_{i_k}) = \Psi_Y^{m-k}(y_1, \dots, y_k).$$

Hence, given the bijection $\phi : \{x_{i_1}, \dots, x_{i_k}\} \longrightarrow \{y_1, \dots, y_k\}$ with $\phi(x_{i_j}) = y_j$, $j = 1, \dots, k$, we have

$$M = \Psi_X^m(x_1, \dots, x_m) = \Psi_Y^m(\phi(x_1), \dots, \phi(x_m)) \in \mathrm{K}_m(Y). \tag{3.27}$$

Then, $\mathrm{K}_m(X) \subseteq \mathrm{K}_m(Y)$ and reasoning in the same way we have $\mathrm{K}_m(Y) \subseteq \mathrm{K}_m(X)$, therefore they are equal. □

Thanks to the last corollary, we can see that reduced curvature sets carry the same information given by the non reduced version. Notice that for finite metric spaces we do not have m-th reduced curvature sets, with m greater than the number of points of the space. In the following corollary, we observe that isometry of finite metric spaces is characterised by the n-th reduced curvature set.

Corollary 4. *Two finite metric spaces (X, d_X) and (Y, d_Y) of cardinality n are isometric if and only if $\tilde{\mathrm{K}}_n(X) = \tilde{\mathrm{K}}_n(Y)$.*

Proof. We need to prove the "⇐" implication only. Since X and Y are finite they are compact and by Theorem 14 we know that they are isometric if and only if $\mathrm{K}_m(X) = \mathrm{K}_m(Y)$ for all $m \in \mathbb{N}$. By Corollary 2, since X and Y are finite, we know that this is true if and only if $\mathrm{K}_n(X) = \mathrm{K}_n(Y)$ and for Corollary 3 this holds if and only if $\tilde{\mathrm{K}}_n(X) = \tilde{\mathrm{K}}_n(Y)$. □

We recall that, thanks to Theorem 13, two spaces are weakly similar if and only if their canonicalizations are isometric, and this condition can now be checked using the above theorem. Hence, we have the following corollary.

Corollary 5. *Two semi-metric spaces (X, d_X), (Y, d_Y) of cardinality n with respective canonicalizations $(\mathcal{C}_X, d_{\mathcal{C}_X})$, $(\mathcal{C}_Y, d_{\mathcal{C}_Y})$ are weakly similar if and only if $\tilde{\mathrm{K}}_n(\mathcal{C}_X) = \tilde{\mathrm{K}}_n(\mathcal{C}_Y)$.*

3.4 A dissimilarity measure for weak similarity

We want to consider some of the possible distances between semi-metric spaces to compare them from the point of view of weak similarity. The purpose of these distances is twofold:

- having a criterion to determine whether or not two spaces are weakly similar;
- measuring how far two spaces are from being weakly similar.

Depending on the purpose of the distance, different properties are desirable. In general, the distances we want to define must discriminate spaces that are not weakly similar, therefore they must attain the value 0 if and only if two spaces are weakly similar. This will immediately address the first desideratum we specified above. In general, to obtain a distance that satisfies this property, it is sufficient to define a distance between metric spaces that is 0 if and only if two spaces are isometric, and then to consider its pullback via the canonicalization. Let us call $\mathcal{C} : \mathrm{FSS} \to \mathrm{FMS}$ the map that assigns to each semi-metric space its canonicalization. Consider a distance d between metric spaces, such that $d(X, Y) = 0$ if and only if two spaces are isometric. Then,

$$\hat{d}(X, Y) = d(\mathcal{C}(X), \mathcal{C}(Y)) \tag{3.28}$$

is a pseudo-distance between finite semi-metric spaces that is 0 if and only if two spaces are weakly similar. We recall that a *pseudo-distance* on a set X is a function d that satisfies all the conditions of a metric, except for the fact that if $d(x, y) = 0$ for $x, y \in X$ does not imply $x = y$.

Example 10. Consider the Gromov-Hausdorff distance d_{GH}. The function

$$\hat{d}_{GH}(X, Y) = d_{GH}(\mathcal{C}(X), \mathcal{C}(Y)) \tag{3.29}$$

is a pseudo-distance between semimetric spaces, that is zero if and only if two spaces are weakly similar. This function is well defined even on spaces of different cardinality.

Example 11. Consider the distance d_∞ on the sub-collection of FMS of spaces with fixed cardinality equal to n. Given $A, B \in \mathrm{FMS}$, with $|A| = |B| = n$, the distance d_∞ between them is defined as

$$d_\infty(A, B) = \inf_{\phi \in \Phi} \max_{a_1, a_2 \in A} |d_A(a_1, a_2) - d_B(\phi(a_1), \phi(a_2))|, \tag{3.30}$$

where Φ is the set of all bijections between A and B. Then, $d_\infty(A, B) = 0$ if and only if A and B are isometric. Therefore, given two semi-metric spaces (X, d_X) and (Y, d_Y), with $|X| = |Y|$, the function

$$\hat{d}_\infty(X, Y) = d_\infty(\mathcal{C}(X), \mathcal{C}(Y)) \tag{3.31}$$

is zero if and only if two spaces are weakly similar. This function is well defined only on spaces of the same cardinality.

The main drawback of the functions defined so far is that they rely on the concept of canonicalization of a semi-metric space. The canonicalization is an arbitrary concept, there may be infinite equivalent ways to define a canonical space.

In [49], for example, the authors obtain canonical spaces ordering the distances with natural numbers. In the approach described in [46] the distances are rescaled so that the maximum value attained by the distance function is always equal to $2\binom{n}{2}$, and the other distances are descending natural numbers starting from that value. Different choices of canonicalizations will lead to different distances. If the purpose of the distance is to decide whether two spaces are weakly similar or not, this is not a problem, since it suffices to ascertain whether or not the distance is zero. On the other hand, if we want to make a quantitative comparison between spaces, it is desirable to have a distance which does not depend on the notion of canonicalization. In literature, there are already distances that can be used to detect weak similarity and that do not depend on the choice of canonicalization. In [65] the authors propose a distance between ordinal spaces of the same cardinality.

Definition 41. Consider two finite ordinal spaces (X, δ_X) and (Y, δ_Y), with $|X| = |Y|$. Given a bijection $f : X \to Y$, we denote by $\Delta_f(X, Y)$ the set

$$\left\{(x, y, w, z) \in X^4 \mid \delta_X(x, y, w, z) \neq \delta_Y(f(x), f(y), f(w), f(z))\right\}. \tag{3.32}$$

The ordinal distance between (X, δ_X) and (Y, δ_Y) is

$$d_{\text{ord}}(X, Y) = \min_{f \in \Phi(X,Y)} \frac{1}{8} |\Delta_f(X, Y)|, \tag{3.33}$$

where $\Phi(X, Y)$ is the set of all bijections between X and Y.

This function is a distance between ordinal spaces and which arises from counting the number of relations between pairs of points that are not preserved under bijections between the two spaces.

Theorem 15 ([65])**.** *d_{ord} is a distance on the set of isomorphic types of finite ordinal spaces with a fixed number of points.*

Each semi-metric space (X, d_X) has an associated ordinal space $\tau((X, d_X)) = (X, \delta_X)$, as in Example 7. Then we can define a pseudo-distance between semi-metric spaces as the pullback of d_{ord} via τ. Explicitly, the function

$$\hat{d}_{\text{ord}}((X, d_X), (Y, d_Y)) = d_{\text{ord}}(\tau(X), \tau(Y)) \tag{3.34}$$

is a pseudo-metric between semi-metric spaces, that is zero if and only if two spaces are weakly similar. The pseudo-distance induced by d_{ord} can be directly defined on semi-metric spaces in the following way. Consider the functions $\text{sign}(x) : \mathbb{R} \to \{-1,0,1\}$ and $\sigma : \{-1,0,1\} \times \{-1,0,1\} \to \{0,1\}$, where

$$\text{sign}(x) = \begin{cases} -1 & \text{if } x < 0, \\ 0 & \text{if } x = 0, \\ 1 & \text{if } x > 0, \end{cases} \tag{3.35}$$

and

$$\sigma(i,j) = \begin{cases} 0 & \text{if } i = j, \\ 1 & \text{if } i \neq j. \end{cases} \tag{3.36}$$

Consider two semi-metric spaces (X, d_X), (Y, d_Y). Given a bijection $\phi : X \to Y$ and four points $x, y, w, z \in X$, let $q_\phi(x, y, w, z)$ be the number

$$\sigma\left(\operatorname{sign}\left(d_X(x,y) - d_X(w,z)\right),\ \operatorname{sign}\left(d_Y(\phi(x),\phi(y)) - d_Y(\phi(w),\phi(z))\right)\right). \tag{3.37}$$

The pseudo-distance $\hat{d}_{\text{ord}}$ between (X, d_X) and (Y, d_Y) is equivalent to

$$\hat{d}_{\text{ord}}(X,Y) = \frac{1}{8} \min_{\phi \in \Phi} \sum_{x,y,w,z \in X^4} q_\phi(x,y,w,z), \tag{3.38}$$

where Φ is the set of all bijections between X and Y.

Given two semi-metric spaces it would be reasonable to ask how much it would cost to transform the two spaces so that they are weakly similar. The distance $\hat{d}_{\text{ord}}$ counts only the number of relations that would have to be modified to reach this goal, but does not take into account the actual values of the distance functions. Moreover, it is defined only between spaces of the same cardinality and it cannot be used to compare spaces with different number of points. Two spaces of different cardinality clearly cannot be weakly similar, but we may wonder how much they have to be changed to become weakly similar.

Example 12. Consider the two spaces in the following figure:

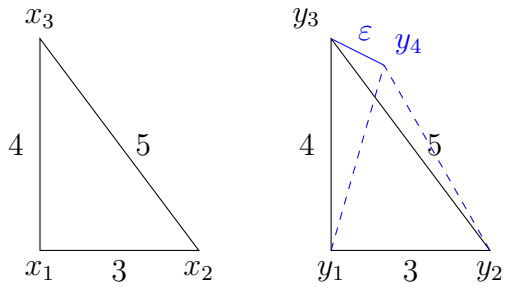


Figure 3.2: Compare different spaces from the point of view of weak similarity.

The two spaces are clearly not weakly similar since they have different cardinalities. On the other hand, if the point y_4 collapsed on y_3, then the two spaces would be isometric. Therefore, with a small change the two spaces become weakly similar.

3.4.1 Dissimilarity for comparing spaces

To formalise the idea in the example above, we introduce a dissimilarity measure for assessing how far the spaces are from being weakly similar. We first recall the notion of dissimilarity [37].

From now on, we will denote by $\mathscr{I}$ the set of strictly increasing functions $\psi : \mathbb{R}^+ \longrightarrow \mathbb{R}^+$, with $\psi(0) = 0$.

Definition 42. Given two semi-metric spaces (X, d_X) and (Y, d_Y), the *weak Gromov-Hausdorff dissimilarity* d_{wGH} is defined as

$$\begin{aligned} d_{wGH}((X, d_X), (Y, d_Y)) = &\inf_{\psi_1 \in \mathscr{I}} d_{GH}((X, \psi_1 \circ d_X), (Y, d_Y)) + \\ &+ \inf_{\psi_2 \in \mathscr{I}} d_{GH}((X, d_X), (Y, \psi_2 \circ d_Y)). \end{aligned} \tag{3.39}$$

This is only a dissimilarity on the collection of finite semi-metric spaces since the triangle inequality is not satisfied and different spaces can have dissimilarity values equal to 0 (as we will prove in proposition 3 ...).

Remark 6. Because of Remark 1, the function d_{wGH} is well-defined, even if one of the functions $\psi \circ d_X$ or $\psi \circ d_Y$ does not satisfy the triangle inequality, since the Gromov-Hausdorff distance, defined using the distortion of a correspondence, is a pseudo-distance on the collection of networks. Notice that for any ψ in $\mathscr{I}$ the space $(X, \psi \circ d_X)$ is a dissimilarity space, since $\psi \circ d_X(x_1, x_2) = 0$ if and only if $x_1 = x_2$. Then, if $d_{GH}((X, \psi \circ d_X), (Y, d_Y)) = 0$, it holds $(X, \psi \circ d_X) \cong^w (Y, d_Y)$.

We have the following result

Proposition 3. *The map d_{wGH} is a dissimilarity on the collection of finite semi-metric spaces.*

Proof. Since d_{GH} is a distance, for all (X, d_X) and (Y, d_Y) it holds that

$$0 \leq \inf_{\psi \in \mathscr{I}} d_{GH}((X, \psi \circ d_X), (Y, d_Y)) < \infty$$

and

$$0 \leq \inf_{\psi \in \mathscr{I}} d_{GH}((X, d_X), (Y, \psi \circ d_Y)) < \infty.$$

Therefore, $0 \leq d_{wGH}(X, d_X), (Y, d_Y)) < \infty$ for all (X, d_X) and (Y, d_Y). It is clear that d_{wGH} is symmetric by its very definition. Hence, d_{wGH} is a dissimilarity. □

Now, we show that that dissimilarity d_{wGH} discriminates spaces that are not weakly similar, i.e., it is 0 if and only if two spaces are weakly similar.

Proposition 4. *Given two semi-metric spaces (X, d_X), (Y, d_Y), we have*

$$d_{wGH}((X, d_X), (Y, d_Y)) = 0 \iff (X, d_X) \cong^w (Y, d_Y).$$

Proof. If $(X, d_X) \cong^w (Y, d_Y)$, there exists a strictly increasing function ψ such that $(X, \psi \circ d_X)$ is isometric to (Y, d_Y) and $(Y, \psi^{-1} \circ d_Y)$ is isometric to (X, d_X). Since the Gromov-Hausdorff distance between two metric spaces is zero if and only if they are isometric, both the infima on the right hand side of Eq. (3.39) are 0, hence $d_{wGH}((X, d_X), (Y, d_Y)) = 0$.
On the other hand, suppose that $d_{wGH}((X, d_X), (Y, d_Y)) = 0$, so that

$$\inf_{\psi \in \mathscr{I}} d_{GH}((X, \psi \circ d_X), (Y, d_Y)) = 0 \text{ and } \inf_{\psi \in \mathscr{I}} d_{GH}((X, d_X), (Y, \psi \circ d_Y)) = 0.$$

Therefore, by the definition of infimum, there exist two sequences $(\psi_n)_{n\in\mathbb{N}} \subseteq \mathscr{I}$ and $(\tilde{\psi}_n)_{n\in\mathbb{N}} \subseteq \mathscr{I}$ such that

$$\begin{aligned} \lim_{n\to\infty} d_{GH}((X, \psi_n \circ d_X), (Y, d_Y)) = 0 \\ \lim_{n\to\infty} d_{GH}((X, d_X), (Y, \tilde{\psi}_n \circ d_Y)) = 0. \end{aligned} \tag{3.40}$$

Let us focus our attention on the first sequence, $(\psi_n)_{n\in\mathbb{N}} \subseteq \mathscr{I}$. By the finiteness of $\mathfrak{C}(X, Y)$, for every n in $\mathbb{N}$ there exists a, possibly non-unique, correspondence R_n in $\mathfrak{C}(X, Y)$ such that $\mathrm{dis}(R_n, \psi_n \circ d_X, d_Y) = d_{GH}((X, \psi_n \circ d_X), (Y, d_Y))$. By the axiom of choice, it is possible to construct a sequence $(\psi_n, R_n)_{n\in\mathbb{N}} \subseteq \mathscr{I} \times \mathfrak{C}(X, Y)$ such that

$$\lim_{n\to\infty} \mathrm{dis}(R_n, \psi_n \circ d_X, d_Y) = 0.$$

The set $\mathfrak{C}(X, Y)$ is finite, henceforth we can find a subsequence $(\hat{\psi}_n, \hat{R}_n)_{n\in\mathbb{N}}$ of $(\psi_n, R_n)_{n\in\mathbb{N}}$ such that there exists a R_1 in $\mathfrak{C}(X, Y)$ and a $\bar{n}$ in $\mathbb{N}$ with $\hat{R}_n = R_1$, for all $n \geq \bar{n}$.

For this correspondence R_1, it holds

$$\lim_{n\to\infty} |\hat{\psi}_n(d_X(x, x')) - d_Y(y, y')| = 0 \quad \forall (x, y), (x', y') \in R_1. \tag{3.41}$$

Hence, the restriction of the sequence $(\hat{\psi}_n)$ on the distance set $\mathrm{D}(X)$ converges to a function $\psi_X : \mathrm{D}(X) \longrightarrow \mathrm{D}(Y)$, such that $d_{GH}((X, \psi_X \circ d_X), (Y, d_Y)) = 0$. In the same way, we can prove the existence of a function $\psi_Y : \mathrm{D}(Y) \longrightarrow \mathrm{D}(X)$ such that $d_{GH}((X, d_X), (Y, \psi_Y \circ d_Y)) = 0$.

We observe that

$$\begin{aligned} d_{GH}((X, \psi_Y \circ \psi_X \circ d_X), (X, d_X)) \leq d_{GH}((X, \psi_Y \circ \psi_X \circ d_X), (Y, \psi_Y \circ d_Y)) + \\ + d_{GH}((Y, \psi_Y \circ d_Y), (X, d_X)). \end{aligned} \tag{3.42}$$

We already know that the second summand on the right hand side of the inequality is 0. For the first one it is easy to see that since $d_{GH}((X, \psi_X \circ d_X), (Y, d_Y)) = 0$,

then also $d_{GH}((X, \psi_Y \circ \psi_X \circ d_X), (Y, \psi_Y \circ d_Y)) = 0$. Therefore, $d_{GH}((X, \psi_Y \circ \psi_X \circ d_X), (X, d_X)) = 0$, and $(X, \psi_Y \circ \psi_X \circ d_X)$ and (X, d_X) are isometric. Since ψ_X and ψ_Y are both non decreasing functions, also their composition is non decreasing. Moreover, $\psi_Y \circ \psi_X$ has to be a bijective function from $\mathrm{D}(X)$ to itself, otherwise the two spaces fail to be isometric. Then, it has to be $\psi_Y \circ \psi_X = \mathrm{id}\,|_{\mathrm{D}(X)}$, therefore ψ_X and ψ_Y are invertible. It is possible to extend, by linear interpolation, the domain and codomain of ψ_X to $\mathbb{R}^+$, thus we have a strictly increasing function ψ_X such that $d_{GH}((X, \psi_X \circ d_X), (Y, d_Y)) = 0$ and this is equivalent to saying that (X, d_X) and (Y, d_Y) are weakly similar. □

Example 13. In this example, we exemplify the computation of d_{wGH} between three finite metric spaces. Let us consider the spaces (X, d_X), (Y, d_Y) and (Z, d_Z) depicted in Fig. 3.3. We can see that $\inf_{\psi \in \mathscr{I}} d_{GH}((X, \psi \circ d_X), (Y, d_Y)) = 0$. In fact, if we take a sequence $(\psi_n)_{n \in \mathbb{N}}$ such that

$$\psi_n(3) = 3,\ \psi_n(4) = 4,\ \psi_n(5) = 4 + \frac{1}{n},$$

clearly $\lim_{n \to \infty} d_{GH}((X, \psi_n \circ d_X), (Y, d_Y)) = 0$. On the other hand, for $\hat{\psi}$ with

$$\hat{\psi}(3) = 3,\ \hat{\psi}(4) = 4.5,$$

it holds $\inf_{\psi \in \mathscr{I}} d_{GH}((X, d_X), (Y, \psi \circ d_Y)) = d_{GH}((X, d_X), (Y, \hat{\psi} \circ d_Y)) = 0.5$. Therefore, $d_{wGH}((X, d_X), (Y, d_Y)) = 0.5$. Reasoning in a similar way it is possible to show that $d_{wGH}((Z, d_Z), (Y, d_Y)) = 1$. We also know by Proposition 4 that since X and Z are weakly similar it has to be $d_{wGH}((X, d_X), (Z, d_Z)) = 0$. Hence,

$$d_{wGH}((Y, d_Y), (Z, d_Z)) = 1 > 0.5 = d_{wGH}((Y, d_Y), (X, d_X)) + d_{wGH}((X, d_X), (Z, d_Z))$$

and the triangle inequality does not hold.

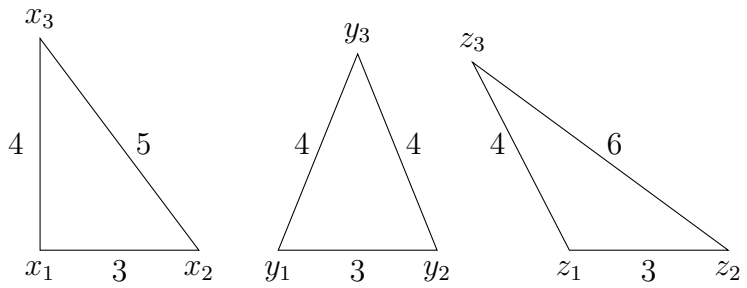


Figure 3.3: Examples for the computation of d_{wGH}.

3.5 Vietoris-Rips filtration and Persistent Homology applied to weak similarities.

We will provide a categorification of the concept of weak similarity of semi-metric spaces, following the ideas for finite metric spaces introduced in [46].

Definition 43 (monotone map between semi-metric spaces)**.** Given two semi-metric spaces (X, d_X) and (Y, d_Y), we say the a map $f : X \longrightarrow Y$ is *monotone* if, for all $x_1, x_2, x_1', x_2' \in X$, we have:

$$d_X(x_1, x_2) \leq d_X(x_1', x_2') \Rightarrow d_Y(f(x_1), f(x_2)) \leq d_Y(f(x_1'), f(x_2')). \tag{3.43}$$

Remark 7 (Proposition 3 - [46]). If $f : X \longrightarrow Y$ is a monotone map, then

$$d_X(x_1, x_2) = d_X(x_1', x_2') \Rightarrow d_Y(f(x_1), f(x_2)) = d_Y(f(x_1'), f(x_2')). \tag{3.44}$$

The converse is not true.

Lemma 5. *A monotone map $f : X \longrightarrow Y$ between two semi-metric spaces (X, d_X) and (Y, d_Y) induces a non-decreasing function between the distance sets $\tilde{f} : \mathrm{D}(X) \longrightarrow \mathrm{D}(Y)$ given by*

$$\tilde{f}(a) = d_Y(f(x_i), f(x_j)) \qquad \text{where } d_X(x_i, x_j) = a. \tag{3.45}$$

Proof. Thanks to Remark 7, we have that the function $\tilde{f}$ is well defined. In fact, for any $a \in \mathrm{D}(X)$, we have that, if $d_X(x_i, x_j) = d_X(x_i', x_j') = a$, we can write $\tilde{f}(a) = d_Y(f(x_i), f(x_j)) = d_Y(f(x_i'), f(x_j'))$. The function is non-decreasing because if $a = d_X(x_1, x_2) \leq b = d_X(x_3, x_4)$, by Definition 43, $\tilde{f}(a) = d_Y(f(x_1), f(x_2)) \leq d_Y(f(x_3), f(x_4)) = \tilde{f}(b)$. □

Lemma 6. *A monotone map $f : X \longrightarrow Y$ between two semi-metric spaces (X, d_X) and (Y, d_Y) induces a non-decreasing function $\hat{f} : \mathbb{R}^+ \longrightarrow \mathbb{R}^+$ whose restriction is $\tilde{f}$ as in Lemma 5.*

Proof. Thanks to Lemma 5, we have that f induces a non-decreasing function $\tilde{f} : \mathrm{D}(X) \longrightarrow \mathrm{D}(Y)$. Such a function can be extended to a non decreasing function $\hat{f} : \mathbb{R}^+ \longrightarrow \mathbb{R}^+$ in the following way. If $\mathrm{D}(X) = \{0, a_1, \ldots, a_k \mid a_i < a_j \text{ if } i < j\}$, we define $\hat{f}$ as

$$\hat{f}(x) = \begin{cases} \dfrac{\tilde{f}(a_1)}{a_1} x & \text{if } x \in [0, a_1] \\ \dfrac{\tilde{f}(a_{i+1}) - \tilde{f}(a_i)}{a_{i+1} - a_i}(x - a_i) + \tilde{f}(a_i) & \text{if } x \in [a_i, a_{i+1}] \\ (x - a_k) + \tilde{f}(a_k) & \text{if } x \in (a_k, \infty). \end{cases} \tag{3.46}$$

By definition, it follows that $\hat{f}\big|_{\mathrm{D}(X)} = \tilde{f}$. □

Definition 44 (category of FSS - [46])**.** We can define a category **FSS** of finite semi-metric space whose objects are finite semi-metric spaces and whose morphisms are monotone maps.

Remark 8. It is possible to observe that two semi-metric spaces are isomorphic in the category **FSS** if and only if they are weakly similar.

We recall the definition of Vietoris-Rips filtration.

Definition 45 (Vietoris-Rips filtration)**.** Given a metric space (X, d_X), we can consider the functor $\mathrm{VR}_\bullet(X) : (\mathbb{R}^+, \leq) \longrightarrow \mathbf{Simp}$ that assigns to each $a \in \mathbb{R}^+$ the Vietoris-Rips complex $\mathrm{VR}_a(X)$ and to each morphism $a \leq b$ the inclusion $\iota^X_{a\leq b} : \mathrm{VR}_a(X) \hookrightarrow \mathrm{VR}_b(X)$.

Definition 46 (rescaling)**.** Given a non-decreasing function $\psi : \mathbb{R}^+ \longrightarrow \mathbb{R}^+$ we call *ψ-rescaling* the functor $R_\psi : (\mathbb{R}^+, \leq) \longrightarrow (\mathbb{R}^+, \leq)$ with

$$\begin{aligned} R_\psi(a) &= \psi(a) \\ R_\psi(a \leq b) &= \psi(a) \leq \psi(b). \end{aligned} \tag{3.47}$$

Lemma 7. *A morphism $f : X \longrightarrow Y$ in* **FSS** *induces a rescaling $R_{\hat{f}}$ and a natural transformation $\eta^f : \mathrm{VR}_\bullet(X) \Longrightarrow \mathrm{VR}_\bullet(Y)R_{\hat{f}}$.*

Proof. Thanks to Lemma 6, we have a non-decreasing function $\hat{f}$ that induces a rescaling $R_{\hat{f}}$. We want to see that, for any $a \in \mathbb{R}^+$, we have a simplicial map $\eta^f_a : \mathrm{VR}_a(X) \longrightarrow \mathrm{VR}_\bullet(Y)R_{\hat{f}}(a) = \mathrm{VR}_{\hat{f}(a)}(Y)$ such that, for all $a, b \in \mathbb{R}^+$ with $a \leq b$, we have a commutative diagram

$$\begin{array}{ccc} \mathrm{VR}_a(X) & \xhookrightarrow{\iota^X} & \mathrm{VR}_b(X) \\ \downarrow{\scriptstyle \eta^f_a} & & \downarrow{\scriptstyle \eta^f_b} \\ \mathrm{VR}_{\hat{f}(a)}(Y) & \xhookrightarrow{\iota^Y} & \mathrm{VR}_{\hat{f}(b)}(Y). \end{array} \tag{3.48}$$

We define η^f_a as

$$\eta^f_a(\{x_{i_0}, \ldots, x_{i_k}\}) = \{f(x_{i_0}), \ldots, f(x_{i_k})\}. \tag{3.49}$$

By the very definition of Vietoris-Rips complex and $\hat{f}$, we have that, for every simplex σ of $\mathrm{VR}_a(X)$, $\eta^f_a(\sigma)$ is a simplex of $\mathrm{VR}_{\hat{f}(a)}(Y)$ and η^f_a is a well-defined simplicial map. We only need to prove that $\iota^Y \circ \eta^f_a = \eta^f_b \circ \iota^X$. Indeed, for any $\sigma \in \mathrm{VR}_a(X)$ with $\sigma = \{x_{i_0}, \ldots, x_{i_k}\}$ we have

$$\begin{aligned} \iota^Y \circ \eta^f_a(\sigma) &= \iota^Y(\{f(x_{i_0}), \ldots, f(x_{i_k})\}) = \{f(x_{i_0}), \ldots, f(x_{i_k})\} \\ \eta^f_b \circ \iota^X(\sigma) &= \eta^f_b(\{x_{i_0}, \ldots, x_{i_k}\}) = \{f(x_{i_0}), \ldots, f(x_{i_k})\}. \end{aligned} \tag{3.50}$$

Hence η^f is a natural transformation. □

Now, we want to show that the Vietoris-Rips filtration can be used as a tool to study weak similarity.

Theorem 16. *Given two semi-metric spaces (X, d_X) and (Y, d_Y), the following statements are equivalent:*

1. *(X, d_X) and (Y, d_Y) are isomorphic in* **FSS**.
2. *There exist a rescaling R_ψ and a natural isomorphism*

$$\eta : \mathrm{VR}_\bullet(X) \Longrightarrow \mathrm{VR}_\bullet(Y)R_\psi.$$

Proof. Suppose that (X, d_X) and (Y, d_Y) are isomorphic in **FSS**. Then we have a monotone map $f : X \longrightarrow Y$ that it is a bijection. Thanks to Lemma 7, we have a rescaling $R_{\hat{f}}$ and a natural transformation $\eta^f : \mathrm{VR}_\bullet(X) \Longrightarrow \mathrm{VR}_\bullet(Y)R_{\hat{f}}$. We want to see that for all $a \in \mathbb{R}^+$, η^f_a is an isomorphism of simplicial complexes. Since f is a bijection, we know that η^f_a is injective. In fact, given $\sigma = \{x_{i_1}, \ldots, x_{i_k}\}$ and $\tau = \{x_{j_1}, \ldots, x_{j_l}\}$ with $\sigma \neq \tau$, we can assume without loss of generality that there is a $x_j \in \tau$ with $x_j \notin \sigma$. Then if $\eta^f_a(\sigma) = \eta^f_a(\tau)$, there is a $x_i \in \sigma$ with $f(x_j) = f(x_i)$ and this is absurd for the injecivity of f. On the other hand η^f_a is also surjective. Suppose that there is a $\rho \in \mathrm{VR}_{\hat{f}(a)}(Y)$ with $\rho = \{y_{i_1}, \ldots, y_{i_k}\}$ that it is not in the image of η^f_a. We can consider the points $f^{-1}(y_{i_1}), \ldots, f^{-1}(y_{i_k})$ of X and see that they form a simplex of $\mathrm{VR}_a(X)$. In fact, since $\rho \in \mathrm{VR}_{\hat{f}(a)}(Y)$, for all $u, v \in \rho$, we have $d_Y(u, v) \leq \hat{f}(a)$. It can be seen that, for all $x_1, x_2 \in X$, we have $\hat{f}(d_X(x_1, x_2)) = d_Y(f(x_1), f(x_2))$. Therefore, for all $u, v \in \rho$, since $\hat{f}^{-1}$ is also a strictly increasing function

$$d_X(f^{-1}(u), f^{-1}(v)) = \hat{f}^{-1}(d_Y(u, v)) \leq \hat{f}^{-1}(\hat{f}(a)) = a. \tag{3.51}$$

Then points $f^{-1}(y_{i_1}), \ldots, f^{-1}(y_{i_k})$ span a simplex of $\mathrm{VR}_a(X)$ whose image under η^f_a is ρ. Therefore η^f_a is also bijective and is an isomorphism of simplicial complexes. Hence, η is a natural isomorphism.
Now, suppose that we have a rescaling R_ψ and a natural isomorphism $\eta : \mathrm{VR}_\bullet(X) \Rightarrow \mathrm{VR}_\bullet(Y)R_\psi$. For each $a \in \mathbb{R}^+$, the restriction of the isomorphism η_a to the vertices of $\mathrm{VR}_a(X)$ yields a bijection $f_a : X \longrightarrow Y$. Moreover, these bijections are all the same because of the commutativity of diagrams that define the natural isomorphism. Hence we have a unique bijection $f : X \longrightarrow Y$ associated with η. We claim that this bijection is an isomorphism of finite metric spaces. In fact, for each pair of points $x_1, x_2 \in X$, call $\bar{a} = d_X(x_1, x_2)$. Since η is a natural isomorphism, we have that $\sigma = \{x_1, x_2\} \in \mathrm{VR}_{\bar{a}}(X)$ and that $\eta_{\bar{a}}(\sigma) = \{f(x_1), f(x_2)\}$ is a simplex of $\mathrm{VR}_{\psi(\bar{a})}(Y)$, that it is not present in any $\mathrm{VR}_b(Y)$, for $b \leq \psi(\bar{a})$. This means that $d_Y(f(x_1), f(x_2)) = \psi(\bar{a}) = \psi(d_X(x_1, x_2))$, for all $x_1, x_2 \in X$, and therefore (X, d_X) and (Y, d_Y) are weakly similar and also isomorphic in **FSS**. □

By using an argument analogous to the one just described, the following Theorem can be proved.

Theorem 17. *Given two metric spaces (X, d_X) and (Y, d_Y), the following statements are equivalent:*

1. *(X, d_X) and (Y, d_Y) are isometric.*
2. *There exists a natural isomorphism*

$$\eta : \mathrm{VR}_\bullet(X) \Longrightarrow \mathrm{VR}_\bullet(Y).$$

The Vietoris-Rips filtration is a complete invariant for isometry. Therefore, it can be used in conjunction with the canonicalization to produce a complete invariant for weak similarity.

Corollary 6. *Two spaces (X, d_X) and (Y, d_Y) are weakly similar if an only if there is a natural isomorphism between $\mathrm{VR}_\bullet(\mathcal{C}_X)$ and $\mathrm{VR}_\bullet(\mathcal{C}_Y)$.*

Remark 9. Given two Vietoris-Rips filtrations $\mathrm{VR}_\bullet(X)$ and $\mathrm{VR}_\bullet(Y)$, the existence of an isomorphism between each $\mathrm{VR}_a(X)$ and $\mathrm{VR}_a(Y)$ is not enough to ensure that X and Y are isometric. Indeed, all these isomorphism have to commute with the inclusions given by the filtrations. For example, consider the two metric spaces, depicted in Fig. 3.4, given by the distance matrices

$$d_X = (d_X(x_i, x_j)) = \begin{pmatrix} 0 & 7 & 9 & 10 \\ 7 & 0 & 8 & 11 \\ 9 & 8 & 0 & 12 \\ 10 & 11 & 12 & 0 \end{pmatrix},$$
$$d_Y = (d_Y(y_i, y_j)) = \begin{pmatrix} 0 & 7 & 9 & 10 \\ 7 & 0 & 8 & 12 \\ 9 & 8 & 0 & 11 \\ 10 & 12 & 11 & 0 \end{pmatrix}. \tag{3.52}$$

They are not isometric, but, for each $a \in \mathbb{R}^+$, we can find an isomorphism between $\mathrm{VR}_a(X)$ and $\mathrm{VR}_a(Y)$. Notice that, on the other hand, if there exists an a in $\mathbb{R}^+$ such that there is no isomorphism between $\mathrm{VR}_a(X)$ and $\mathrm{VR}_a(Y)$, then the two spaces are for sure not isometric.

3.5.1 Persistent homology as an incomplete invariant for weak similarity

As we have seen in the background chapter, it is possible to apply the homology functor to filtration in order to obtain a persistence module. Then, the following corollary holds.

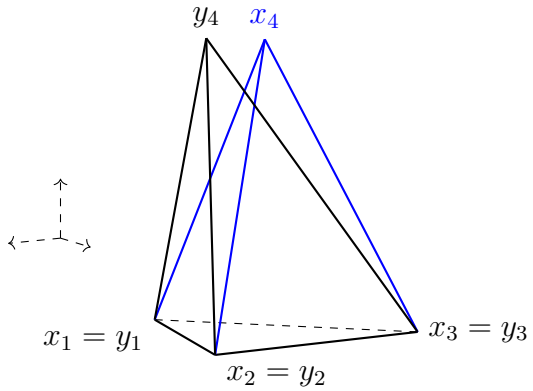


Figure 3.4: Embedding of the spaces of Remark 9 in $\mathbb{R}^3$.

Corollary 7. *The persistent homology of the Vietoris-Rips filtration is an incomplete invariant for isometry.*

Proof. The two spaces in Remark 9 have persistence modules with interleaving distance 0, yet they are not isometric. □

Persistent homology can be used to discriminate non-weakly similar spaces. Given two spaces $(X, d_X), (Y, d_Y)$, the Vietoris-Rips filtration of their canonicalization is computed. Applying the homology functor yields two persistent modules that can be compared with a distance. If such a distance is greater then 0, the two spaces cannot be weakly similar. The most common used distances are the bottleneck distance (equivalent to the interleaving distance) and the Wasserstein distance. This is the approach examined in [49]. Otherwise, the persistence modules can be further processed into a vectorized version, like persistence images [2] or persistence landscapes [14], for which a richer family of metrics is available.

3.5.2 A dissimilarity measure for persistence modules

In a spirit similar to Section 3.4, we can define a dissimilarity between persistence modules. Its aim is to compare persistence modules from the point of view of weak similarity. Comparing rescaling of persistence diagram is not new in literature. In [59] the authors consider and compare log-scaled persistence diagrams. The shift-invariant bottleneck distance [21] has been introduced to compare this kind of persistence diagrams, so that diagrams that are equivalent up to a multiplicative constant can be identified. Recall that we defined $\mathscr{I}$ as the set of strictly increasing functions $\psi : \mathbb{R}^+ \longrightarrow \mathbb{R}^+$, with $\psi(0) = 0$.

Definition 47. Given two persistence modules H_kF_1 and H_kF_2, the *weak interleaving* dissimilarity d_{wI} between them is

$$d_{wI}(H_kF_1, H_kF_2) = \inf_{\psi_1 \in \mathscr{I}} d_I(H_kF_1R_{\psi_1}, H_kF_2) + \inf_{\psi_2 \in \mathscr{I}} d_I(H_kF_1, H_kF_2R_{\psi_2}), \quad (3.53)$$

where d_I is the interleaving distance between persistence modules.

One of the key properties desired for distances between persistence modules is that they satisfy a form of stability theorem, that is, there must be a distance between the original semi-metric spaces, that bounds from above the distance between the obtained persistence modules. We have a stability theorem of this kind for the Vietoris-Rips filtration and the interleaving distance [24], in fact

$$d_I(H_k\mathrm{VR}_\bullet(X,d_X),H_k\mathrm{VR}_\bullet(Y,d_Y)) \leq 2d_{GH}((X,d_X),(Y,d_Y)).$$

We will provide a similar stability theorem also for the dissimilarities introduced in this chapter.

Theorem 18 (stability theorem for weak similarity - (page 50))**.** *Let (X,d_X), (Y,d_Y) be two finite semi-metric spaces. We denote by $H_k\mathrm{VR}_\bullet(X,d_X)$ and $H_k\mathrm{VR}_\bullet(Y,d_Y)$ the k-th persistence modules obtained from the Vietoris-Rips filtration associated with the two spaces. Then for all $k \in \mathbb{N}$,*

$$d_{wI}(H_k\mathrm{VR}_\bullet(X,d_X),H_k\mathrm{VR}_\bullet(Y,d_Y)) \leq 2d_{wGH}((X,d_X),(Y,d_Y)). \tag{3.54}$$

We can see that thanks to this theorem and Proposition 4, if the persistence modules obtained by two Vietoris-Rips filtration have distance d_{wI} greater than 0, then the corresponding finite metric spaces cannot be weakly similar.

Example 14. Consider the finite semi-metric spaces (X,d_X) and (Y,d_Y) in Fig. 3.5 with

$$d_X = (d_X(x_i,x_j)) = \begin{pmatrix} 0 & 7 & 12 & 8 \\ 7 & 0 & 10 & 11 \\ 12 & 10 & 0 & 9 \\ 8 & 11 & 9 & 0 \end{pmatrix},$$
$$d_Y = (d_Y(y_i,y_j)) = \begin{pmatrix} 0 & 7 & 12 & 8 \\ 7 & 0 & 10 & 9 \\ 12 & 10 & 0 & 11 \\ 8 & 9 & 11 & 0 \end{pmatrix}. \tag{3.55}$$

We will use $\mathbb{Z}_2$ as the field of coefficients with which we will compute homology. We can see that the persistence module $H_1\mathrm{VR}_\bullet(X,d_X)$ is the functor

$$H_1\mathrm{VR}_\varepsilon(X,d_X) = \begin{cases} \mathbb{Z}_2 & \text{if } 10 \leq \varepsilon \leq 11 \\ 0 & \text{otherwise,} \end{cases}$$
$$H_1\mathrm{VR}_\bullet(X,d_X)(a \leq b) =$$
$$= \begin{cases} \mathrm{id} : \mathbb{Z}_2 \to \mathbb{Z}_2 & \text{if } 10 \leq a \leq b \leq 11 \\ 0 : H_1\mathrm{VR}_a(X,d_X) \to H_1\mathrm{VR}_b(X,d_X) & \text{otherwise.} \end{cases}$$

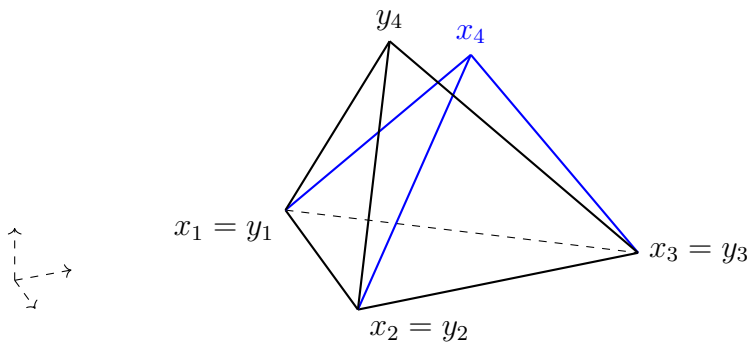


Figure 3.5: Embedding of the spaces of Example 14 in $\mathbb{R}^3$. It holds $d_X(x_i, x_j) = d_Y(y_i, y_j)$ for all $i, j = 1, \ldots, 4$, except for $d_X(x_1, x_4) = d_Y(y_2, y_4)$ and $d_X(x_2, x_4) = d_Y(y_1, y_4)$. This asymmetry will make the spaces non-weakly isometric.

The persistence module $H_1\mathrm{VR}_\bullet(Y, d_Y)$ is

$$H_1\mathrm{VR}_\varepsilon(Y, d_Y) = 0 \quad \forall \varepsilon \geq 0$$
$$H_1\mathrm{VR}_\bullet(Y, d_Y)(a \leq b) = 0 : H_1\mathrm{VR}_a(Y, d_Y) \to H_1\mathrm{VR}_b(Y, d_Y) \quad \forall a \leq b.$$

Let us consider the sequence $(\psi_n)_{n\in\mathbb{N}}$ with

$$\psi_n(x) = \begin{cases} x & \text{if } 0 \leq x \leq 10 \\ n(x-10)+10 & \text{if } 10 < x \leq 10 + \frac{1}{n} \\ x + 1 - \frac{1}{n} & \text{if } 10 + \frac{1}{n} < x. \end{cases}$$

It is easy to see that

$$\lim_{n\to\infty} d_I(H_1\mathrm{VR}_\bullet(Y, d_Y), H_1\mathrm{VR}_\bullet(X, d_X)R_{\psi_n}) = 0.$$

On the other hand, for all strictly increasing functions $\psi : \mathbb{R}^+ \to \mathbb{R}^+$ it holds

$$d_I(H_1\mathrm{VR}_\bullet(X, d_X), H_1\mathrm{VR}_\bullet(Y, d_Y)R_\psi) = \frac{1}{2}.$$

Therefore, $d_{wI}(H_1\mathrm{VR}_\bullet(X, d_X), H_1\mathrm{VR}_\bullet(Y, d_Y)) = \frac{1}{2} > 0$, and the two spaces are not weakly similar.

Proof of Theorem 18. We have that, for every $\psi \in \mathscr{I}$,

$$H_k\mathrm{VR}_\bullet(X, \psi \circ d_X))R_\psi = H_k\mathrm{VR}_\bullet(X, d_X).$$

This is true because $(X, \psi \circ d_X)$ and (X, d_X) are weakly similar and, as in the proof of Theorem 16, there are a natural isomorphism η and a rescaling R_ψ such that $\eta : \mathrm{VR}_\bullet(X, d_X) \Longrightarrow \mathrm{VR}_\bullet(X, \psi \circ d_X)R_\psi$. In this case, the natural isomorphism is

simply the identity between every $\mathrm{VR}_a(X, d_X)$ and $\mathrm{VR}_\psi(a)(X, \psi \circ d_X)$. Therefore, the two functors $\mathrm{VR}_\bullet(X, d_X)$ and $\mathrm{VR}_\bullet(X, \psi \circ d_X)R_\psi$ are equal.
Now, let us take a sequence $(\psi_n)_{n\in\mathbb{N}}$ in $\mathscr{I}$ with

$$\lim_{n\to\infty} d_{GH}((X, \psi_n \circ d_X), (Y, d_Y)) = \inf_{\psi\in\mathscr{I}} d_{GH}((X, \psi \circ d_X), (Y, d_Y))$$

and a sequence $(\bar{\psi}_n)_{n\in\mathbb{N}}$ with

$$\lim_{n\to\infty} d_{GH}((X, d_X), (Y, \bar{\psi}_n \circ d_Y)) = \inf_{\psi\in\mathscr{I}} d_{GH}((X, d_X), (Y, \psi \circ d_Y)).$$

By the classical stability theorem [24], for all n in $\mathbb{N}$, we have

$$\begin{aligned} d_I(H_k\mathrm{VR}_\bullet(X, \psi_n \circ d_X), H_k\mathrm{VR}_\bullet(Y, d_Y)) &\leq 2d_{GH}((X, \psi_n \circ d_X), (Y, d_Y)), \\ d_I(H_k\mathrm{VR}_\bullet(X, d_X), H_k\mathrm{VR}_\bullet(Y, \bar{\psi}_n \circ d_Y)) &\leq 2d_{GH}((X, d_X), (Y, \bar{\psi}_n \circ d_Y)). \end{aligned} \tag{3.56}$$

Notice that the inequalities in Eq. (3.56) hold even if the functions $\psi_n \circ d_X$ and $\bar{\psi}_n \circ d_Y$ do not satisfy the triangular inequality. Recall that, for all n in $\mathbb{N}$ by the definition of infimum

$$\begin{aligned} &\inf_{\psi\in\mathscr{I}} d_I(H_k\mathrm{VR}_\bullet(X, \psi \circ d_X), H_k\mathrm{VR}_\bullet(Y, d_Y)) \leq \\ &\leq d_I(H_k\mathrm{VR}_\bullet(X, \psi_n \circ d_X), H_k\mathrm{VR}_\bullet(Y, d_Y)), \\ &\text{and} \\ &\inf_{\psi\in\mathscr{I}} d_I(H_k\mathrm{VR}_\bullet(X, d_X), H_k\mathrm{VR}_\bullet(Y, \psi \circ d_Y)) \leq \\ &\leq d_I(H_k\mathrm{VR}_\bullet(X, d_X), H_k\mathrm{VR}_\bullet(Y, \bar{\psi}_n \circ d_Y)). \end{aligned} \tag{3.57}$$

By the definition of d_{wI} and because of the previous inequalities it holds

$$\begin{aligned} &d_{wI}(H_k\mathrm{VR}_\bullet(X, d_X), H_k\mathrm{VR}_\bullet(Y, d_Y)) \leq \\ &\lim_{n\to\infty} (d_I(H_k\mathrm{VR}_\bullet(X, \psi_n \circ d_X), H_k\mathrm{VR}_\bullet(Y, d_Y)) + \\ &+ d_I(H_k\mathrm{VR}_\bullet(X, d_X), H_k\mathrm{VR}_\bullet(Y, \bar{\psi}_n \circ d_Y)). \end{aligned} \tag{3.58}$$

Because of the inequalities in Eq. (3.56) and the definition of d_{wGH}, the following is true

$$\begin{aligned} &\lim_{n\to\infty} (d_I(H_k\mathrm{VR}_\bullet(X, \psi_n \circ d_X), H_k\mathrm{VR}_\bullet(Y, d_Y)) + \\ &+ d_I(H_k\mathrm{VR}_\bullet(X, d_X), H_k\mathrm{VR}_\bullet(Y, \bar{\psi}_n \circ d_Y)) \leq \\ &\leq \lim_{n\to\infty} (2d_{GH}((X, \psi_n \circ d_X), (Y, d_Y)) + 2d_{GH}((X, d_X), (Y, \bar{\psi}_n \circ d_Y)) = \\ &= 2d_{wGH}((X, d_X), (Y, d_Y)). \end{aligned} \tag{3.59}$$

Therefore,

$$d_{wI}(H_k\mathrm{VR}_\bullet(X, d_X), H_k\mathrm{VR}_\bullet(Y, d_Y)) \leq 2d_{wGH}((X, d_X), (Y, d_Y)). \tag{3.60}$$

□

From the computational point of view it is more appealing to consider the bottleneck distance between persistence diagrams rather then the interleaving distance. For one-parameter finite type persistence modules, the two distances attain the same value, but the bottleneck distance can be computed with polynomial time algorithms [44].

Definition 48 ([72])**.** Two filtering function $f, g : K \to \mathbb{R}$ are said to be *ordering equivalent* if f and g induce the same pre-order on the simplices of K.

Notice that composing a filtering function f with a strictly increasing function ψ yields a filtering function $\psi \circ f$ that is ordering equivalent to f.

Given a persistence module H_pF, we denote by $PD(H_pF)$ its associated persistence diagram. For any function $\psi : \mathbb{R} \to \mathbb{R}$ and persistence diagram PD we define the multiset $\psi(PD)$ as

$$\psi(PD) = \{(\psi(a), \psi(b)) \mid (a, b) \in PD\}.$$

Proposition 5. *Given a filtering function $f : K \to \mathbb{R}$ on a simplicial complex K and a strictly increasing function $\psi : \mathbb{R} \to \mathbb{R}$, for any $p \in \mathbb{N}$ it holds*

$$\psi(PD(H_p(f))) = PD(H_p(\psi \circ f)).$$

Proof. The functions f and $\psi \circ f$ are ordering equivalent functions since ψ is strictly increasing. Call F the filtration functor induced by f and ψF the filtration functor induced by $\psi \circ f$. For any $u, v \in \mathbb{R}$, with $u \leq v$, it holds $F(u) = \psi F(\psi(u))$ and the following diagram commutes

$$\begin{array}{ccc} F(u) & \xrightarrow{F(u \leq v)} & F(v) \\ \downarrow{=} & & \downarrow{=} \\ \psi F(\psi(u)) & \xrightarrow{\psi F(\psi(u) \leq \psi(v))} & \psi F(\psi(v)). \end{array} \tag{3.61}$$

Therefore, if H_pF is decomposable as $\bigoplus_{j=1}^{N} \chi_{[b_i, d_i]}$ then a decomposition of $H_p\psi F$ is given by $\bigoplus_{j=1}^{N} \chi_{[\psi(b_i), \psi(d_i)]}$. □

Remark 10. From the previous Proposition if follows that given a filtration F and a strictly increasing function ψ, the persistence diagram of H_kFR_ψ is equal to $\psi^{-1}(PD(H_kF))$.

In the same way of the weak interleaving dissimilarity we define a dissimilarity based on the bottleneck distance.

Definition 49 (weak bottleneck dissimilarity)**.** Given two persistence diagrams PD_1, PD_2, the *weak bottleneck dissimilarity* between them is defined as

$$d_{wB}(PD_1, PD_2) = \inf_{\psi_1 \in \mathscr{I}} d_B(\psi_1(PD_1), PD_2) + \inf_{\psi_2 \in \mathscr{I}} d_B(PD_1, \psi_2(PD_2)) \tag{3.62}$$

To compute such a dissimilarity it would be necessary to consider an infinite number of strictly increasing functions. An upper bound to this dissimilarity can be found using the framework introduced by Legonye et al. [72].

Let D be a persistence diagram with finite number of points that are not on the diagonal Δ^∞. The space of diagrams $\{\psi(D) \mid \psi \in \mathcal{I}\}$ can be parametrised in the following way. Consider a filtering function f on a finite simplicial complex K, such that $PD(H_p f) = D$, for a certain $p \in \mathbb{N}$. Now, consider the image of the function f. This is a finite set $\operatorname{im} f = \{a_0, a_1, \dots, a_k \mid a_i < a_j \text{ if } i < j\}$. Define $\Theta(f) \subset \mathbb{R}^{k+1}$ to be the set of vectors $\theta = (\theta_0, \dots, \theta_k)$ such the the following inequalities are satisfied:

$$\begin{aligned} a_0 + \theta_0 &< a_1 + \theta_1, \\ a_1 + \theta_1 &< a_2 + \theta_2, \\ &\dots \\ a_{k-1} + \theta_{k-1} &< a_k + \theta_k. \end{aligned} \tag{3.63}$$

Any such θ induces a strictly increasing function $\psi_\theta : \operatorname{im} f \to \mathbb{R}$, such that $\psi_\theta(a_i) = a_i + \theta_i$

Proposition 6. *Let D be a persistence diagram, realised by a filtering function $f : K \to \mathbb{R}$ on a simplicial complex K and a homology functor H_p. For any $\psi : \mathbb{R} \to \mathbb{R}$ strictly increasing function, there exists a $\theta \in \Theta(f)$ such that*

$$\psi(D) = H_p(\psi_\theta \circ f). \tag{3.64}$$

Proof. Let $\operatorname{im} f$ be the set $\{a_0, a_1, \dots, a_k\}$. Then, it is sufficient to take $\theta \in \mathbb{R}^{k+1}$ such that $\theta_i = \psi(a_i) - a_i$. The statement follows because of Proposition 5. In fact, $PD(H_p(\psi_\theta \circ f)) = \psi_\theta(D)$ and for any $(b_i, d_i) \in D$ it holds $(\psi(b_i), \psi(d_i)) = (\psi_\theta(b_i), \psi_\theta(d_i))$ by the definition of ψ_θ. □

The following Theorem will be useful to approximate d_{wB}.

Theorem 19. *Consider a finite persistence diagram D_0, and a finite persistence diagram D, such that $D = PD(H_p f)$, for a filtering function f. The function $J : \Theta(f) \to \mathbb{R}$, $J(\theta) = d_B(\psi_\theta(D), D_0) = d_B(PD(H_p \psi_\theta \circ f), D_0)$ is smooth almost everywhere.*

Proof. J is the composition of two functions, $B_p : \Theta(f) \to Bar$ and $T : Bar \to \mathbb{R}$, where Bar is the collection of persistence diagrams, $B_p(\theta) = PD(H_p \psi_\theta \circ f)$ and $T(D) = d_B(D, D_0)$. For all $\theta \in \Theta(f)$ the functions $\psi_\theta \circ f$ are ordering equivalent, therefore [72, Theorem 4.9] holds, and B_p is a C^∞ function. Because of [72, Proposition B.1], the function T is generically smooth. The set of barcodes on which T is not smooth is the image of a zero-measure set of $\Theta(f)$. Therefore, J is smooth almost everywhere. □

Given the differentiability of $J(\theta)$ and the fact that Θ is a convex set, a gradient descend algorithm can be used to find a local minimum of the function J. This will not be in general a global minimum of the function, since J is not convex.

Example 15 (J is not convex)**.** Consider $D_0 = \{(2,3)\} \cup \Delta^\infty$ and $D = \{(1,2),(3,4)\} \cup \Delta^\infty$. There are a finite simplicial complex K and a filtering function f such that $D = H_p f$ and $\operatorname{im} f = \{1,2,3,4\}$. Given an arbitrary small ε, consider $\theta_1 = (0, -1+\varepsilon, -1, -1)$ and $\theta_2 = (1,1,0, -1+\varepsilon)$. It holds $J(\theta_1) = J(\theta_2) = \varepsilon/2$. On the other hand, consider the point $\theta_3 = \frac{1}{2}\theta_1 + \frac{1}{2}\theta_2 = (1/2, \varepsilon/2, -1/2, -1+\varepsilon/2)$. Then, for a small enough ε, $J(\theta_3) = 1/2 > \varepsilon/2 = \frac{1}{2}J(\theta_1) + \frac{1}{2}J(\theta_2)$, hence J is not convex.

Nevertheless, this optimization problem makes it possible to compute an upper bound for the dissimilarity d_{wB}. Being able to approximate d_{wB} makes it possible to have a proxy for d_{wGH}. From the computational point of view, this is appealing, since computing the Gromov-Hausdorff distance boils down to solving an NP-problem [77], whereas the bottleneck distance has a polynomial-time formulation.

3.6 Conclusions and future work

We have constructed suitable representative elements for the equivalence classes of the relation of weak similarity for finite semi-metric spaces. Thanks to these representatives we can check in a simple way whether two spaces are weakly similar or not, through approaches based on curvature sets or distances. We studied how to define pseudo-distances between semi-metric spaces to understand if two spaces are weakly similar or not. We introduced the weak Gromov-Hausdorff dissimilarity to measure how much two spaces have to be modified in order for them to be weakly similar. We have shown that Vietoris-Rips filtrations can help us in characterizing the classes of weak similarity. We have seen that we can use persistent homology to try to discriminate non-weakly similar finite metric spaces, and we have defined a weak interleaving dissimilarity between persistence modules and shown that it satisfies a stability theorem for weak similarity. This dissimilarity has the advantage of being based on the bottleneck distance between persistence diagrams, for which polynomial time algorithm are available. We have shown how to approximate it using the framework introduced by Leygonie et al. [72]. In the future, we would like to try to find other and simpler invariants for weak similarity and make a comparison between them. We have also seen the usefulness of persistent homology, and in future work we would like to study the problem of finding distances between persistence modules that are meaningful from the point of view of weak similarity.

Chapter 4

Homological scaffold via minimal homology basis

4.1 Introduction

This chapter is an elaboration of a joint work with Ulderico Fugacci, Marco Guerra, Giovanni Petri and Francesco Vaccarino and it resulted in the paper [55]. Network science has long represented the cornerstone theory in dealing with complex, heterogeneous multi-agent systems. Network descriptions have found wide applications and had a significant impact on a wide range of fields ([80, 7]), including social networks ([51, 98]), epidemiology ([83, 29]), biology ([48, 4]), and neuroscience ([8, 16, 9]). In recent years, new approaches to the analysis of networks and, more generally, complex interacting systems have emerged which leverage topological techniques ([57, 84, 70, 87]). The theory of (or around) persistence has recently been proposed as a framework for the topological skeletonization of spaces, particularly weighted graphs and networks ([68, 63, 47, 23]).
In [85], the generators of persistent homology are used to build one instance of network skeletonization called *homological scaffold*. However, the method has a serious drawback, consisting in the large degree of arbitrariness in the choice of one representative cycle from the many equivalent generating cycles of the same homology class. This is unfortunately a direct consequence of the homology classes being equivalence classes and affects all attempts to localize cycles ([90, 73]). In this work, we set out to address this issue by searching for a form of *canonicity* in the choice of generators, namely by computing *minimal representatives* of homology bases.
Minimal homology bases have long been investigated ([81, 36]), with a breakthrough only coming thanks to the introduction of a first efficient algorithm for the computation of bases in dimension one ([35]). Here, we leverage said minimal bases to propose a new approach to network skeletonization, the *minimal scaffold*, which overcomes the limitation of the previous one. While the minimal scaffold is not

unique in the most general case possible, we provide strong guarantees and caveats on when and to what degree it is well-defined. We then show a few applications of the novel method, concluding the chapter with a comparison between our and the previous construction.

4.2 Homological Scaffold

The homological scaffold originated from the intuition that traditional, graph-theoretical tools in network analysis were naturally able to capture significant properties ([6]), but proved not as effective in detecting multi-agent and large-scale interactions. Interest in searching for alternative descriptors of network relations arose, and soon works were published which leveraged invariants offered by computational topology ([74, 70, 84]).
In proposing the scaffold ([85]), the authors pointed out that homology might be able to summarize well network *mesoscale* structures, i.e., features living between the purely local connections and the global statistics, to which previous methodologies were blind. Furthermore, this structure could be analyzed over the continuous, full range of interaction intensities, without the need for ad-hoc domain-specific thresholds.
Homological cycles intuitively describe obstruction patterns. The presence of non-trivial homology within a given region of a network highlights its structure as non-contractible, binding signals to flow over constrained channels, which in turn play the role of bridges.
To test the method, the homological scaffold was computed from resting-state fMRI data for 15 healthy volunteers who were either infused with placebo or psilocybin: the scaffold discriminated the two groups, as well as providing meaningful insight as to the impact of the psychoactive substance onto the pattern of information flow in the brain [85].

Consider a non-negatively weighted finite graph $W = (V, E, w)$, where $w : E \to \mathbb{R}^+$. We can construct a filtration of simplicial complexes $\mathcal{F} = \{K_\varepsilon\}_{\varepsilon \in \mathbb{R}^+}$ in the following way. First, we take for each $\varepsilon \in \mathbb{R}^+$ the subgraph $(V_\varepsilon, E_\varepsilon)$, where $V_\varepsilon = V$ and $E_\varepsilon = \{e \in E \mid w(e) \leq \varepsilon\}$. Then, we define K_ε as the flag complex of $(V_\varepsilon, E_\varepsilon)$. Notice that, given a non-negatively weighted finite graph $W = (V, E, w)$, we can define an extended dissimilarity $\bar{w} : V \times V \to \mathbb{R}^+$, such that $\bar{w}(x, y) = w(x, y)$ if $(x, y) \in E$ and $w(x, y) = +\infty$ if $(x, y) \notin E$. Then, the filtration of flag complexes associated with W is equivalent to the Vietoris-Rips filtration of $(V, \bar{w})$
Once this filtration is obtained, we can compute its persistent homology. As in most applications, to have an easier interpretability of the results, we use the fields $\mathbb{Z}_2$ as the coefficient group for the homology functor. Let $\{b_i\}$ be a set of 1-dimensional

generator cycles of the persistent homology, i.e. representatives of each persistent homology class. Since we are over $\mathbb{Z}_2$, each of the b_i's is completely identified by its support, which is a set of edges of E. In particular, we can depict set $\{b_i\}$ as a matrix whose rows are indexed by E and having the b_i's as columns. The row sums, as natural numbers, form a new weighting function on the edges of W, the new weights counting precisely in how many persistent cycles an edge appears along the filtration.

Definition 50. Suppose W and $\mathcal{F}$ as above, and consider a set $\{b_i\}$ of 1-dimensional generator cycles of the persistent homology. Consider the function $h_W : E \mapsto \mathbb{R}^+$

$$h_W := \sum_i \mathbb{1}_{e \in b_i} \tag{4.1}$$

where by $\mathbb{1}_{e \in b_i}$ we denote the indicator function $E \mapsto \mathbb{R}^+$ such that $\mathbb{1}_{e \in b_i}(e') = 1$ if e' appears in b_i, and 0 otherwise.

Then the **homological scaffold** of W is the weighted graph $\mathcal{H}(W)$ such that

- its vertex set coincides with the vertex set of W
- its edge set $E_{\mathcal{H}}$ is a subset of the edge set of W, consisting of edges with nonzero value for h_W
- its weight function is the restriction of h_W to $E_{\mathcal{H}}$.

In accordance with the above definition, building the homological scaffold of a weighted network W is a method of *network compression* or *skeletonization*. The definition also implies that edge weights are assigned by the number of basis cycles the edge belongs to.

We provide an example, referring to Fig. 4.1. In panel (a), a filtration of simplicial complexes arising from a point cloud is depicted. At each step, highlighted in purple is a representative of a persistent cycle (i.e. of a bar in the barcode), each at the scale at which it is born.
In panel (b), the corresponding homological scaffold is represented: it amounts to taking the union of the cycles of panel (a), i.e. stacking generators of PH_1, each contributing unitary weight.

In the following, we shall sometimes refer to the homological scaffold as the *loose*, or *original* scaffold, to contrast it with the new definition of scaffold to follow.

As anticipated in the introduction, it is apparent that there is a substantial source of arbitrariness in this definition.
Several different representative cycles exist which form a basis of the persistent

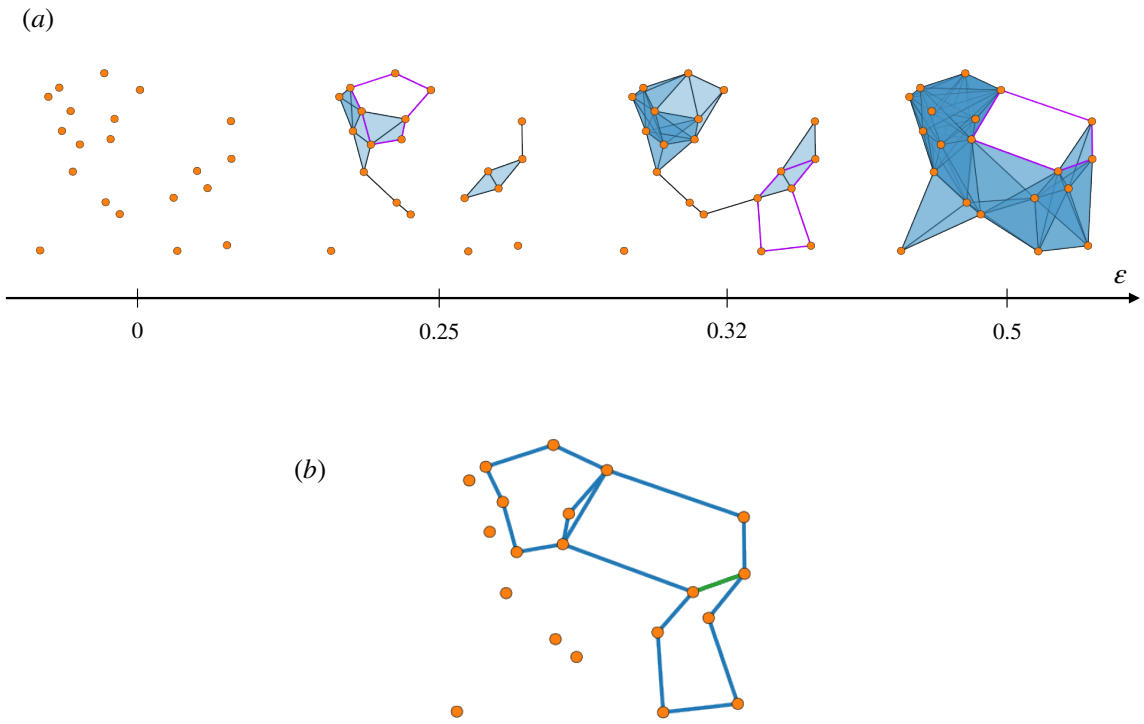


Figure 4.1: (a) A point cloud in $[0,1]^2$ and the generators of PH_1, plotted on the filtration step they appear at (scale reported on the axis below). (b) The resulting homological scaffold. Edges in blue have weight 1, each belonging to only one generator. The edge in green has weight 2, as it belongs to two generators.

homology (as a consequence of several different cycles belonging to the same homology class), and hence one must make a choice. For example, Fig. 4.2(a) depicts one specific cycle whose homology class generates (part of) the persistent homology group of the point cloud. At the same time, any other choice of edges forming a cycle around the hole is homologically equivalent and, in principle, legitimate.
In the original paper, the authors resorted to using the cycles as output by the *JavaPlex* implementation ([94]) of the persistent homology algorithm (based on the original implementation of [34]), and a posteriori checked the selected cycles for consistency. However, in principle, this means that the same simplicial complex written with two different orderings of the simplices could lead to different choices of generators, and therefore, to different scaffolds.
As such, we must be careful in the choice of nodes and edges output by the algorithm; while the presence of a generator denotes undeniably that an obstruction

pattern exists, we cannot be as confident about its precise location in the network or the constituents that provide bridges around it. The homological scaffold defined in this way introduces noise in the localization of mesoscale patterns onto individual nodes and edges, a process which, if accurate, could provide valuable insight as to the functional role of single players in a network.
In this work, we try to work around the problem of cycle choice and give a stricter definition, by requiring that, among all possible representatives, those of *minimal total length* are chosen (e.g., Fig. 4.2(b)).
The original algorithm reported a computational complexity of the order $O(n^3)$ to obtain representatives of basis cycles.

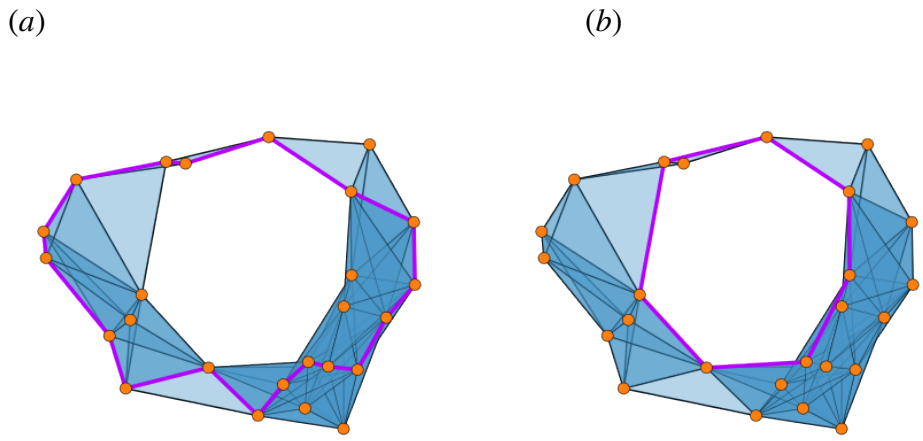


Figure 4.2: A simplicial complex K with $\dim H_1(K) = 1$. Its homological scaffold (on a subset of the filtration steps, for clarity) is reported in panel (a): the chosen generator meanders around the hole. Furthermore, a different ordering of the list of simplices fed to the algorithm could return a different cycle. In panel (b), the shortest representative cycle is chosen: this choice is stable with respect to any ordering of the input, while at the same time endowing the generator with some metric and geometric meaning.

4.3 Minimal Bases

The search for minimality in the computation of the scaffold was made feasible by the introduction of efficient algorithms to compute the minimal representatives

of a homology basis in dimension one.
It is known that in dimension higher than one, minimal representatives of a homology basis will remain elusive. Indeed, Chen and Freedman ([25]) proved that the problem of obtaining these minimal representatives is computationally intractable, being at least as hard as the notoriously NP-Hard Nearest Codeword Problem. Furthermore, it is even NP-Hard to approximate within any constant factor, meaning that no polynomial-time algorithm exists to obtain an approximate minimal basis that differs from the exact one by at most a multiplicative constant. In the light of this, we must necessarily restrict our attention to the 1-dimensional case, i.e., computing minimal representatives of a basis of H_1.

4.3.1 Minimal Bases and Dey's Algorithm

Given a simplicial complex K, let us consider C_1 the vector space generated by the 1-simplices of K and Z_1 the vector space of 1-cycles, i.e., $Z_1 = \ker \partial_1$.

Definition 51. Given a 1-cycle $b \in Z_1$, let $\mu(b)$ be its length, i.e., the sum of the weights of the 1-simplices that form it, and denote by $[b]$ the homology class b belongs to. Let $\beta_1 := \dim H_1(K)$. A *minimal homology basis* is a set of 1-cycles $\in Z_1$, of cardinality β_1, such that

$$\{b_1, ..., b_{\beta_1}\} = \underset{\text{Span}\{[b_i]\}=H_1}{\text{argmin}} \sum_i \mu(b_i). \tag{4.2}$$

The name minimal homology basis, comes with a slight abuse of terminology, as it would be more appropriate to call it a *minimally-represented homology basis*. In 2018, Dey et al. ([35]) introduced a polynomial-time algorithm to obtain said representatives. Building on the work of Horton ([58]), de Pina ([86]), and Mehlhorn et al. ([64]), the algorithm sets off to compute a basis of the space of cycles. Then, it applies a cohomological technique called *simplex annotation* ([17]) to lift a basis of cycles to a basis of the homology group H_1, while at the same time enforcing the minimal length constraint. A sketch of the algorithm follows.
Algorithm: MinBasis(K)

- A basis of the cycles group Z_1 is found via a spanning tree. Each edge in the complement of the spanning tree identifies a candidate cycle ([58]).
- An annotation of the edges is computed via matrix reduction ([17]). This yields the dimension β_1 of H_1, as well as an efficient tool to determine if two cycles b_1 and b_2 are linearly dependent in H_1 ($[b_1] = [b_2]$).
- A set of *support vectors* is generated which maintains a basis of the orthogonal complement in H_1 of the minimal basis cycles.

- Iteratively for each dimension of H_1, the candidate set of cycles is parsed in search of cycles b's that are linearly independent in homology from the previous ones (exploiting the support vectors). Among these, the μ-shortest one is added to the minimal basis.
- The set of support vectors is updated for the remaining dimensions to enforce it remain a basis of the orthogonal complement of the basis.
- The last two steps above are repeated until completion of the minimal basis.

Call $B = \{b_i\}$ the output of MinBasis on input K.
Theorem (3.1, [35]) Cycles in B form a minimal homology basis of $H_1(K)$.

Notice that the minimal homology basis is guaranteed to exist, as we only work with finite simplicial complexes, which implies the existence of a finite number of bases. However, it needs not, in general, be unique. Several different cycles of the same minimal length may all belong to the same homology class of a basis cycle. Heuristically, this is especially true in case the input complex is unweighted (equivalently, has equal weights for every edge), in which case the length of a cycle is the number of edges that form it. Furthermore, there exist cases when different sets of cycles of minimal length generate the same homology space, and are not even pairwise homologous. We will treat the problem of the uniqueness of the minimal basis in more detail in the following, and account for it explicitly in the construction of the minimal scaffold.

The computational complexity of the above procedure is evaluated ([35]) to $O(n^2\beta_1 + n^\omega)$ where n is the number of simplices in K and ω is the fast matrix multiplication exponent, which as of 2014 is bounded by 2.37 ([35, 30, 69]). This yields a worst-case complexity of $O(n^3)$ in the number of simplices for general complexes, which we recall is itself of order 3 in the number of points in the worst case.

4.4 Minimal Scaffold

In this section, we introduce an alternative definition for the homological scaffold, which we call minimal, based on the minimal representatives obtained above, and aims at overcoming the arbitrariness in the cycle choice of the previous definition. After addressing the simplest case, we analyze its uniqueness properties and introduce a second, more refined, definition.

Let $\mathcal{F}$ be the filtration of simplicial complexes induced by a non-negatively weighted finite graph W. For all filtration steps ε, define, as per (Eq. (4.2)),

$B^\varepsilon := \{b_i^\varepsilon\}$ the minimal homology basis of $H_1(K^\varepsilon)$. Take the disjoint union of minimal bases for ε varying on all filtration steps

$$B^* := \coprod_\varepsilon B^\varepsilon$$

Definition 52. Suppose W, $\mathcal{F}$ and B^* as above. Similarly to the loose case, define the function $h_{W,min} : E \mapsto \mathbb{R}^+$ as

$$h_{W,min} := \sum_{b \in B^*} \mathbb{1}_{e \in b} \tag{4.3}$$

Then, we define the **minimal scaffold** of W as the weighted graph $\mathcal{H}_{min}(W)$ whose:

- vertex set coincides with the vertex set of W
- edge set E_m is a subset of the edge set of W, consisting of edges with nonzero value for $h_{W,min}$
- weight function is the restriction of $h_{W,min}$ to E_m.

The minimal scaffold amounts, again, to the stacking of generator cycles across a filtration. However, two differences are to be noted with respect to the loose definition. First, we require the representative cycles to be minimal. Second, we point out that while the loose scaffold is built by aggregating the generator cycles of $PH_1(\mathcal{F})$, the minimal scaffold is built by independently computing a minimal basis for each $H_1(K^\varepsilon)$, for all ε. Notice that, since cycles are modified throughout a filtration, it would be meaningless to talk about a minimal representative over a certain persistence interval. This also means that its computation can be effectively parallelized by assigning different filtration steps to different jobs, and later recombining the outputs.
An interesting phenomenon that descends directly from the above peculiarity is that the minimal scaffold of random point clouds tends to display a more pronounced triangular structure (clustering) around cycles. Indeed, as longer (or, in non-metrical filtrations, later) edges are introduced, a cycle can be shortened (by the triangular inequality) by a longer edge which cuts a corner. Since at each step the algorithm records the minimal representative, upon aggregating the minimal scaffold one finds each cycle in its progressively shorter version, and the *history* of the shortening is visible as a padding of triangles around it.

Considering the example of Fig. 4.3, in panel (a) we observe an example of a filtration of simplicial complexes. At each step, highlighted in purple we may see the minimal representative of a homology class, together with its evolution history. At filtration value 0.26, we observe a pentagon being reduced to a shorter,

quadrilateral cycle by the addition of a longer edge. This is an example of the phenomenon explained above. Fig. 4.2 gives a visual description of the difference between a minimal and generic cycle.
The union of these progressively shorter cycles for all steps (weighted according to Eq. (4.3)) is the minimal scaffold, as seen in Fig. 4.3 panel (b).

We remark that, if there is no ambiguity in the construction of a filtration of simplicial complexes from a point cloud, or from a weighted graph, we will indifferently speak of the scaffold as a function of either of them ($\mathcal{H}_{min}(C)$, or $\mathcal{H}_{min}(W)$, or $\mathcal{H}_{min}(\mathcal{F})$).

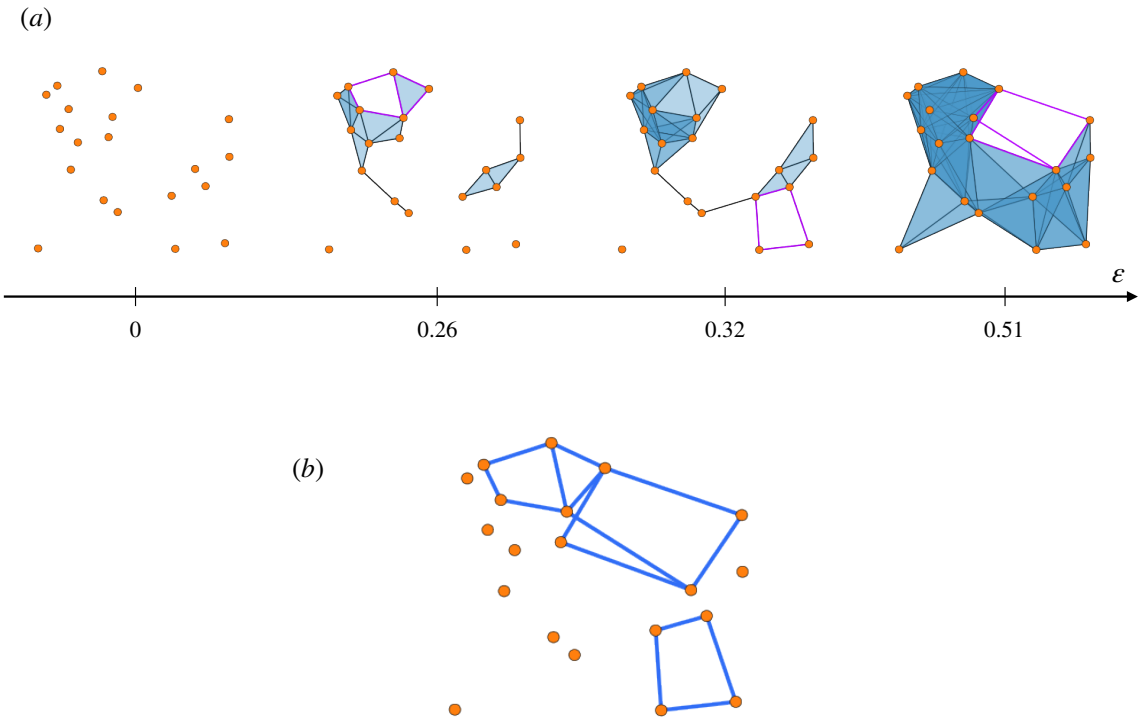


Figure 4.3: (a) The same point cloud of Fig. 4.1. Along the filtration we show the evolution of minimal generators, which can get progressively shorter as new edges are introduced. For example, at $\varepsilon = 0.26$, the pentagonal cycle gets cut to a shorter quadrilateral, albeit with an individual longer edge. This evolution is accounted for in the minimal scaffold, which displays the triangle-rich structure mentioned above. (b) The resulting minimal scaffold (weights not reported).

We have mentioned that the scaffold amounts to a change in weighting in the input graph

$$h_{W,min} \ : \ E \mapsto \mathbb{R}^+$$

altering the original weights of the edges. Additionally, considering *node strength* (i.e. the sum of the weights of the edges incident to a given node), it can equally be considered as a function

$$\mathcal{H}_{min} \ : \ V \mapsto \mathbb{R}^+$$

assigning weights to nodes. Considering the reliability of the choice of edges in the procedure, this explains why the minimal scaffold can be utilized to associate mesoscopic features with single nodes and links.

Computational Complexity

For large input sizes, the cost of assembling the minimal basis cycles into the scaffold is negligible with respect to the cost of computing such minimal basis. We know that each run of Dey's algorithm costs $O(|K|^3)$ in the worst case ([35]), and in the worst case $|K|$ is itself $O(n^3)$ where n is the number of points.
The number of filtration steps has an upper bound of $O(n^2)$ (i.e., the number of edges) in the worst case, as in general every edge may carry a different weight. Hence Dey's algorithm has to be run once for each edge in the worst case.
This yields a theoretical worst-case complexity of order $O(n^9 n^2) = O(n^{11})$. Therefore, while the minimal scaffold is undeniably a polynomial-time algorithm, its practical computation is often hindered by its dire lack of scalability, especially if compared against the loose version, which has a far more favourable complexity.
A comparison of running times is carried out in Fig. 4.4, which clearly shows that computing the minimal scaffold on an ordinary machine can quickly become troublesome.

Implementation

We have written a Python implementation of Dey's algorithm, together with a library for the computation of the minimal scaffold. The code is available on GitHub at [54], with some usage examples. It allows for shared-memory multi-threaded parallelism across filtration steps to improve computation times, while still being suitable for ordinary desktop workstations.

4.5 Uniqueness of the minimal scaffold

The uniqueness of the minimal scaffold depends on the uniqueness of the minimal basis. Indeed, if there exists only one possible set B^* of cycles forming a

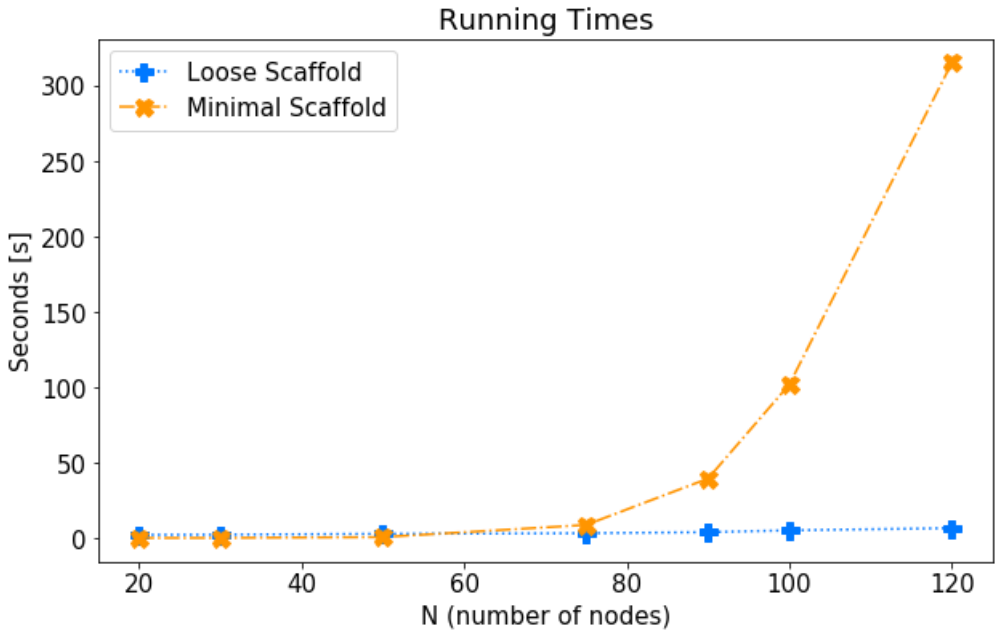


Figure 4.4: The running times of computing the minimal and loose scaffolds for Watts-Strogatz weighted random graphs. For all instances, number of nodes N is indicated on the x-axis. Number of stubs k is $N/2$, and rewiring probability is $p = 0.025$.

minimal basis, then the scaffold is uniquely determined. Two issues affect the uniqueness of set B^*.

Draws

The first one arises when two or more different and homologous basis cycles are of the same minimal length. This case is relatively simple to work around: we modify the definition of minimal scaffold to keep track of all variants of minimal basis cycles, dividing the weight equally among them.
Specifically, to account for this issue we have slightly modified Dey's algorithm. In its last step described above, one is concerned with finding all cycles whose annotation is not orthogonal to the given support vector: among these, the one with minimal length is chosen as a basis cycle. Instead, we keep track of *all* such cycles with the same minimal length. This does not alter the complexity, as one needs to check all possible cycles anyway. We call this case a *draw*. Therefore, we modify set B to become a set of *sets of cycles*.

Definition 53. Given complex K, we define a minimal basis *with draws*

$$\tilde{B} := \bigcup_{i=1}^{\beta_1(K)} \{b_{i,1}, ..., b_{i,n_i}\}$$

where for all $i = 1, ..., \beta_1(K)$, the cycles $b_{i,j}$ with $j = 1, ..., n_i$ are homologous and have the same minimal length. Furthermore, for every choice of $j_i \in \{1, ..., n_i\}$, $\text{Span}_i\{b_{i,j_i}\} = H_1(K)$. Call $V_i := \{b_{i,1}, ..., b_{i,n_i}\}$ each set of draws, i.e., *variants* of the i^{th} minimal basis cycle, $\forall i = 1, ..., \beta_1(K)$.

In the example of Fig. 4.5(a) and (b), we have set $\tilde{B} = \{ \{b_{1,1}, b_{1,2}\} \}$, whereas set B might have indifferently been equal to $\{b_{1,1}\}$ or to $\{b_{1,2}\}$, whichever happened to come first in the search. The minimal scaffold is modified accordingly.

Definition 54. Given the usual filtration $\mathcal{F}$, let $\tilde{B}^\varepsilon$ be the minimal basis with draws of $H_1(K^\varepsilon)$. Again, we aggregate all variants of minimal basis cycles along the filtration

$$\tilde{B}^* := \coprod_\varepsilon \tilde{B}^\varepsilon$$

Then, we define the weighting function *with draws* $\tilde{h}_{W,min} : E \mapsto R^+$

$$\tilde{h}_{W,min} := \sum_{V \subset \tilde{B}^*} \frac{1}{|V|} \sum_{b \in V} \mathbb{1}_{e \in b} \tag{4.4}$$

and the resulting *minimal scaffold with draws* $\tilde{\mathcal{H}}_{min}(W)$ is built from $\tilde{h}_{W,min}$ as in Eq. (4.3).

The meaning of the above definition is that all variants of all minimal basis cycles are taken into account when building the scaffold, and the weights are assigned dividing each variant's contribution by its cardinality, for each filtration step. In the example of Fig. 4.5(c), the two cycles forming the variant of the only generator are multiplied by a factor of $\frac{1}{2}$ and then summed: therefore, common edges outside the diamond are assigned weight 1, consistently with the minimal scaffold in (Eq. (4.3)), whereas the four edges forming the perimeter of the diamond each get assigned weight $\frac{1}{2}$.

With the introduction of draws, we settle the case when ambiguity arises among individual cycles, without interactions. As an example, we can state the following result.

Proposition If $\mathcal{F}$ is such that, for all ε in the filtration, each basis cycle belongs to a different connected component of K^ε, then the minimal scaffold with draws $\tilde{H}_{min}(\mathcal{F})$ is unique.

Pathological cases

The other issue arises when there exist sets of minimal cycles that are representatives for homology classes that are not linearly independent. Suppose that three different cycles generate a homology group of dimension two, i.e., when three minimal cycles are pairwise independent in homology, but threewise dependent. In this case, two generators are sufficient to span H_1 and, if their lengths are arranged pathologically, there is no principled way to choose two out of the three.
Suppose for example that three cycles b_1, b_2 and b_3 are such that

$$\mu(b_1) < \mu(b_2) = \mu(b_3) \quad \text{and} \quad [b_1] = [b_2] + [b_3]$$

In this case, both bases $\{b_1, b_2\}$ and $\{b_1, b_3\}$ span the same homology space, and are of equal minimal length. The minimality criterion fails in this case.
One could believe that such a configuration can only happen in the most general spaces, and that by imposing some mild hypotheses on the input data one could rule the pathology out. In fact the opposite is true, this degeneracy being possible even after enforcing very strong conditions on the data.

Counterexample Even if W is planar and an isometric embedding $W \hookrightarrow \mathbb{R}^2$ exists (i.e., the input planar weighted graph can be accurately drawn onto the plane), the minimal scaffold $\tilde{H}_{min}(W)$ needs not be unique.
In fact, consider complex K arising from the geometric, planar graph in Fig. 4.5(d). Its homology $H_1(K)$ is generated by two cycles; possible generators are depicted in Fig. 4.5(e). Since the outer cycle b_1 is the shortest, and the two inner ones b_2 and b_3 are of equal length, the minimality criterion can not solve between $\{b_1, b_2\}$ and $\{b_1, b_3\}$, as both are acceptable minimal bases. The minimal scaffold (with or without draws) is not unique in this case.

Clearly, the same could happen with more than three cycles, with a larger number of possibly ambiguous configuration. Therefore, if we allow for a high degree of symmetry in the input, this pathology could arise even in the rather tame context of planar graphs on $\mathbb{R}^2$. A similar pathology may arise when the weights of the graph are integer numbers and there are several cycles of the same length. This issue is rather delicate, in the sense that not only the algorithm is unable to make a principled choice; it is not even capable of detecting when such a configuration takes place. In fact, this is more of a feature of homology than a flaw in the skeletonization framework: what our eyes see as different cycles are in fact homologically equivalent, and it is impossible to use homology to tell them apart.

We however remark that, for complexes arising from real-world data, this type of configuration is actually pathological. Indeed, the following generality result holds

Proposition Assume a point cloud $C = \{X_i\}$ such that the X_is are drawn independently from a uniform distribution over $[0,1]^d$. Then, almost surely, the minimal scaffold $\mathcal{H}_{min}(W)$ (with or without draws) is unique.

If the input point cloud is sampled uniformly at random in some $\mathbb{R}^d$, then edge lengths are distributed according to an absolutely continuous probability law. Therefore, given two edges e_1 and e_2, $\mathbb{P}[\mu(e_1) = \mu(e_2)] = 0$. The same holds for any two non-identical cycles, and any two homology bases (being but finite sets of edges): the probability of them sharing the exact same length is zero. By finiteness of the input, at least one minimal homology basis exists and, by the above reasoning, almost surely this basis is unique for each filtration step. Then, with probability 1 the minimal scaffold is unique.

This result is actually quite general: whenever we can assume our input data to be subject to noise, then we are in principle allowed to rule out pathological same-length cycles. In these cases, the minimal scaffold is unique.

We remark that this uniqueness result is compatible with the phenomenon of the concentration of measure: while for a very high-dimensional space or a very large number of points we know from theory that the distribution of length of edges concentrates towards its mean value, the probability of two edges (and hence two cycles) having the same length is still zero. One needs to be careful, however, that the probability of two cycles differing in length by less than some $\epsilon > 0$ could grow very rapidly with ϵ.

In summary, the minimal scaffold with draws $\tilde{\mathcal{H}}_{min}$ is well-defined up to some pathological circumstances, where it may depend on the ordering of the input.

4.6 Applications

As illustrative examples, we show here a few applications of the minimal scaffold. Through it, we obtain meaningful subsets of known networks in neuroscience, and rank their constituents by their “topological importance".

The C. Elegans dataset is a correlation network of neural activations of the nematode worm Caenorhabditis Elegans. C. Elegans has become a model organism due to the unique characteristic of each individual sharing the exact same nervous system structure.
The input consists of a symmetric weighted adjacency matrix over 297 nodes, each

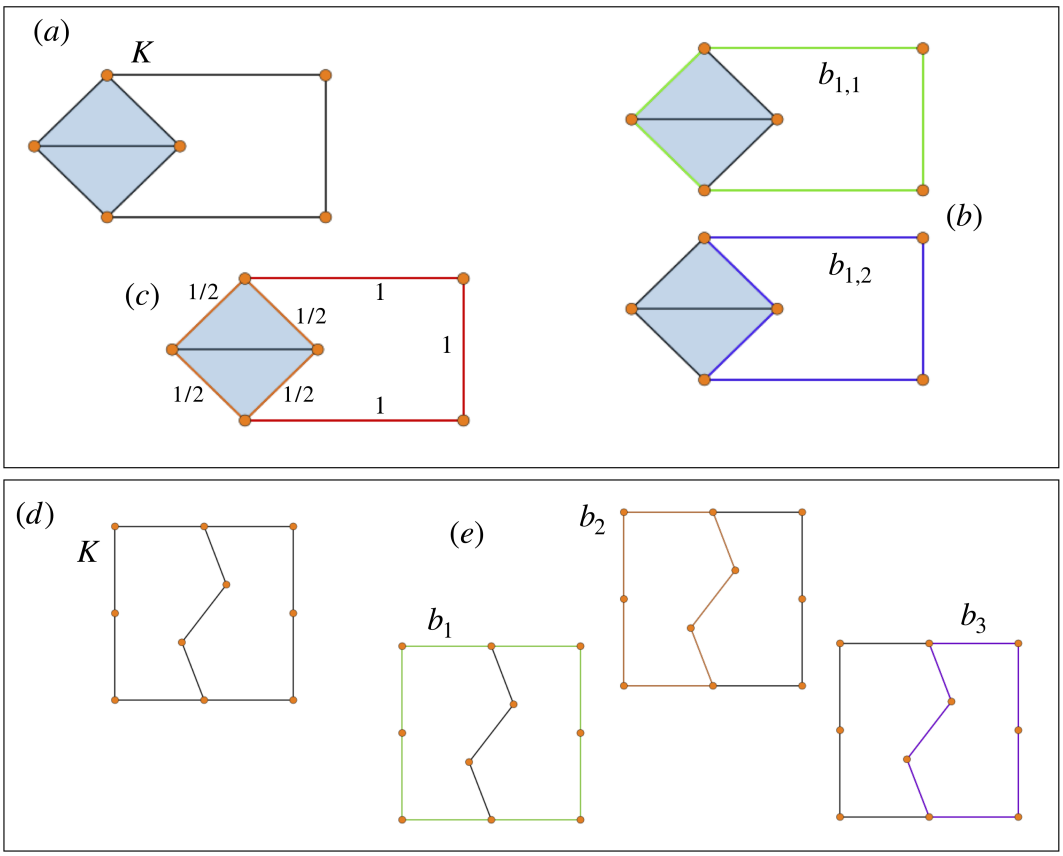


Figure 4.5: Top panel: (a) A simplicial complex K. (b) Two homologous and equally minimal generators of $H_1(K)$. (c) The minimal scaffold with draws $\tilde{\mathcal{H}}_{min}(K)$. The weight is equally divided among the variants of the minimal representative. Bottom panel: (d) A simplicial complex K on the represented point cloud. $H_1(K)$ has dimension 2. (e) $\mu(b_1) < \mu(b_2) = \mu(b_3)$. A minimal basis can either be composed of $\{b_1, b_2\}$ or $\{b_1, b_3\}$, hence it is not unique.

representing a neuron. Edge weights represent (quantized) time correlations between the firing of neurons, ranging from 1 to 70.
The minimal homological scaffold of its brain map highlights the *geometry* of the obstruction patterns, i.e., the precise areas where nervous stimuli are less likely to flow. We stress the improvement obtained by the minimal scaffold over the loose one, in that it is not only able to identify the *presence* of a “grey area" in the network, but it can as well provide a reliable boundary for it, and identify which neurons and inter-neuron links are responsible for information flowing around the obstruction.
As an interesting example, we see in Fig. 4.6 the top 25 neurons ranked in descending order of relative node strength (sum of weights of incident edges) with respect

to the average node strength. We can identify four nodes, labeled 81, 260, 36, and 37, which hold a significantly higher relative strength than the rest. This implies their presence in many minimal cycles across several scales, hence suggesting that they play a crucial role in the fabric of information flow within the nematode's brain.

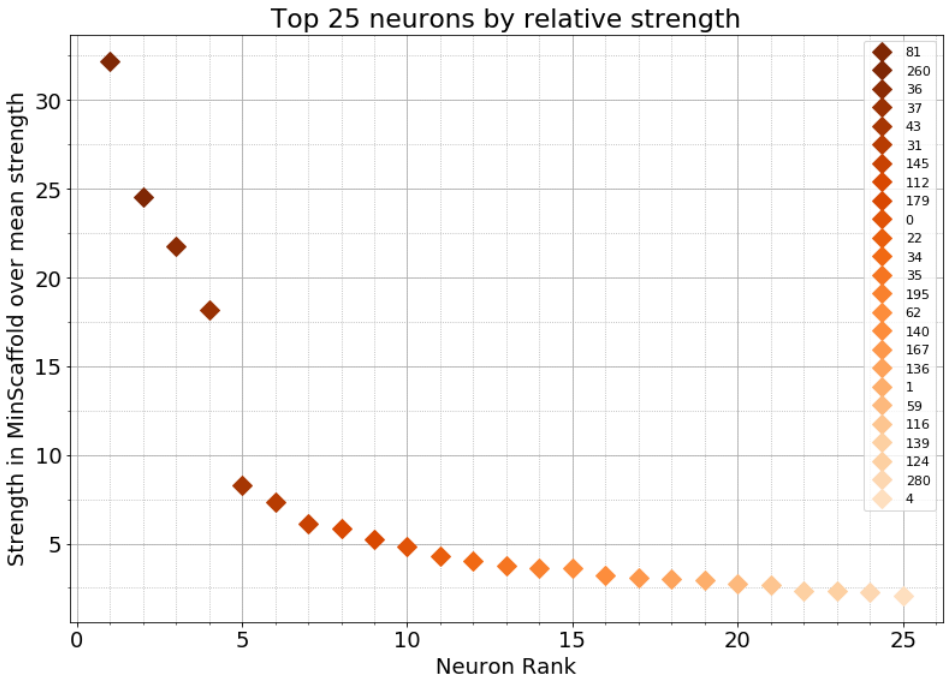


Figure 4.6: The top 25 neurons by relative node strength in the minimal scaffold over average strength in C. Elegans (mean 36.41). Four neurons show a significantly higher relative strength than the others.

The same type of analysis was repeated on the correlation network of brain activities in an 88-parcel atlas of the human brain, obtained through fMRI imaging at resting state. The data is courtesy of the Human Connectome Project ([75]).
Again, the minimal scaffold identifies which regions and links in the human brain are key bridges for the flow of information. Two parcels stand out (Fig. 4.7(a)) as particularly relevant for network topology.
For a relatively small network such as this, we can visualize the scaffold as a proper subnetwork by a chord diagram (Fig. 4.7(b)), with edge weight represented by color intensity and node strength by the size and color of the vertex. We stress that, starting from a virtually complete graph over 88 nodes, we reduce the size from 3828 edges to just 191, while preserving the topological structure.
We can, as well, leverage libraries in computational neuroscience ([1]) to embed the scaffold in the actual human brain, with regions correctly located, projected on the three coordinated planes. In Fig. 4.7(c), for visualization purposes color intensities

represent log-weight in the scaffold.
To better highlight the value of the scaffold in signalling brain network function, we constructed a suitable null model of the functional network, as was done in [76]. The technique consists in reshuffling the correlation matrix subject to the constraint of keeping a fixed spectrum, i.e. applying a random rotation, which guarantees the matrix remains positive semidefinite and hence a proper correlation matrix. An implementation of such a procedure can be found in [32].
The resulting randomized adjacency matrix is characterized by a vastly larger number of homological cycles than the original; so much so in fact that the computation of its minimal scaffold becomes cumbersome. However, even without computing them explicitly, we know for sure that the scaffolds of the original and randomized networks are totally different, specifically because they are built by aggregating two completely different persistence structures, i.e. the minimal scaffold does indeed highlight the functional information in the original dataset.

The possible applications in which the minimal scaffold could provide novel insight into the structure of brain data are many: any relatively small correlation matrix could be either compressed or its patterns analyzed, as is often the case in EEG [67, 61, 93, 60] or neuronal [50] studies, and in fMRI ones when using rather coarse atlases (e.g. [95, 5]).

4.7 Comparison of Scaffolds

As the last contribution for this work, we consider a comparison between the minimal and loose scaffolds.
We have already pointed out that the minimal scaffold in general offers superior guarantees as a tool, both for network analysis and network skeletonization. On the other hand, the loose scaffold clearly has an advantage in terms of computational complexity: while it is in principle viable for most of the applications where persistent homology has been employed, the minimal scaffold, even adopting filtration-wise parallelization, requires a vastly larger amount of computational power, which effectively limits its range of application, unless run on dedicated, high-performance infrastructures.
A reasonable question to ask is the following. If one is interested not in the exact structure of the scaffold, but only in its statistical behaviour, could the loose scaffold provide a sufficient approximation of the minimal one? In a more concrete example, if instead of wondering exactly which nodes in a network are the most topologically important one is interested in the distribution of the degree sequence of the minimal scaffold, could the loose one come to one's help?
To answer this question, we have performed comparisons of several graph metrics in the two scaffolds of C. Elegans. Further, to gain insight into the general case, we

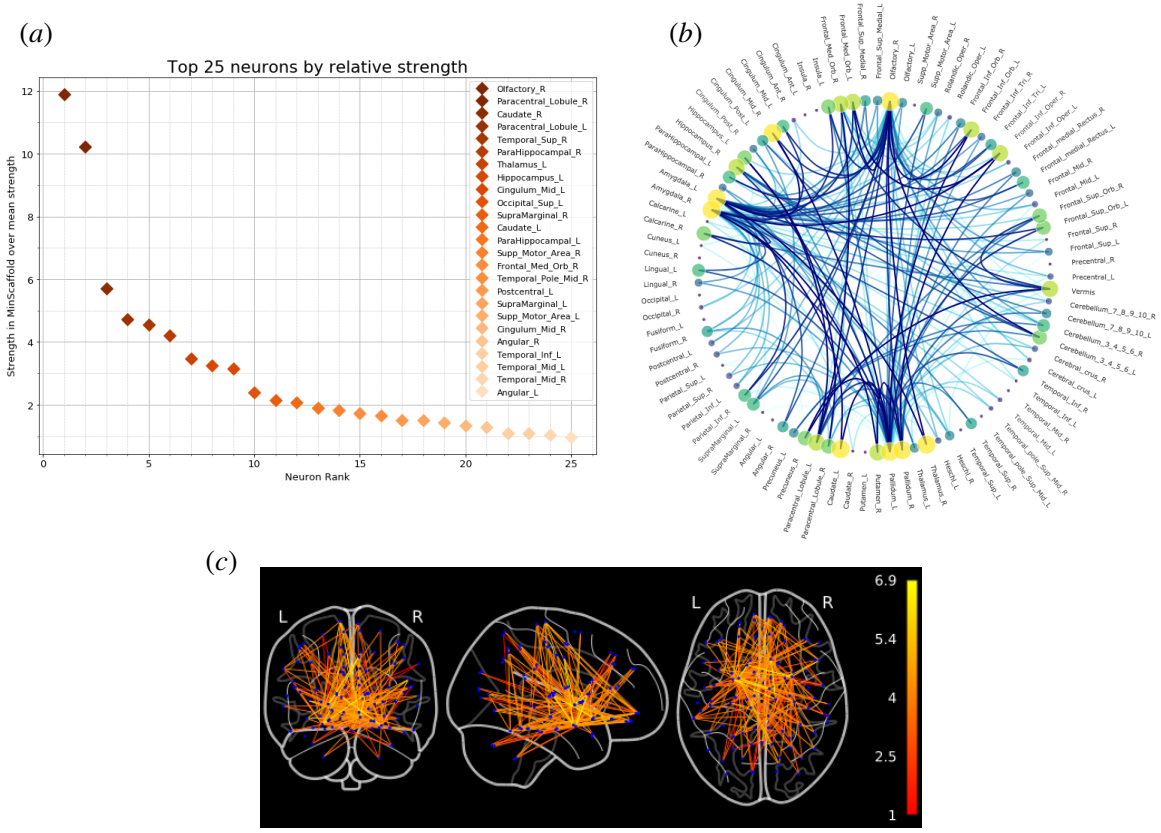


Figure 4.7: (a) The top 25 brain regions in the human brain by relative node strength in the minimal scaffold over average strength (mean 546.7). Two neurons show significantly higher importance. (b) The chord diagram of the minimal scaffold. Node size represents node strength, edge color intensity represents weight in the scaffold. (c) The minimal scaffold embedded in the human brain, with regions accurately located, projected on the three coordinated planes. Edge color represents log-weight in the minimal scaffold (Log-scale for visualization purposes).

have sampled two families of random graphs at different parameter values, one for geometric graphs (Random Geometric Graph), and one for non-geometric graphs (Weigthed Watts-Strogatz).

C. Elegans

For the C. Elegans dataset, we have compared the following graph metrics of the minimal and loose scaffolds:

1. Degree Sequence

2. Node Strength
3. Betweeness Centrality
4. Closeness Centrality
5. Eigenvector Centrality
6. Clustering Coefficients
7. Edge weights

Results (reported in the Table of Fig. 4.8(c)) indicate that, for metrics 1 to 5, the two scaffolds are very well correlated. So for example the cheap, loose scaffold is a reliable proxy of the distribution of the "true" degree sequence (scatterplot in Fig. 4.8(d)).
We instead observe poor correlation of edge weights and clustering coefficients. The first one is not unexpected, since the edge weighting procedure is conceptually different in the two scaffolds: while in the minimal one we consider a different basis for each filtration step, the loose scaffold considers bases of the persistent homology space, drastically reducing the number of cycles considered. To make it clearer, in general set B^* has cardinality much larger than the dimension of PH_1. It is therefore explicable that the distributions of edge weights do not generally agree.
Clustering coefficients, on the other hand, are a measure of how "triangular" a graph is around a given node. As remarked in Section 4.4, another consequence of assembling the scaffold from the minimal bases of the H_1's is that a large number of artificial triangles appear around cycles. In this case too, therefore, the poor correlation is easily explained.

Random Graphs

Drawing inspiration from [91], we repeat the analysis on random graph samples. [91] divides random networks into two categories: those created from edge weighting schemes and those created from points in the Euclidean space. We have chosen to analyze the weighted Watts-Strogatz (WS) model as representative of the first class, and the geometric random model as representative of the second. We remark that weighting needs to be introduced in order to compute persistence; while for geometric graphs this simply requires computing the Euclidean distance, for the Watts-Strogatz model it requires an ad-hoc procedure that is described in detail in the supplemental material of [91].
We briefly recall that a WS graph is parametrized by the number of nodes, by the number of stubs to rewire, and by the rewiring probability. A random geometric graph is instead parametrized by the number of points to sample (uniformly) in

$[0,1]^d$, and by a cutoff value that acts as distance threshold, beyond which no edge is introduced.
In both cases, we observe good agreement on key statistics, as reported in Fig. 4.8(a) and (b). Each bar is obtained by computing the correlation of the reported statistic on a sample of 30 random graphs of the reported model, with parameters as indicated on the x-axis.
For comparison, two null models are built for each instance of the minimal and loose scaffolds in the sample, by constructing an Erdős-Rényi random graph on the same vertex set, one with the same number of edges as the minimal scaffold, and one with the same number as the loose one. The correlation is computed of each statistic between the minimal scaffold and the loose null model and between the loose scaffold and the minimal null model. The average of these correlations is reported on the boxplots to act as a baseline value, highlighting that the two scaffolding procedures agree with each other by more than just statistical noise.
For a finer analysis, we have performed a two-sample Kolmogorov-Smirnov test comparing the distribution of the given metrics in the minimal and loose scaffolds, for all parameter values of the two random models. We consider the Kolmogorov-Smirnov test to be inconclusive if its p value exceeds a threshold of 0.05, in which case one cannot confidently reject the null hypothesis that the samples are drawn from the same distribution. In Fig. 4.8 panels (a) and (b), the darker boxes report for each parameter choice and metric the fraction of samples for which the test was inconclusive: in all cases except one, the KS test could not distinguish between the distribution of the graph statistic between the minimal and loose scaffolds, strengthening the indication of a good agreement between the two.

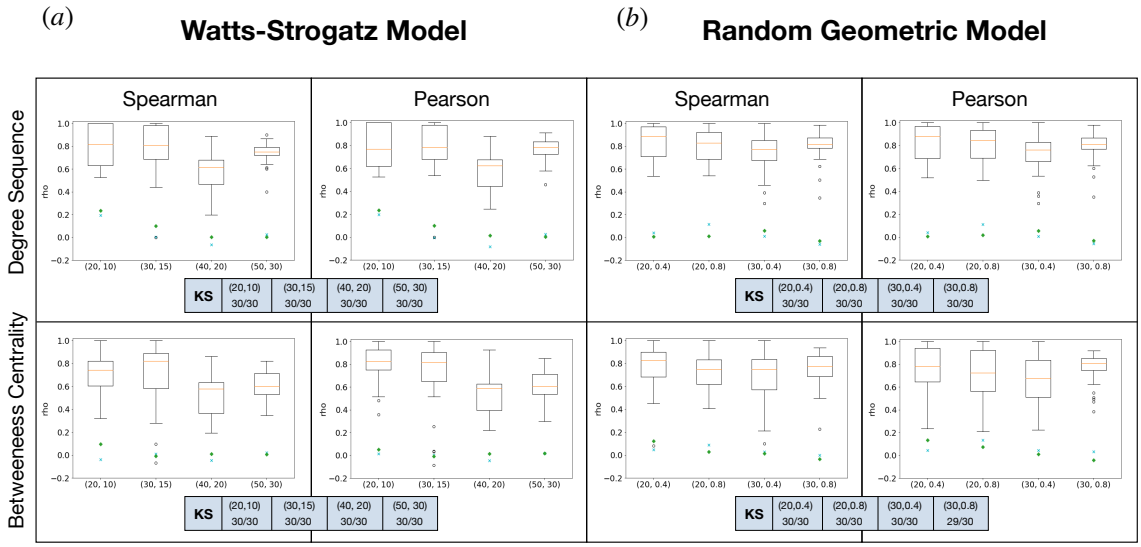
nPSO Random Graph Model

A modern random graph model, which has recently gained traction in network science for its ability to concurrently tune several parameters of interest in modeling real networks, is the Nonuniform Popularity-Similarity model. Introduced in [79], it builds upon a sequence of increasingly refined generative models to provide all the key structural properties of real-world graphs, such as scale-freeness, small-worldness and community structure. We therefore set out to employ it as benchmark in our comparison of the minimal and loose scaffolds.
In general, networks which display hyperbolic geometries tend to have a rather tree-like structure, with a certain scarcity of cycles. It is straightforward that, in the absence of a significant structure of persistent homology, the loose and minimal scaffolds will agree to high degree for at least two reasons: the low number of cycles forces the loose scaffold to localize onto the few available holes, hence resembling the minimal, and secondly the scarcity of homology makes for a comparison between two mostly empty sets.

Following the lead of [3], we tuned the nPSO model parameters in order to empirically maximize the persistent homology structure, so as to make the comparison the most significant possible. As reported in Fig. 4.9, we observe again good ability of the scaffolds to proxy each other across the metrics analyzed, significantly higher than with respect to a null model, for a sample with parameters $N = 50, m = 2, T = 5, \gamma = 3$ and uniform distribution. A Kolmogorov-Smirnov test was also performed, as in the previous section, where a p-value higher that 0.05 indicates that the distribution of degrees and betweeness centralities in the minimal and loose scaffold cannot be confidently distinguished. This was the case for all the samples we tested.

4.8 Conclusions

We provided a new method of network analysis and skeletonization, based on the computation of minimal homology bases. This new construction fills a significant gap in previous literature, in that it yields, in all but some pathological cases, a well-defined and unique subgraph, acting as a reasonable ground truth for comparison with the previous construction. It can be employed in a range of applications, both to identify crucial and weak links in a network, and to obtain compressed and topologically sound representations of the input. It also allows to evaluate the reliability of other scaffolding procedures with respect to said ground truth: we have observed that, for some applications, the loose scaffold can be deemed a sufficiently accurate tool, while not incurring in as cumbersome a computational load.
We foresee that the subject of homological skeletonization is not yet concluded. Other approaches to finding canonical generators of homology are possible (for example in [68] and [12]), and we plan to investigate them further in subsequent works.
A question which remains open and could be worthy of further work is the following: could one construct a sensible "entropy" functional on the space of cycles, so as to obtain a strictly unique, minimally-represented basis that is in the most likely?



(c)

	Spearman		Pearson	
Metric	Corr	p-val	Corr	p-val
Node Degree	0.953148	3.1842e-155	0.975559	3.4463e-196
Node Strength	0.772330	4.3712e-60	0.700653	3.7250e-45
Betw. Centrality	0.952098	7.7348e-154	0.986412	1.8813e-233
Closeness Centrality	0.921274	5.1143e-123	0.960413	8.7695e-166
Eigenvector Centrality	0.880711	9.5943e-98	0.858564	1.3911e-87
Clustering Coefficients	0.412889	1.1778e-13	0.358577	1.9337e-10
Edge Weights	0.226321	1.3586e-09	0.086226	0.0224

(d)

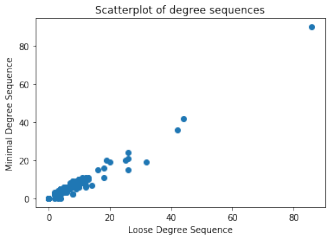


Figure 4.8: Correlations between the minimal and loose scaffold. (a) Comparison in the weighted Watts-Strogatz model. Degree sequence and betweenness centrality in the two scaffolds are compared, using Pearson and Spearman correlation coefficients. Each box is computed over a sample of 30 weighted Watts-Strogatz random graphs, with parameters as reported on the x-axis: the pair (N, k) indicates a WS model on N nodes, with k stubs to rewire. The rewiring probability is 0.025. The cyan crosses and the green diamonds represent the average correlation value against the loose and minimal null models, respectively. (b) Comparison in the random geometric model. Each box is computed over a sample of 30 random geometric graphs, with parameters as reported on the x-axis: the pair (N, t) indicates a graph on N nodes sampled uniformly at random in the $[0,1]^2$ square. t is the connectivity distance threshold. The cyan x's and the green diamonds represent the average correlation value against the loose and minimal null models, respectively. The darker boxes in panels (a) and (b) report, for their respective model and for each metric and parameter values, the fraction of the sampled instances for which the Kolmogorov-Smirnov test was inconclusive (p value > 0.05). (c) Correlation tests for several network metrics on the C.Elegans network. (d) Scatterplot of the degree sequence of neurons of C. Elegans in the minimal scaffold versus in the loose one.

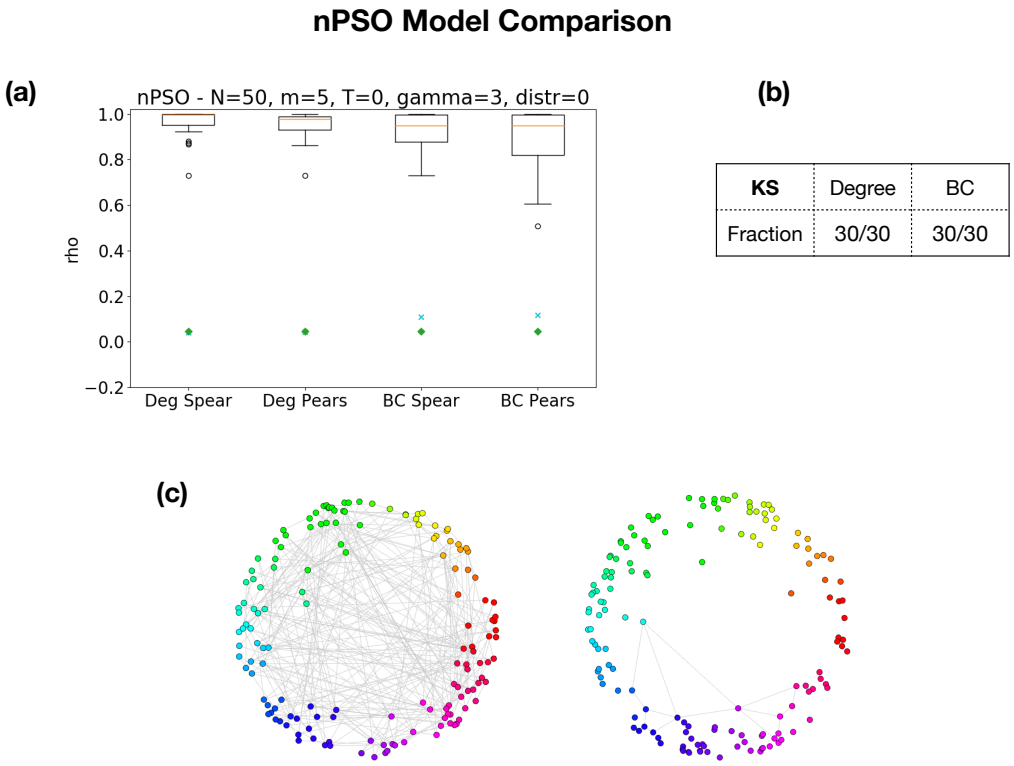


Figure 4.9: Comparison of the minimal and loose scaffold for nPSO random model. (a) Degree sequence and betweenness centrality in the two scaffolds are compared using Pearson and Spearman correlation coefficients. Each box is computed over a sample of 30 nPSO instances, with the following parameters: 50 nodes, average degree 10 ($m = 5$), 0 temperature, power-law exponent $\gamma = 3$, and uniform distribution of angular coordinates. The cyan crosses and green diamonds represent the average correlations against the loose and minimal null models respectively, as in Fig. 4.8. In panel (b), the table reports, for the degree and betweenness centrality distributions, the fraction of Kolmogorov-Smirnov test that could not reject the hypothesis of the two samples coming from the same distribution. This has always been the case for each sampled instance and both metrics. (c) A graphical depiction of an instance of the nPSO model with parameters $N = 150, m = 2, T = 5, \gamma = 3$ and uniform distribution on the left. On the right, the corresponding minimal scaffold.

Chapter 5

Landscapes of data sets and functoriality of persistent homology

5.1 Introduction

This chapter comes from a joint work with Wojciech Chachólski, Nicola Quercioli and Francesca Tombari which resulted in the paper [22]. The purpose of this Chapter is to investigate a new way to look at data sets, in order to enlighten some of the characteristics given by the symmetries of a problem. We will consider data sets as finite sets of real valued functions on a finite set X, which will be called *domain* of the data set. Each datum can be seen as a measurement performed on the domain. Even at this very basic level, we can see that a data set enriches us with knowledge about its domain. Since the data set can be seen as a vector valued function $\Phi : X \to \mathbb{R}^n$, where n is the number of functions in the data set, the pullback of a metric on $\mathbb{R}^n$ through Φ endows X with a pseudo-metric that will make it possible to extract non-trivial homological information in form of persistent homology. A single measurement does not contain any higher non-trivial homological information. Sets of measurements however do. Thus it is essential that measurements, on a given set X, are grouped together to form various data sets. In this case persistent homology becomes a non-expansive (1-Lipschitz) function $PH_d^{\Phi} \colon \Phi \to \mathbf{Tame}\left(\mathbf{Vect}^{[0,\infty)\times\mathbb{R}}\right)$, assigning to each measurement in the data set Φ a tame persistence module parametrized by $[0,\infty) \times \mathbb{R}$. It is important to notice that the choice of a set of measurements on X affects the pseudo-metric defined on it. One can use this fact to change the metric on X in order to extract more meaningful information from persistent homology. Fore example, consider the domain $X = \{(\cos(\frac{2\pi t}{4n}), \sin(\frac{2\pi t}{4n})) \mid t = 0, \dots, 4n-1\}$ and the data set $\Phi = \{\phi\}$, where $\phi : X \to \mathbb{R}$ is defined as $\phi(x, y) = x$. The persistent homology of ϕ in degree

greater than 0 is always trivial, since all the homology groups are the trivial vector space. Let us consider $g : X \to X$, the rotation by 90 degrees, $g(x, y) = (-y, x)$. If we add the measurement $\phi \circ g$ to Φ, the pseudo-distance d_Φ changes, and there are values of $\varepsilon > 0$ such that $H_1\mathrm{VR}_\varepsilon(X, d_\Phi)$ is non-trivial. Therefore, under the addition of $\phi \circ g$ to Φ, the persistent homology of ϕ is non-trivial in degree 1. This illustrates how our knowledge of an object is affected by the number and the type of measurements done on it. We remark that in this example we did not add a "brand new" measurement. The data set is enlarged using its elements and their composition with an endomorphism of X. If we consider the action of such endomorphisms on the data set, it is possibile to inject geometrical features of our choice on the data set. For exhibiting and extracting interesting homological features of data sets, such actions are therefore important.

A data set Φ is naturally equipped with an action of the monoid of its operations $\mathrm{End}_\Phi(X)$, which are endomorphisms of X preserving Φ. This action gives the set Φ a structure of Grothendieck construction, that we will summarise with what we call a *Grothendieck graph*. Persistent homology turns out to be a functor indexed by this graph, rather than simply a function. Thus, not only persistent homology can be assigned to individual measurements in a data set, but operations can be used to compare persistent homologies of different measurements. That is what we call local functorial properties of persistent homology.

Persistent homology also has certain global functorial properties. There are various ways of representing data in the form of sets of measurements, we might choose different units or different parametrizations of a domain of measurements, or we might need to focus only on certain sets of transformations that act on the dataset, such as rotations. Furthermore, the same measurements might be part of different data sets. These are some of the reasons why it is essential to be able to compare data sets equipped with different structures. For that purpose we introduce the notion of incarnations of data sets to encode different actions, and SEOs to compare incarnations. An incarnation of a data set Φ is an action of a subset $M \subset \mathrm{End}_\Phi(X)$. A SEO (set equivariant operator) between two incarnations (Φ, M) and (Ψ, N) is a pair consisting of a map $T : M \to N$ and an equivariant (with respect to T) function $\alpha : \Phi \to \Psi$. The use of this kind of operators for the comparison of incarnations of data sets has been inspired by [10, 11] , where GENEOs (group equivariant non-expansive operators) are introduced and used for applications to neural networks. We will see how different kind of SEOs make it possible or not to obtain a comparison between the persistent homologies of two incarnations. When such a comparison is not possible, the information given by the SEO and the one given by persistent homology are somehow complementary.

Consider a SEO obtained by composing measurements in a data set by a given real valued function defined on the real numbers. Multiplication by -1 is an example of such a SEO. It has the effect of turning the sub-level sets persistent homology of a measurement into its super-level sets persistent homology, leading in general to

a completely different information about the data set. The outcome consists of two different points of view on the same object, that are not functorially comparable, but together may enhance the accuracy of the analysis of the object of interest.

5.2 Data sets

For us a data set, which we regard as a point in the data landscape, is given by a finite set of real valued functions on some finite set X also called measurements.

Definition 55. Given a finite set X, a data set Φ is a finite set of functions, called *measurements*,

$$\Phi = \{\phi_i \colon X \to \mathbb{R} \mid i = 1, \cdots, m\}.$$

We define $\mathrm{dom}(\Phi)$, the **domain** of data set Φ, to be the set X which is the domain of the functions in Φ.

The most fundamental aspect of a data set Φ is that it is a set.

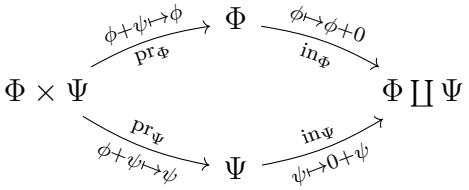
Definition 56. We denote by **Data** the category whose objects are data sets, with functions between data sets as morphisms.

This is the most primitive landscape of data sets. The nature of our data sets however can be used to impose more intricate structures and more meaningful landscapes. This is reminiscent of the case of groups. The most fundamental aspect of a group is that it is a set. However the category whose morphisms are group homomorphism is a much more meaningful landscape in which to study relationships between groups. To understand relationships between topological groups, the category with continuous group homomorphisms provides an even more meaningful landscape.

In this most primitive landscape however we can already perform products and coproducts.

Definition 57. Let $\phi \colon X \to \mathbb{R}$ and $\psi \colon Y \to \mathbb{R}$ be functions. Define $\phi + \psi \colon X \coprod Y \to \mathbb{R}$ to be the function that maps x in X to $\phi(x)$ and y in Y to $\psi(y)$. The **coproduct** of two data sets Φ and Ψ, denoted by $\Phi \coprod \Psi$, is defined to be the data set given by the measurements $\{\phi + 0 \mid \phi \in \Phi\} \cup \{0 + \psi \mid \psi \in \Psi\}$ on $X \coprod Y$. Their **product**, denoted by $\Phi \times \Psi$, is defined to be the data set given by the measurements $\{\phi + \psi \mid \phi \in \Phi \text{ and } \psi \in \Psi\}$ on $X \coprod Y$.

The functions:



satisfy the following universal properties, which justify the names coproduct and product:

- for any data set Π, and any two functions $\alpha\colon \Phi \to \Pi$ and $\beta\colon \Psi \to \Pi$, there is a unique function $\mu\colon \Phi \coprod \Psi \to \Pi$ for which $\mu \operatorname{in}_\Phi = \alpha$ and $\mu \operatorname{in}_\Psi = \beta$;
- for any data set Π, and any two functions $\alpha\colon \Pi \to \Phi$ and $\beta\colon \Pi \to \Psi$, there is a unique function $\mu\colon \Pi \to \Phi \times \Psi$ for which $\operatorname{pr}_\Phi \mu = \alpha$ and $\operatorname{pr}_\Psi \mu = \beta$.

In the rest of the Chapter we will consider some examples of transformations between data sets.

Definition 58. Let $f\colon \mathbb{R} \to \mathbb{R}$ be a function. By composing with f, a data set Φ is transformed into a new data set $f\Phi := \{f\phi \mid \phi \in \Phi\}$. This operation is called **change of units** along f. The symbol $f-\colon \Phi \to f\Phi$ denotes the function mapping ϕ to $f\phi$.

For example let $s\colon \mathbb{R} \to \mathbb{R}$ be the map

$$s(x) = \begin{cases} 1 & \text{if } x \geq 0 \\ -1 & \text{if } x < 0. \end{cases}$$

Consider $X = \{x_1, x_2\}$ and two data sets $\Phi = \{\phi_1, \phi_2\}$ and $\Psi = \{\psi_1, \psi_2\}$ given by the constant functions

$$\begin{aligned} \phi_1(x_1) &= \phi_1(x_2) = 1 \\ \phi_2(x_1) &= \phi_2(x_2) = 2 \\ \psi_1(x_1) &= \psi_1(x_2) = -1 \\ \psi_2(x_1) &= \psi_2(x_2) = 1. \end{aligned}$$

Consider also the function $\alpha : \Phi \to \Psi$ such that $\alpha(\phi_1) = \psi_1$ and $\alpha(\phi_2) = \psi_2$. Then $s\Phi = \{\phi_1\}$ and $s-\colon \Psi \to s\Psi$ is the identity. Thus, there is no function $s\Phi \to s\Psi$ making the following diagram commutative:

$$\begin{array}{ccc} \Phi & \xrightarrow{s-} & s\Phi = \{1\} \\ {\scriptstyle\alpha}\downarrow & & \downarrow \\ \Psi & \xrightarrow{s-=\mathrm{id}} & s\Psi = \Psi \end{array}$$

Consequently, for that s there is no functor S assigning to a data set Φ its change of units $s\Phi$ along s, for which $s-\colon \Phi \to s\Phi$ is a natural transformation between S and the identity functor. If the function f of a change of units is invertible, then $f-\colon \Phi \to f\Phi$ is a bijection whose inverse is given by $f^{-1}-$. The association $(\alpha\colon \Phi \to \Psi) \mapsto ((f-)\alpha(f^{-1}-)\colon f\Phi \to f\Psi)$ is a functor for which $f-\colon \Phi \to f\Phi$ is a natural transformation between this functor and the identity functor. Changing the units along any function preserves products and coproducts i.e., $f(\Phi \coprod \Psi)$ is

isomorphic to $f(\Phi) \coprod f(\Psi)$, and $f(\Phi \times \Psi)$ is isomorphic to $f(\Phi) \times f(\Psi)$. As we have seen in the second Chapter of the thesis, it is often interesting to consider transformations of a data set given by a rescaling of the measurements, and this is exactly the idea behind change of units.

Another simple way to transform a data set is precomposing every measurement of the data set with a function with codomain equal to the domain of the data set.

Definition 59. Let Φ be a data set with the domain X. By composing a function $f\colon Y \to X$ with the measurements in Φ, we obtain a new data set $\Phi f := \{\phi f \mid \phi \in \Phi\}$ with the domain Y. This operation is called **domain change** along f. The symbol $-f\colon \Phi \to \Phi f$ denotes the function that maps ϕ to ϕf.

Let $f_1\colon Z_1 \to X$ and $f_2\colon Z_2 \to Y$ be functions and $f_1 \coprod f_2\colon Z_1 \coprod Z_2 \to X \coprod Y$ be their coproduct. For any datasets Φ and Ψ with $\mathrm{dom}(\Phi) = X$ and $\mathrm{dom}(\Psi) = Y$, the following equalities hold:

$$(\Phi \coprod \Psi)(f_1 \coprod f_2) = \Phi f_1 \coprod \Psi f_2, \qquad (\Phi \times \Psi)(f_1 \coprod f_2) = \Phi f_1 \times \Psi f_2.$$

5.3 Metrics and persistent homology

We can think about a data set Φ as a subset $\Phi \subset \mathbb{R}^{|X|}$. Via this inclusion Φ inherits a metric induced by the infinity norm $\|v\|_\infty = \max\{|v_i|\}$ on $\mathbb{R}^{|X|}$. We use the symbol $\|\phi - \psi\|_\infty$ to denote the distance between ϕ and ψ in Φ. The considered data sets are not just sets anymore but metric spaces. Therefore, non-expansive (i.e. 1-Lipschitz) functions between data sets play a special role. For example, let $f\colon \mathbb{R} \to \mathbb{R}$ be a function. If f is non-expansive, then so is the change of units along f, $f-\colon \Phi \to f\Phi$, that maps ϕ to $f\phi$. The domain change $-h\colon \Phi \to \Phi h$ is non-expansive along any h. Non-expansiveness is an important assumption to prove some stability results in [10] and it is also reasonable in applications, since it is important that these functions between data sets do not alter the information too much.

By taking all the measurements of Φ together, we can form a function $\varphi : [\phi_1 \cdots \phi_m]\colon X \to \mathbb{R}^m$. Via this function, X inherits a pseudo-metric d_Φ defined as the pullback of a distance on $\mathbb{R}^m$. Given a metric δ on $\mathbb{R}^m$, we can define a pesudo-distance $d_{\varphi^{-1}\delta}$ as

$$d_{\varphi^{-1}\delta}(x, y) = \delta(\varphi(x), \varphi(y)). \tag{5.1}$$

Here, we will focus on the pseudo-metric induced by the Chebyshev metric on $\mathbb{R}^m$. Explicitly, this pseudo-distance is $d_\Phi(x, y) := \max_{1 \le i \le m} |\phi_i(x) - \phi_i(y)|$.

Proposition 7. *Given a finite metric space (X, d_X) with n points, there is a data set Φ with n measurements such that the metric d_Φ coincides with d_X.*

Proof. For each point $x_i \in X$ consider the function $\phi_i : X \to \mathbb{R}$, defined as

$$\phi_i(x_j) = d_X(x_i, x_j).$$

Take $\Phi = \{\phi_i\}_{i=1,\dots,n}$. For any $x_i, x_j \in X$, it holds $d_\Phi(x_i, x_j) = d_X(x_i, x_j)$. In fact,

$$d_\Phi(x_i, x_j) = \max_{1\leq k\leq n} |\phi_k(x_i) - \phi_k(x_j)| = \max_{1\leq k\leq n} |d_X(x_k, x_i) - d_X(x_k, x_j)|.$$

Then, because of the triangular inequality, for any $k = 1, \dots, n$ it holds

$$|d_X(x_k, x_i) - d_X(x_k, x_j)| \leq d_X(x_j, x_i) = |\phi_j(x_i) - \phi_j(x_j)|$$

and the claim follows. □

This metric plays a fundamental role as it permits us to extract persistent homologies (see [20, 40]) for each measurement in the dataset. To obtain a meaningful filtration, of which we will compute the persistent homology, we will need to consider more than one parameter. If we try to use only the sublevel sets of each measurement, we obtain only a filtration of vertices, since there are no other simplices defined on X. If, on the other hand, we use only the metric structure given by d_Φ, we loose the information encoded by the considered measurement. We will combine the two approaches to obtain a useful filtration. For each $\phi \in \Phi$ and $(r, s) \in [0, \infty) \times \mathbb{R}$ we will consider the simplicial complex $K_{r,s} = VR_r(\phi \leq s, d_\Phi)$, where $\phi \leq s$ is the subset of points of X where ϕ assumes values less or equal of s. Therefore, $VR_r(\phi \leq s, d_\Phi)$ is the simplicial complex whose simplices are the subsets of $\phi \leq s$ with diameter less or equal to r. We will apply to this filtration the homology functor in degree d, H_d, to obtain the persistent homology:

$$PH_d^\Phi(\phi)_{r,s} := H_d\left(K_{r,s}\right). \tag{5.2}$$

Such a persistence module is indexed in the category $([0, \infty) \times R, \leq)$, where $(r, s) \leq (r', s')$ if and only if $r \leq r'$ and $s \leq s'$. The persistence module is well defined, since if $s \leq s'$ and $r \leq r'$, then $(\phi \leq s) \subset (\phi \leq s')$ and therefore $\mathrm{VR}_r(\phi \leq s) \subset \mathrm{VR}_{r'}(\phi \leq s')$. The linear function induced on homology by this inclusion is denoted by:

$$PH_d^\Phi(\phi)_{(r,s)\leq(r',s')} \colon PH_d^\Phi(\phi)_{r,s} \to PH_d^\Phi(\phi)_{r',s'}.$$

These functions form a functor $PH_d^\Phi(\phi)$ indexed by the poset $[0, \infty) \times \mathbb{R}$ with values in the category of vector spaces. Since X is finite, $PH_d^\Phi(\phi)$ is **tame** (see [89]). This means that values of $PH_d^\Phi(\phi)$ are finite dimensional, and there are two finite sequences $0 = r_0 < r_1 < \cdots < r_m$ in $[0, \infty)$ and $s_0 < s_1 < \cdots < s_l = \infty$ in $\mathbb{R}$ such that $PH_d^\Phi(\phi)$, restricted to subposets of the form $[r_i, r_{i+1}) \times (\infty, s_0) \subset [0, \infty) \times \mathbb{R}$ and $[r_i, r_{i+1}) \times [s_j, s_{j+1}) \subset [0, \infty) \times \mathbb{R}$, is constant. The category of such functors

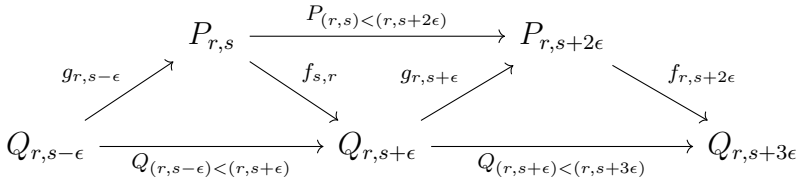
is denoted by $\mathbf{Tame}\left(\mathbf{Vect}^{[0,\infty)\times\mathbb{R}}\right)$.Thus a data set Φ leads to a function assigning to each measurement ϕ its persistent homology in a given degree:

$$PH_d^{\Phi}\colon \Phi \to \mathbf{Tame}\left(\mathbf{Vect}^{[0,\infty)\times\mathbb{R}}\right).$$

To compare different persistence modules we cannot use the interleaving distance as it was defined in the first Chapter. We will recall here another notion of interleaving, in order to define an interleaving distance on $\mathbf{Tame}\left(\mathbf{Vect}^{[0,\infty)\times\mathbb{R}}\right)$ (see [71]).

Definition 60. Let P and Q be in $\mathbf{Tame}\left(\mathbf{Vect}^{[0,\infty)\times\mathbb{R}}\right)$.

- P and Q are ϵ-**interleaved** if, for all (r,s) in $[0,\infty)\times\mathbb{R}$, there are linear functions $f_{s,r}\colon P_{r,s} \to Q_{r,s+\epsilon}$ and $g_{s,r}\colon Q_{r,s} \to P_{r,s+\epsilon}$ making the following diagram commutative:



- $d_{\bowtie}(P,Q) := \inf\{\epsilon \in [0,\infty) \mid P \text{ and } Q \text{ are } \epsilon\text{-interleaved}\}$.

The function $P,Q \mapsto d_{\bowtie}(P,Q)$ is an extended (∞ is allowed) metric on the set $\mathbf{Tame}\left(\mathbf{Vect}^{[0,\infty)\times\mathbb{R}}\right)$ called interleaving metric in the direction of the vector (0,1).

Proposition 8. *The function* $PH_d^{\Phi}\colon \Phi \to \mathbf{Tame}\left(\mathbf{Vect}^{[0,\infty)\times\mathbb{R}}\right)$ *is non-expansive if the set* Φ *is equipped with* ∞*-norm metric* $\|\phi-\psi\|_{\infty}$ *and the set* $\mathbf{Tame}\left(\mathbf{Vect}^{[0,\infty)\times\mathbb{R}}\right)$ *is equipped with the interleaving metric in the direction of the vector* (0,1).

Proof. Let $\phi,\psi\colon X \to \mathbb{R}$ be measurements in Φ and $\epsilon = \|\phi-\psi\|_{\infty}$. For every s in $\mathbb{R}$, the sublevel set $\phi \leq s$ is a subset of $\psi \leq s+\epsilon$, and $\psi \leq s$ is a subset of $\phi \leq s+\epsilon$. This translates into inclusions:

$$\mathrm{VR}_r(\phi \leq s, d_{\Phi}) \subset \mathrm{VR}_r(\psi \leq s+\epsilon, d_{\Phi}) \qquad \mathrm{VR}_r(\psi \leq s, d_{\Phi}) \subset \mathrm{VR}_r(\phi \leq s+\epsilon, d_{\Phi})$$

leading to functions:

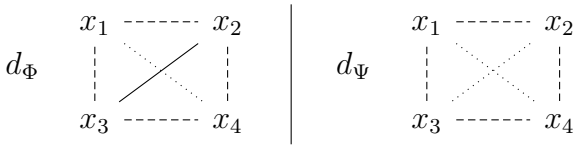
$$f_{s,r}\colon PH_d^{\Phi}(\phi)_{r,s} \to PH_d^{\Phi}(\psi)_{r,s+\epsilon} \qquad g_{s,r}\colon PH_d^{\Phi}(\psi)_{r,s} \to PH_d^{\Phi}(\phi)_{r,s+\epsilon}.$$

These functions provide ϵ interleaving between $PH_d^{\Phi}(\phi)$ and $PH_d^{\Phi}(\psi)$, giving $\|\phi-\psi\|_{\infty} \geq d_{\bowtie}(PH_d^{\Phi}(\phi), PH_d^{\Phi}(\psi))$. □

A measurement $\phi\colon X \to \mathbb{R}$ can be part of many data sets and its persistent homology depends on what data set this function is part of. For example, let $X = \{x_1, x_2, x_3, x_4\}$ and $\phi, \psi\colon X \to \mathbb{R}$ be measurements defined as follows:

$$\begin{array}{lll} \phi(x_1) = -1 & \phi(x_2) = \phi(x_3) = 0 & \phi(x_4) = 1 \\ \hline \psi(x_3) = -1 & \psi(x_1) = \psi(x_4) = 0 & \psi(x_2) = 1 \end{array}$$

The measurement ϕ is part of two data sets $\Phi = \{\phi\}$ and $\Psi = \{\phi, \psi\}$. The induced pseudometrics d_Φ and d_Ψ on X can be depicted by the following diagrams where the continuous, dashed, and dotted lines indicate distance 0, 1 and 2 respectively:

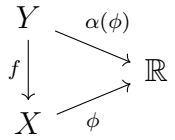


In this case $PH_1^\Phi(\phi)_{r,s} = 0$ for all r and s, however:

$$\dim PH_1^\Psi(\phi)_{r,s} = \begin{cases} 1 & \text{if } 1 \leq s \text{ and } 1 \leq r < 2 \\ 0 & \text{otherwise} \end{cases}$$

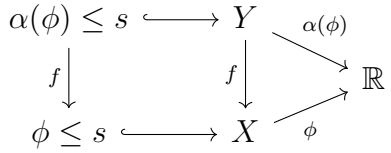
To understand persistent homology, it is therefore paramount to understand how it changes when data sets change and here functoriality plays an essential role. This is not always achieved, but we can see that under particular conditions this is the case.

Definition 61. Let Φ and Ψ be data sets consisting of measurements on X and Y respectively. A function $\alpha\colon \Phi \to \Psi$ is called **geometric** if there is a function $f\colon Y \to X$, called a **realization** of α, making the following diagram commutative for every ϕ in Φ:



For example $-f\colon \Phi \to \Phi f$ is geometric, as it is realized by f.

The commutativity of the triangle above has two consequences. First, f is non-expansive with respect to the pseudometrics d_Φ on X and d_Ψ on Y. Second, for s in $\mathbb{R}$ and ϕ in Φ, the subset $(\alpha(\phi) \leq s) \subset Y$ is mapped via f into $(\phi \leq s) \subset X$, i.e., the following diagram commutes:



Therefore, the realization f induces a map of Vietoris-Rips complexes and their homologies:

$$f_{s,r}\colon \mathrm{VR}_r(\alpha(\phi) \leq s, d_\Psi) \to \mathrm{VR}_r(\phi \leq s, d_\Phi);$$

$$\begin{array}{ccc} PH_d^\Psi(\alpha(\phi))_{r,s} & & PH_d^\Phi(\phi)_{r,s} \\ \| & & \| \\ H_d\left(\mathrm{VR}_r(\alpha(\phi) \leq s, d_\Psi)\right) & \xrightarrow{H_d(f_{r,s})} & H_d\left(\mathrm{VR}_r(\phi \leq s, d_\Phi)\right). \end{array}$$

If $f, f'\colon Y \to X$ are two realizations of α, then for y in Y, $d_\Phi(f(y), f'(y)) = 0$, hence they are points of the same simplex in the Vietoris-Rips complex, implying that $f_{r,s}$ and $f'_{r,s}$ are homotopic for all r and s. Consequently, $H_d(f_{r,s}) = H_d(f'_{r,s})$. The linear function $H_d(f_{r,s})$ depends therefore only on α and it is independent on the choice of its realization f. We denote this function by:

$$PH_d^\alpha(\phi)_{r,s}\colon PH_d^\Psi(\alpha(\phi))_{r,s} \to PH_d^\Phi(\phi)_{r,s}.$$

These functions are natural in r and s and induce a morphism in the category $\mathbf{Tame}\left(\mathbf{Vect}^{[0,\infty)\times\mathbb{R}}\right)$ between persistent homologies:

$$PH_d^\alpha(\phi)\colon PH_d^\Psi(\alpha(\phi)) \to PH_d^\Phi(\phi).$$

If $\alpha\colon \Phi \to \Psi$ and $\beta\colon \Psi \to \Xi$ are geometric functions realized by $f\colon Y \to X$ and $g\colon Z \to Y$, then the composition $\beta\alpha\colon \Phi \to \Xi$ is also geometric, and realized by the composition $fg\colon Z \to X$. Consequently, for every measurement ϕ in Φ, $PH_d^{\beta\alpha}(\phi) = PH_d^\alpha(\phi)PH_d^\beta(\alpha(\phi))$, that assures the commutativity of the diagram:

$$PH_d^\Xi(\beta\alpha(\phi)) \xrightarrow{PH_d^\beta(\alpha(\phi))} PH_d^\Psi(\alpha(\phi)) \xrightarrow{PH_d^\alpha(\phi)} PH_d^\Phi(\phi)$$
$$\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad \searrow \quad \nearrow$$
$$\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad PH_d^{\beta\alpha}(\phi)$$

For any $\alpha\colon \Phi \to \Psi$, taking persistent homology leads to two functions on Φ:

$$\begin{array}{ccccc} & & \xrightarrow{PH_d^\Phi} & & \mathbf{Tame}\left(\mathbf{Vect}^{[0,\infty)\times\mathbb{R}}\right) \\ \Phi & \nearrow & & & \\ & \searrow^{\alpha} & \Psi & \xrightarrow{PH_d^\Psi} & \mathbf{Tame}\left(\mathbf{Vect}^{[0,\infty)\times\mathbb{R}}\right) \end{array}$$

These functions rarely coincide. However, when α is geometric, we can use the morphisms $PH_d^\alpha(\phi)\colon PH_d^\Psi(\alpha(\phi)) \to PH_d^\Phi(\phi)$ to compare the values of these two functions on Φ. For non-geometric α, we are not equipped with such comparison morphisms and there is no reason for such a comparison to even exist. For example,

consider the change of unit along the function $f\colon \mathbb{R} \to \mathbb{R}$, $f(x) := -x$. Then $f-\colon \Phi \to f\Phi$ is an isomorphism. In this case

$$PH_d^{\Phi}(\phi)_{r,s} := H_d\left(\mathrm{VR}_r(\phi \leq s, d_{\Phi})\right) \Big| (f-)PH_d^{f\Phi}(\phi) = H_d\left(\mathrm{VR}_r(\phi \geq -s, d_{\Phi})\right).$$

Thus PH_d^{Φ} encodes information about sub-level sets of the measurements in Φ and $(f-)PH_d^{f\Phi}$ encodes information about super-level sets of the measurements. These persistent homologies encode therefore the same information as the so called extended persistence (see [27, 82]).

5.4 Actions

To describe symmetries of a data set Φ with domain X, we consider operations on X that convert measurements into measurements. By definition a Φ-**operation** is a function $g\colon X \to X$ such that, for every measurement ϕ in Φ, the composition ϕg also belongs to Φ. If $g\colon X \to X$ is such an operation, then, for all ϕ and ψ in Φ:

$$\|\phi - \psi\|_{\infty} = \max_{x\in X}|\phi(x) - \psi(x)| \geq \max_{x\in \mathrm{im}(g)}|\phi(x) - \psi(x)| = \|\phi g - \psi g\|_{\infty}.$$

Thus the function $-g\colon \Phi \to \Phi$ that maps ϕ to ϕg is non-expansive.

The composition of Φ-operations is again a Φ-operation, and the identity function id_X is also a Φ-operation. In this way the set of Φ-operations with the composition becomes a unitary monoid, called the **structure monoid** of Φ, and denoted by:

$$\mathrm{End}_{\Phi}(X) = \{g : X \to X \mid \phi g \in \Phi \text{ for every } \phi \in \Phi\} \subset \mathrm{End}(X).$$

A Φ-operation g is invertible if there is a Φ-operation h such that $gh = hg = \mathrm{id}_X$. Since Φ is finite, a Φ-operation is invertible if and only if it is a bijection. Their collection is denoted by:

$$\mathrm{Aut}_{\Phi}(X) = \{g : X \to X \mid \ g \text{ is a bijection, and } \phi g \in \Phi \text{ for every } \phi \in \Phi\}.$$

With the composition operation, $\mathrm{Aut}_{\Phi}(X)$ becomes a group for which the inclusion $\mathrm{Aut}_{\Phi}(X) \subset \mathrm{End}_{\Phi}(X)$ is a monoid homomorphism.

A data set Φ is equipped with an associative right action:

$$\Phi \times \mathrm{End}_{\Phi}(X) \to \Phi, \qquad (\phi, g) \mapsto \phi g.$$

Thus Φ is not just a set, but a set with an action of the monoid $\mathrm{End}_{\Phi}(X)$. To encode the symmetries of Φ induced by this action, we consider its incarnations.

Definition 62. An **incarnation** of Φ is a choice of a subset $M \subset \mathrm{End}_{\Phi}(X)$ (in general, not necessarily a submonoid). An incarnation is denoted by a pair (Φ, M). We think about M as an additional structure on Φ. An incarnation of the form (Φ, M) is called an M-incarnation.

We also refer to an M-incarnation as an M-action. The choice of an M-action on Φ encodes certain symmetries of Φ. Different choices of M can encode different symmetries. This flexibility is important in applications. For example in data sets that represent images, we might want to focus on rotational symmetries, so we may use an appropriate action on the data set to inject the corresponding geometry. The incarnation $(\Phi, \mathrm{End}_\Phi(X))$ is an example of a incarnation called *universal*.

An incarnation (Φ, M) is called a **monoid incarnation** if $M \subset \mathrm{End}_\Phi$ is a submonoid, and our convention here is that all such submonoids contain the identity element. If (Φ, M) is an incarnation, we use the symbol $(\Phi, \langle M \rangle)$ to denote the monoid incarnation induced by M, where $\langle M \rangle \subset \mathrm{End}_\Phi(X)$ is the submonid generated by M.

If a submonoid $M \subset \mathrm{End}_\Phi(X)$ is a group, then (Φ, M) is called a **group incarnation**. The incarnation $(\Phi, \mathrm{Aut}_\Phi(X))$ is an example of a group incarnation called universal.

Let (Φ, M) be an incarnation for which any element g in M is a bijection. Such incarnations are called **group-like**. For group like incarnations (Φ, M) the finiteness implies that the monoid $\langle M \rangle$ is in fact a subgroup of $\mathrm{Aut}_\Phi(X)$. Thus any group-like incarnation (Φ, M) leads to a group incarnation $(\Phi, \langle M \rangle)$.

Let (Φ, M) be an incarnation. For a subset $\Omega \subset \Phi$, the symbol ΩM denotes the set of all the measurements in Φ which either belong to Ω or are of the form $\omega g_1 \cdots g_k$, for some ω in Ω and some sequence of elements $g_1, \ldots g_k$ in M. If $\Omega M = \Phi$, then Ω is said to **generate** the incarnation (Φ, M). In the case (Φ, M) is a monoid incarnation, then any element in ΩM is of the form ωg for some ω in Ω and g in M. Note that $\Omega M = \Omega \langle M \rangle$ for every incarnation (Φ, M).

If ψ belongs to $\phi M := \{\phi\} M$, then ψ is said to be a **deformation** of ϕ. If (Φ, M) is a group incarnation, then the relation of being a deformation is an equivalence relation. For a general incarnation however being a deformation can fail to be even a symmetric relation. Two measurements in Φ are said to be **connected** if they are related by the equivalence relation generated by the relation of being a deformation, that is the smallest equivalence relation that contains the relation of deformation. The symbol Φ/M denotes the partition of Φ induced by this equivalence relation. We refer to Φ/M as the **quotient** of the incarnation (Φ, M). The partitions Φ/M and $\Phi/\langle M \rangle$ coincide. If (Φ, M) is a group incarnation, then Φ/M coincide with the orbit partition of the usual group action of M on Φ.

Let (Φ, M) be an incarnation. For a measurement ψ in Φ, the symbol $[\psi]$ denotes the *block* in Φ/M containing ψ. Explicitly, $[\psi]$ is the subset of Φ consisting of all the measurements connected to ψ. Note that, for all g in M, if ϕ is connected

to ψ, then ϕg is also connected to ψ. We thus have the following inclusions:

$$\begin{array}{ccc} M & \hookrightarrow & \mathrm{End}_{\Phi}(X) \\ \downarrow & & \downarrow \\ \mathrm{End}_{[\psi]}(X) & \hookrightarrow & \mathrm{End}(X) \end{array}$$

The M incarnation $([\psi], M)$ of the block $[\psi]$, given by the above inclusions $M \subset \mathrm{End}_{[\psi]}$, is called a **block incarnation** of (Φ, M). In this way we can think about $[\psi]$ and $([\psi], M)$ as a new data set.

An incarnation (Φ, M) is called **transitive** if all the elements in Φ are connected to each other. For example, let M be a finite submonoid of $\mathrm{End}(X)$. For a given function $\phi\colon X \to \mathbb{R}$, define a data set $\phi M := \{\phi g \mid g \in M\}$ to consist of all functions of the form $x \mapsto \phi(g(x))$ for all g in M. Then every $g\colon X \to X$ in M is a ϕM-operation. The obtained incarnation $(\phi M, M)$ is transitive. Any transitive group incarnation is of such form. For all measurements ϕ in any incarnation (Φ, M), the block incarnation $([\phi], M)$ is transitive. Any transitive incarnation is of this form.

Definition 63. Let (Φ, M) be an incarnation. A subset $\Omega \subset \Phi$ is called **independent** if no element in Ω is a deformation of any other element in Ω, explicitly: $\omega \notin \omega' M$ for all $\omega \neq \omega'$ in Ω.

A **basis** of (Φ, M) is an independent subset $\Omega \subset \Phi$ such that $\Omega M = \Phi$ (Ω generates (Φ, M)).

Two measurements ψ and ϕ are called **indistinguishable** if ψ is a deformation of ϕ and ϕ is a deformation of ψ. If (Φ, M) is a group incarnation, then ψ and ϕ are indistinguishable if and only if $\psi = \phi g$ for some g in M, i.e., if ψ is a deformation of ϕ.

Proposition 9. *1. Every incarnation has a basis.*

2. Let $\Omega, \Omega' \subset \Phi$ be two bases of an incarnation (Φ, M). Then there is a bijection $\sigma\colon \Omega \to \Omega'$ such that ω and $\sigma(\omega)$ are indistinguishable for every ω in Ω.

Proof. (1): Let (Φ, M) be an incarnation. Choose $\Omega \subset \Phi$ to be an independent subset for which ΩM is maximal. Existence of Ω is guaranteed by finiteness of Φ. We claim that $\Omega M = \Phi$ and hence Ω is a basis. If this is not the case, let ψ be in $\Phi \setminus \Omega M$. Define $\Omega' = \{\psi\} \cup \{\omega \in \Omega \mid \omega \notin \{\psi\} M\}$. Then $\Omega' M$ contains Ω and hence ΩM. It also contains ψ. Since Ω' is independent, we would obtain a contradiction to the maximality assumption about ΩM, and thus the claim holds.

(2): Let ω be in Ω. Since $\Omega M = \Phi = \Omega' M$, there is ω' in Ω' such that $\omega \in \omega' M$. Let ω_1 in Ω be such that $\omega' \in \omega_1 M$. Then $\omega \in \omega' M \subset \omega_1 M$, and hence $\omega = \omega_1$ by the independence of Ω. The desired bijection is then given by $\omega \mapsto \omega'$. □

According to Proposition 9, any two bases of an incarnation have the same number of elements. We define the **dimension** of an incarnation to be the cardinality of its bases. For example a transitive group incarnation has dimension 1. In fact for a transitive group incarnation any single measurement forms a basis. More generally, the dimension of a group incarnation (Φ, M) equals the cardinality of Φ/M. In this case $\Omega \subset \Phi$ is a basis if and only if, for every block Ψ in Φ/M, the intersection $\Omega \cap \Psi$ has only one element. Since being a basis depends only on the monoid $\langle M \rangle$, the dimension of a group-like incarnation (Φ, M) equals also the cardinality of Φ/M, and similarly a subset $\Omega \subset \Phi$ is a basis if and only if, for every block Ψ in the partition Φ/M, the intersection $\Omega \cap \Psi$ has only one element.

The dimension of a transitive monoid incarnation can be bigger than 1. For example, let $X = \{x_1, x_2, x_3\}$ and consider functions $\phi_1, \phi_2, \phi_3 \colon X \to \mathbb{R}$ and $g_1, g_2, g_3 \colon X \to X$ defined as follows:

$\phi_1(x_1) = 2$	$\phi_2(x_1) = 2$	$\phi_3(x_1) = 1$	$g_1(x_1) = x_2$	$g_2(x_1) = x_2$	$g_3(x_1) = x_1$
$\phi_1(x_2) = 2$	$\phi_2(x_2) = 2$	$\phi_3(x_2) = 2$	$g_1(x_2) = x_2$	$g_2(x_2) = x_2$	$g_3(x_2) = x_2$
$\phi_1(x_3) = 3$	$\phi_2(x_3) = 2$	$\phi_3(x_3) = 2$	$g_1(x_3) = x_3$	$g_2(x_3) = x_2$	$g_3(x_3) = x_2$

The compositions $g_i g_j$ and $\phi_i g_j$ are described by the following tables:

	g_1	g_2	g_3
g_1	g_1	g_2	g_2
g_2	g_2	g_2	g_2
g_3	g_2	g_2	g_3

	g_1	g_2	g_3
ϕ_1	ϕ_1	ϕ_2	ϕ_2
ϕ_2	ϕ_2	ϕ_2	ϕ_2
ϕ_3	ϕ_2	ϕ_2	ϕ_3

Thus the functions g_1, g_2, and g_3 are $\Phi := \{\phi_1, \phi_2, \phi_3\}$-operations. Furthermore the subset $M := \{\mathrm{id}, g_1, g_2, g_3\} \subset \mathrm{End}_\Phi(X)$ is a submonoid. The incarnation (Φ, M) is a transitive monoid incarnation. Since the set $\{\phi_1, \phi_3\}$ is independent and generates (Φ, M), it is a basis. Thus (Φ, M) is an example of a transitive monoid incarnation of dimension 2.

5.5 Nirvana

To compare incarnations of various data sets we are going to use maps that preserve the incarnation structure.

Definition 64. A set equivariant operator (**SEO**) from an incarnation (Φ, M) to an incarnation (Ψ, N), denoted by $(\alpha, T) \colon (\Phi, M) \to (\Psi, N)$, is a pair of functions $(\alpha \colon \Phi \to \Psi, T \colon M \to N)$ for which the following diagram commutes:

$$\begin{array}{ccccc} \Phi \times M & \hookrightarrow & \Phi \times \mathrm{End}_\Phi(X) & \xrightarrow{\text{action}} & \Phi \\ \downarrow{\scriptstyle \alpha \times T} & & & & \downarrow{\scriptstyle \alpha} \\ \Psi \times N & \hookrightarrow & \Psi \times \mathrm{End}_\Psi(Y) & \xrightarrow{\text{action}} & \Psi \end{array}$$

Explicitly, for ϕ in Φ and g in M, it holds $\alpha(\phi g) = \alpha(\phi)T(g)$.

This implies that for ϕ in Φ and every sequence of elements $g_1, \ldots, g_k$ in M, it holds:

$$\alpha(\phi g_1 \cdots g_k) = \alpha(\phi)T(g_1) \cdots T(g_k).$$

Be however aware that in general there may not be a homomorphism $T\colon \langle M\rangle \to \langle N\rangle$ of monoids which extends $T\colon M \to N$ and makes the following diagram commutative:

$$\begin{array}{ccccccc} \Phi \times M & \hookrightarrow & \Phi \times \langle M\rangle & \hookrightarrow & \Phi \times \mathrm{End}_\Phi(X) & \xrightarrow{\text{action}} & \Phi \\ \downarrow{\scriptstyle \alpha\times T} & & \downarrow{\scriptstyle \alpha\times T} & & & & \downarrow{\scriptstyle \alpha} \\ \Psi \times N & \hookrightarrow & \Psi \times \langle N\rangle & \hookrightarrow & \Psi \times \mathrm{End}_\Psi(Y) & \xrightarrow{\text{action}} & \Psi \end{array}$$

A SEO between monoid incarnations $(\alpha, T)\colon (\Phi, M) \to (\Psi, N)$ is called a MEO (monoid equivariant operators) if $T\colon M \to N$ is a monoid homomorphism. A MEO between group incarnations is also called a GEO (group equivariant operators).

Let $(\alpha_0, T_0)\colon (\Phi_0, M_0) \to (\Phi_1, M_1)$ and $(\alpha_1, T_1)\colon (\Phi_1, M_1) \to (\Phi_2, M_2)$ be SEOs. Then the compositions $(\alpha_1\alpha_0, T_1T_0)$ form a SEO. Furthermore the pair $(\mathrm{id}_\Phi, \mathrm{id}_M)\colon (\Phi, M) \to (\Phi, M)$ is also a SEO. The composition of SEOs is an associative operation and defines a category structure on the collection of data set incarnations with SEOs as morphisms. This category is called **Nirvana**.

A SEO $(\alpha, T)\colon (\Phi, M) \to (\Psi, N)$ is an isomorphism if and only if both of the functions α and T are bijections. Isomorphisms preserve independence and being a basis:

Proposition 10. *If $(\alpha, T)\colon (\Phi, M) \to (\Psi, N)$ is an isomorphism, then a subset $\Omega \subset \Phi$ is independent or a basis if and only if its image $\alpha(\Omega) \subset \Psi$ is independent or a basis.*

Proof. Assume α and T are bijections. This assumption imply that ϕ_1 belongs to $\phi_2 M$ if and only if $\alpha(\phi_1)$ belongs to $\alpha(\phi_2)N$. It follows that two elements in Φ are (in)dependent if and only if their images via α are (in)dependent in Ψ. By the same argument, $\Omega M = \Phi$ if and only $\alpha(\Omega)T(M) = \alpha(\Phi)$. □

According to Proposition 10 two isomorphic incarnations have the same dimension.

The universal incarnations $(\Phi, \mathrm{End}_\Phi(X))$ and $(\Phi, \mathrm{Aut}_\Phi(X))$ are special in the category Nirvana. For any (Φ, M), the pair $(\mathrm{id}, i\colon M \hookrightarrow \mathrm{End}_\Phi(X))$ defines a SEO $(\Phi, M) \to (\Phi, \mathrm{End}_\Phi(X))$ called **canonical**. If (Φ, M) is a group incarnation, then the pair $(\mathrm{id}, i\colon M \hookrightarrow \mathrm{Aut}_\phi(X))$ defines a GEO $(\Phi, M) \to (\Phi, \mathrm{Aut}_\phi(X))$ also called canonical.

The rest of this section is devoted to present three ways of constructing SEOs.

5.5.1 Change of units

Consider a function $f\colon \mathbb{R} \to \mathbb{R}$. For any incarnation (Φ, M), consider the data set $f\Phi$ (see Section 5.2). If g is a Φ-operation, then it is also a $f\Phi$-operation. Thus, there is an inclusion $\mathrm{End}_\Phi(X) \subset \mathrm{End}_{f\Phi}(X)$, which is an equality if f is invertible, therefore we have an incarnation $(f\Phi, M)$. If (Φ, M) is a monoid or a group incarnation, then so is $(f\Phi, M)$. The pair $(f-, \mathrm{id}_M)\colon (\Phi, M) \to (f\Phi, M)$ is a SEO called the change of units along f.

Assume f is invertible. If $(\alpha, T)\colon (\Phi, M) \to (\Psi, N)$ is a SEO, then the pair of functions $((f-)\alpha(f^{-1}-), T)$ forms a SEO between $(f\Phi, M)$ and $(f\Psi, N)$. The assignment $(\alpha, T) \mapsto ((f-)\alpha(f^{-1}-), T)$ is a self functor $\mathrm{C}(f)$ of Nirvana also called the change of units along f. It is an equivalence of categories. Indeed,

$$\begin{aligned} \mathrm{C}(f)\mathrm{C}(f^{-1})((\Phi, M)) &= \mathrm{C}(f)(f^{-1}\Phi, M) = (\Phi, M) \\ \mathrm{C}(f)\mathrm{C}(f^{-1})((\alpha, T)) &= \mathrm{C}(f)((f^{-1}-)\alpha(f-), T)) \\ &= ((f-)(f^{-1}-)\alpha(f-)(f^{-1}-), T) = (\alpha, T). \end{aligned}$$

The same holds for $\mathrm{C}(f^{-1})\mathrm{C}(f)$, hence $\mathrm{C}(f)$ is an equivalence of categories. The SEOs $(f-, \mathrm{id}_M)\colon (\Phi, M) \to (f\Phi, M)$, for all incarnations (Φ, M), form a natural transformation between the identity functor on **Nirvana** and the change of units along f functor.

5.5.2 Domain change

Let (Φ, M) and (Ψ, N) be incarnations of data sets consisting of measurements on X and Y respectively. A SEO $(\alpha, T)\colon (\Phi, M) \to (\Psi, N)$ is called **geometric** if there is a function $f\colon Y \to X$, called a realization of (α, T), making the following diagram commutative for every ϕ in Φ and g in M:

$$\begin{array}{ccccc} Y & \xrightarrow{T(g)} & Y & \xrightarrow{\alpha(\phi)} & \\ {\scriptstyle f}\downarrow & & \downarrow{\scriptstyle f} & & \mathbb{R} \\ X & \xrightarrow[g]{} & X & \xrightarrow{\phi} & \end{array}$$

For example, let (Φ, M) be an incarnation of a data set consisting of measurements on X. Then the SEO $(\mathrm{id}_\Phi, \mathrm{id}_M)\colon (\Phi, M) \to (\Phi, M)$ is geometric. The identity function $\mathrm{id}_X\colon X \to X$ is one of its realizations.

Let $Y \subset X$ be M-invariant: $g(y)$ belongs to Y for all y in Y and g in M. Consider the data set $\Phi|_Y$ given by the domain change along the inclusion $Y \subset X$. The restriction of g to Y is a $\Phi|_Y$-operation for every g in M. We use the symbol $T_Y\colon M \to \mathrm{End}_{\Phi|_Y}(Y)$ to denote the function that maps g in M to the restriction of g to Y. The incarnation $(\Phi|_Y, T_Y(M))$ is called the **restriction** of (Φ, M) to the

invariant subset Y. The pair $(\Phi \twoheadrightarrow \Phi|_Y, T_Y)$ forms a geometric SEO. The inclusion $i_Y \colon Y \hookrightarrow X$ is one of its realizations.

Let $f \colon Y \to X$ be a bijection. Consider the data set Φf. For any g in M, the function $f^{-1}gf \colon Y \to Y$ is a Φf-operation. Define $T \colon M \to \mathrm{End}_{\Phi f}(Y)$ to map g in M to $f^{-1}gf$. The incarnation $(\Phi f, T(M))$ is called the domain change of (Φ, M) along f. The pair $(-f \colon \Phi \to \Phi f, T)$ forms a geometric SEO and $f \colon Y \to X$ is one of its realizations.

5.5.3 Extending from a basis

SEOs can be effectively constructed using bases.

Proposition 11. *Let (Φ, M) and (Ψ, N) be incarnations and Ω be a basis of (Φ, M). Then two SEOs $(\alpha, T), (\alpha', T') \colon (\Phi, M) \to (\Psi, N)$ are equal if and only if $T = T'$ and $\alpha(\omega) = \alpha'(\omega)$ for any ω in Ω.*

Proof. The only non trivial thing to prove in the statement of the proposition is that $\alpha = \alpha'$ when their restrictions to Ω are equal. Assume $T = T'$ and $\alpha(\omega) = \alpha'(\omega)$ for any ω in Ω. Since Ω generates (Φ, M), any element in Φ is of the form $\phi = \omega g_1 \cdots g_k$ for some ω in Ω and a sequence of elements $g_1, \ldots, g_k$ in M. The assumption and the fact that (α, T) and (α', T) are SEOs, imply:

$$\alpha(\phi) = \alpha(\omega g_1 \cdots g_k) = \alpha(\omega) T(g_1) \cdots T(g_k) =$$

$$= \alpha'(\omega) T(g_1) \cdots T(g_k) = \alpha'(\omega g_1 \cdots g_k) = \alpha'(\phi).$$

Consequently $\alpha = \alpha'$. □

According to Proposition 11, a SEO is determined by what it does on a basis of the domain. This is analogous to a linear map between vector spaces being determined by its values on a basis. However unlike for linear maps, we cannot freely map elements of a basis of an incarnation to obtain a SEO. To obtain a SEO certain relations have to be preserved. Let (Φ, M) be an incarnation. A **relation** between measurements ϕ and ψ in Φ is by definition a pair of sequences $((g_1, \ldots, g_k), (h_1, \ldots, h_l))$ of elements in M for which the following equality holds: $\phi g_1 \cdots g_k = \psi h_1 \cdots h_l$.

Proposition 12. *Let (Φ, M) and (Ψ, N) be incarnations, Ω be a basis of (Φ, M), and $\bar{\alpha} \colon \Omega \to \Psi$ and $T \colon M \to N$ be functions.*

1. *Assume that for every relation $((g_1, \ldots, g_k), (h_1, \ldots, h_l))$ between any two elements ω, ω' in Ω, the pair $((T(g_1), \ldots, T(g_k)), (T(h_1), \ldots, T(h_l)))$ is a relation between $\alpha(\omega)$ and $\alpha(\omega')$ in Ψ. Under this assumption, there is a unique SEO $(\alpha, T) \colon (\Phi, M) \to (\Psi, N)$ for which the restriction of $\alpha \colon \Phi \to \Psi$ to Ω is $\bar{\alpha}$.*

2. *Assume* (Φ, M) *and* (Ψ, N)*, are monoid incarnations,* T *is a monoid homomorphism, and if* $\omega g = \omega' h$ *for some* ω, ω' *in* Ω *and* g, h *in* M*, then* $\alpha(\omega)T(g) = \alpha(\omega')T(h)$*. Under these assumptions, there is a unique MEO* $(\alpha, T)\colon (\Phi, M) \to (\Psi, N)$ *for which the restriction of* $\alpha\colon \Phi \to \Psi$ *to* Ω *is* $\bar{\alpha}$*.*

3. *Assume* (Φ, M) *and* (Ψ, N) *are group incarnations,* T *is a group homomorphism, and if* $\omega = \omega g$*, for some* ω *in* Ω *and* g *in* M*, then* $\alpha(\omega) = \alpha(\omega)T(g)$*. Under these assumptions, there is a unique GEO* $(\alpha, T)\colon (\Phi, M) \to (\Psi, N)$ *for which the restriction of* $\alpha\colon \Phi \to \Psi$ *to* Ω *is* $\bar{\alpha}$*.*

Proof. Since the proofs are analogous, we illustrate only how to show statement (2). For every ϕ in Φ, there exist (not necessarily unique) ω in Ω and g in M such that $\phi = \omega g$. The assumption implies that the expression $\alpha(\omega)T(g)$ depends on ϕ and not on the choices of ω and g for which $\phi = \omega g$. Thus by mapping ϕ in Φ to $\alpha(\omega)T(g)$ in Ψ, we obtain a well defined function also denoted by $\alpha\colon \Phi \to \Psi$. The pair (α, T) is the desired MEO. The uniqueness is a consequence of Proposition 11. □

For example assume (Φ, M) is a transitive group incarnation and (Ψ, N) is a group incarnation. Choose an element ω in Φ. Recall that any such element is a basis of (Φ, M). Fix a group homomorphism $T\colon M \to N$. Then any GEO $(\alpha, T)\colon (\Phi, M) \to (\Psi, N)$ is uniquely determined by the element $\alpha(\omega)$ in Ψ. Thus by choosing a basis element ω in Φ, we can identify the collection of GEOs of the form $(\alpha, T)\colon (\Phi, M) \to (\Psi, N)$ with a subset of Ψ. To describe this subset explicitly, we apply Proposition 12.2. It states that there is a GEO $(\alpha, T)\colon (\Phi, M) \to (\Psi, N)$ (necessarily unique) such that $\alpha(\omega) = \psi$ if and only if the following implication holds: if $\omega = \omega g$, then $\psi = \psi T(g)$. The collection $M_\omega := \{g \in M \mid \omega = \omega g\}$ is the isotropy subgroup of ω consisting of all the elements in M that fix ω. Thus GEOs of the form $(\alpha, T)\colon (\Phi, M) \to (\Psi, N)$ can be identified with the subset of all the elements in Ψ whose isotropy group contains $T(M_\omega)$.

5.6 Decomposition

We want to see that in order to study an incarnation, we have to focus only on the blocks given by the action of the set of transformations on the data set. Let (Φ, M) be an incarnation of a data set Φ. Consider its quotient Φ/M, which is a partition of Φ, and the block incarnations (Ψ, M) for every block Ψ in Φ/M (see Section 5.4). Let X be the domain of Φ. Recall that the domain of the data set $\coprod_{\Psi \in \Phi/M} \Psi$ is given by the disjoint union $\coprod_{\Psi \in \Phi/M} X$, and that this data set consists of functions $\coprod_{\Psi \in \Phi/M} X \to \mathbb{R}$ whose restrictions to all but one summands X in $\coprod_{\Psi \in \Phi/M} X$ is the 0 function and the restriction to the remaining summand belongs

to the corresponding block of the partition Φ/M. Define:

$$M' = \left\{ \coprod_{\Psi\in\Phi/M} g\colon \coprod_{\Psi\in\Phi/M} X \to \coprod_{\Psi\in\Phi/M} X \quad | \quad g \in M \right\}.$$

Then $M' \subset \mathrm{End}_{\coprod_{\Psi\in\Phi/M}\Psi}(\coprod_{\Psi\in\Phi/M} X)$. We call $(\coprod_{\Psi\in\Phi/M}\Psi, M')$ the diagonal incarnation. Define $T\colon M \to M'$ to map $g\colon X \to X$ in M to $\coprod_{\Psi\in\Phi/M} g$ in M'. Define $\alpha\colon \Phi \to \coprod_{\Psi\in\Phi/M}\Psi$ to map ϕ to the function $\coprod_{\Psi\in\Phi/M} X \to \mathbb{R}$ whose restriction to the summand X corresponding to the block $[\phi]$ is ϕ and that maps all other summands to 0. Note that both of the functions α and T are bijections. Furthermore they form a SEO between (Φ, M) and $(\coprod_{\Psi\in\Phi/M}\Psi, M')$.

Proposition 13. *The SEO* $(\alpha, T)\colon (\Phi, M) \to (\coprod_{\Psi\in\Phi/M}\Psi, M')$ *is an isomorphism.*

5.7 Grothendieck graphs

In this section we explain a convenient data structure to encode incarnations of data sets.

Definition 65. A **Grothendieck graph** is a triple (V, M, E) consisting of a finite set V whose elements are called vertices, a finite set M whose elements are called colors or operations, and a subset $E \subset V \times M \times V$ whose elements are called edges, such that, for every vertex v in V, the following composition is a bijection:

$$(\{v\} \times M \times V) \cap E \hookrightarrow E \hookrightarrow V \times M \times V \xrightarrow{\mathrm{pr}_M} M.$$

This last condition assures that, for every v in V and g in M, there is a unique element in V, denoted by vg, such that (v, g, vg) is an edge in E. For example let (Φ, M) be an incarnation of a data set Φ. Define:

$$E_{\Phi,M} := \{(\phi, g, \psi) \in \Phi \times M \times \Phi \mid \phi g = \psi\}.$$

Then the triple $(\Phi, M, E_{\Phi,M})$ is a Grothendieck graph. We think about this graph as a convenient data structure representing the incarnation (Φ, M).

Grothendieck graphs are also convenient to represent SEOs.

Definition 66. Define a **morphism between Grothendieck graphs** (V, M, E) and (W, N, F) to be a pair of functions $\alpha\colon V \to W$ and $T\colon M \to N$ such that, if (v, g, w) belongs to E, then $(\alpha(v), T(g), \alpha(w))$ belongs to F. Such a morphism is denoted by $(\alpha, T)\colon (V, M, E) \to (W, N, F)$.

Componentwise composition defines a category structure on the collection of Grothendieck graphs and we use the symbol GGraph to denote this category. If

$(\alpha, T)\colon (\Phi, M) \to (\Psi, N)$ is a SEO, then $(\alpha, T)\colon (\Phi, M, E_{\Phi,M}) \to (\Psi, N, E_{\Psi,N})$ is a morphism between the associated Grothendieck graphs. By assigning to a SEO (α, T) the graph morphism given by the same pair (α, T), we obtain a fully faithful functor from the category **Nirvana** to **GGraph**.

Grothendieck graphs can also be used to encode pseudometric information on incarnations. A pseudometric on a Grothendieck graph (V, M, E) is a pseudometric d on V such that $d(v, w) \geq d(vg, wg)$ for all v and w in V, and g in M. For example, the pseudometric $\|\phi - \psi\|_\infty$ on Φ is a pseudometric on the graph $(\Phi, M, E_{\Phi,M})$.

A Grothendieck graph (V, M, E) is said to be compatible with a monoid structure on M if $(v,1,v)$ is in E, and whenever (v_0, g_0, v_1) and (v_1, g_1, v_2) belong to E, then so does $(v_0, g_1 g_0, v_2)$. In this case the composition operation given by the association $(v_0, g_0, v_1)(v_1, g_1, v_2) \mapsto (v_0, g_1 g_0, v_2)$ defines a category structure, denoted by $\mathrm{Gr}_M V$, with V as the set of objects and E as the set of morphisms. This category is a familiar Grothendieck construction [39, 96]. For example, the Grothendieck graph associated with a monoid incarnation (Φ, M) is compatible with the monoid structure on M. We think about $\mathrm{Gr}_M \Phi$ as an additional structure on the data set Φ: objects are the measurements in Φ, morphisms are triples $(\phi, g, \phi g)$, where ϕ is in Φ, g is in M, and the composition of $(\phi, g, \phi g)$ and $(\phi g, h, \phi g h)$ is given by $(\phi, gh, \phi gh)$.

Definition 67. A **contravariant functor** indexed by a Grothendieck graph (V, M, E) with values in a category $\mathcal{C}$, denoted by $P\colon (V, M, E) \to \mathcal{C}$, is by definition a sequence of objects $\{P(v) \mid v \in V\}$ and a sequence of morphisms $\{P(v_0, g, v_1)\colon P(v_1) \to P(v_0) \mid (v_0, g, v_1) \in E\}$ in $\mathcal{C}$ subject to: if (v_0, g_0, v_1), (v_1, g_1, v_2), and (v_0, h, v_2) are edges in E, then $P(v_2, h, v_0) = P(v_2, g_1, v_1)P(v_1, g_0, v_0)$.

If (V, M, E) is compatible with a monoid structure on M, then a contravariant functor indexed by (V, M, E) is simply a contravariant functor indexed by the category $\mathrm{Gr}_M V$.

Let (Φ, M) be an incarnation of a data set Φ consisting of measurements on X, and $(\Phi, M, E_{\Phi,M})$ be the associated Grothendieck graph. For every g in M, the function $-g\colon \Phi \to \Phi$, mapping ϕ to ϕg, is geometric and realized by $g\colon X \to X$ (see Section 5.3). Persistent homology leads therefore to the following collections of objects and morphisms in $\mathbf{Tame}\left(\mathbf{Vect}^{[0,\infty)\times\mathbb{R}}\right)$ as explained in Section 5.3:

$$\left\{PH_d^\Phi(\phi) \mid \phi \in \Phi\right\},$$

$$\left\{PH_d^{-g}(\phi)\colon PH_d^\Phi(\phi g) \to PH_d^\Phi(\phi) \mid (\phi, g, \phi g) \in E_{\Phi,M}\right\}.$$

These sequences form a functor $PH_d^\Phi\colon (\Phi, M, E_{\Phi,M}) \to \mathbf{Tame}\left(\mathbf{Vect}^{[0,\infty)\times\mathbb{R}}\right)$ also referred to as the persistent homology functor of the incarnation (Φ, M).

Let $(\alpha, T)\colon (W, N, F) \to (V, M, E)$ be a morphism and $P\colon (V, M, E) \to \mathcal{C}$ be a functor. The following sequences of objects and morphisms in $\mathcal{C}$ form a contravariant functor denoted by $P(\alpha, T)\colon (W, N, F) \to \mathcal{C}$ and called the **composition** of (α, T) with P:

$$\{P(\alpha(v)) \mid v \in V\},$$

$$\{P(w_0, g, w_1)\colon P(\alpha(w_1)) \to P(\alpha(w_0)) \mid (w_0, g, w_1) \in F\}.$$

For example, let $(\mathrm{id}_\Phi, i)\colon (\Phi, M) \to (\Phi, \mathrm{End}_\Phi(X))$ be the canonical SEO (see Section 5.5). Consider the induced morphism of the associated Grothendieck graphs:

$$(\mathrm{id}_\Phi, i_M)\colon (\Phi, M, E_{\Phi,M}) \to (\Phi, \mathrm{End}_\Phi(X), E_{\Phi,\mathrm{End}_\Phi(X)}).$$

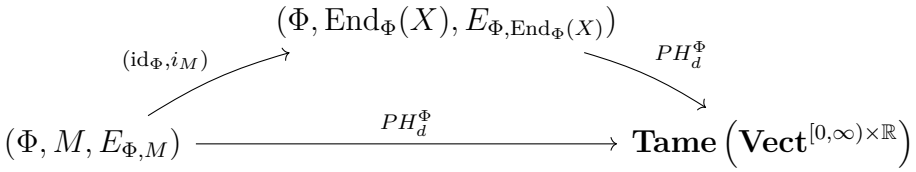
Consider also the persistent homology of the universal incarnation:

$$PH_d^\Phi\colon (\Phi, \mathrm{End}_\Phi(X), E_{\Phi,\mathrm{End}_\Phi(X)}) \to \mathbf{Tame}\left(\mathbf{Vect}^{[0,\infty)\times\mathbb{R}}\right).$$

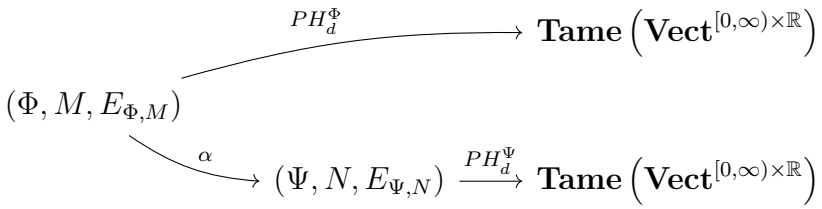
The composition of these two functors coincides with the persistent homology of the incarnation (Φ, M):

$$PH_d^\Phi\colon (\Phi, M, E_{\Phi,M}) \to \mathbf{Tame}\left(\mathbf{Vect}^{[0,\infty)\times\mathbb{R}}\right).$$

In this way we obtain a commutative diagram:



Such a commutativity does not hold for arbitrary SEOs. Consider a SEO $(\alpha, T)\colon (\Phi, M) \to (\Psi, N)$. We can form two functors indexed by the graph $(\Phi, M, E_{\Phi,M})$:



These functors rarely coincide. However, in the case (α, T) is geometric, the morphisms $PH_d^\alpha(\phi)\colon PH_d^\Psi(\alpha(\phi)) \to PH_d^\Phi(\phi)$ (see Section 5.3), for all ϕ in Φ, form a natural transformation.

5.8 Conclusions

We defined datasets as set of measurements on a finite space X. We have seen how multiparameter persistent homology can be used to study the functions in a dataset. We enriched datasets with the action of a set of the endomorphisms of X on the set of functions defined on X. We defined the incarnations to take this structure into account and introduced the category Nirvana, along with set equivariant operators, to have the proper setting to compare different incarnations of datasets. We have given three examples of SEOs, change of units, domain change and extension from a basis. Lastly, we introduced the structure of Grothendieck graphs to relate data sets' incarnations and set equivariant operators between them. Persistent homology becomes a controvariant functor indexed by a Grothedieck graph. We see that given a SEO between Grothendieck graphs, it is not always possible to compare the associated persistent homologies, therefore set equivariant operators and persistent homology yield different information.

Bibliography

[1] Alexandre Abraham et al. “Machine learning for neuroimaging with scikit-learn”. In: *Frontiers in Neuroinformatics* 8 (2014), p. 14. ISSN: 1662-5196. DOI: 10.3389/fninf.2014.00014. URL: https://www.frontiersin.org/article/10.3389/fninf.2014.00014.

[2] Henry Adams et al. “Persistence Images: A Stable Vector Representation of Persistent Homology”. In: *Journal of Machine Learning Research* 18.8 (2017), pp. 1–35. URL: http://jmlr.org/papers/v18/16-337.html.

[3] Muscoloni Alessandro and Cannistraci Carlo Vittorio. “Leveraging the nonuniform PSO network model as a benchmark for performance evaluation in community detection and link prediction”. In: *New Journal of Physics* 20.6 (2018), p. 063022.

[4] Uri Alon. “Biological networks: the tinkerer as an engineer”. In: *Science* 301.5641 (2003), pp. 1866–1867.

[5] Miroslav Andjelković, Bosiljka Tadić, and Roderick Melnik. “The topology of higher-order complexes associated with brain hubs in human connectomes”. In: *Scientific reports* 10.1 (2020), pp. 1–10.

[6] Andrea Baronchelli et al. “Networks in cognitive science”. In: *Trends in cognitive sciences* 17.7 (2013), pp. 348–360.

[7] Alain Barrat et al. “The architecture of complex weighted networks”. In: *Proceedings of the national academy of sciences* 101.11 (2004), pp. 3747–3752.

[8] Danielle S Bassett and Olaf Sporns. “Network neuroscience”. In: *Nature neuroscience* 20.3 (2017), p. 353.

[9] Danielle Smith Bassett and ED Bullmore. “Small-world brain networks”. In: *The neuroscientist* 12.6 (2006), pp. 512–523.

[10] Mattia G. Bergomi et al. “Towards a topological-geometrical theory of group equivariant non-expansive operators for data analysis and machine learning”. In: *CoRR* abs/1812.11832 (2018). arXiv: 1812.11832. URL: http://arxiv.org/abs/1812.11832.

[11] Mattia G. Bergomi et al. “Towards a topological-geometrical theory of group equivariant non-expansive operators for data analysis and machine learning”. In: *Nature Machine Intelligence* 1.9 (2019), pp. 423–433. DOI: 10.1038/s42256-019-0087-3.

[12] Jean-Daniel Boissonnat, Siddharth Pritam, and Divyansh Pareek. “Strong Collapse for Persistence”. In: *26th Annual European Symposium on Algorithms (ESA 2018)*. Ed. by Yossi Azar, Hannah Bast, and Grzegorz Herman. Vol. 112. Leibniz International Proceedings in Informatics (LIPIcs). Schloss Dagstuhl–Leibniz-Zentrum fuer Informatik, 2018, 67:1–67:13. ISBN: 978-3-95977-081-1. DOI: 10.4230/LIPIcs.ESA.2018.67.

[13] Mireille Boutin and Gregor Kemper. “On reconstructing n-point configurations from the distribution of distances or areas”. In: *Adv. in Appl. Math.* 32.4 (2004), pp. 709–735. ISSN: 0196-8858. DOI: 10.1016/S0196-8858(03)00101-5.

[14] Peter Bubenik. “Statistical Topological Data Analysis using Persistence Landscapes”. In: *Journal of Machine Learning Research* 16.3 (2015), pp. 77–102. URL: http://jmlr.org/papers/v16/bubenik15a.html.

[15] Peter Bubenik and Jonathan A. Scott. “Categorification of Persistent Homology”. In: *Discrete Comput. Geom.* 51.3 (2014), pp. 600–627. DOI: 10.1007/s00454-014-9573-x.

[16] Ed Bullmore and Olaf Sporns. “Complex brain networks: graph theoretical analysis of structural and functional systems”. In: *Nature reviews neuroscience* 10.3 (2009), pp. 186–198.

[17] Oleksiy Busaryev et al. “Annotating simplices with a homology basis and its applications”. In: *Scandinavian workshop on algorithm theory*. Springer. 2012, pp. 189–200.

[18] G. Carlsson et al. “Axiomatic construction of hierarchical clustering in asymmetric networks”. In: *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*. 2013, pp. 5219–5223.

[19] Gunnar Carlsson. “Topology and data”. In: *Bull. Amer. Math. Soc. (N.S.)* 46.2 (2009), pp. 255–308. ISSN: 0273-0979. DOI: 10.1090/S0273-0979-09-01249-X.

[20] Gunnar Carlsson and Afra Zomorodian. “The theory of multidimensional persistence”. In: *Discrete Comput. Geom.* 42.1 (2009), pp. 71–93. ISSN: 0179-5376. DOI: 10.1007/s00454-009-9176-0. URL: https://doi.org/10.1007/s00454-009-9176-0.

[21] Nicholas J. Cavanna, Oliver Kiselius, and Donald R. Sheehy. “Computing the Shift-Invariant Bottleneck Distance for Persistence Diagrams”. In: *Proceedings of the Canadian Conference on Computational Geometry*. 2018.

[22] Wojciech Chacholski et al. *Landscapes of data sets and functoriality of persistent homology*. 2020. arXiv: 2002.05972 [math.AT].

[23] Frédéric Chazal, Ruqi Huang, and Jian Sun. "Gromov–hausdorff approximation of filamentary structures using reeb-type graphs". In: *Discrete & Computational Geometry* 53.3 (2015), pp. 621–649.

[24] Frédéric Chazal, Vin de Silva, and Steve Oudot. "Persistence stability for geometric complexes". In: *Geom. Dedicata* 173.1 (2014), pp. 193–214. DOI: 10.1007/s10711-013-9937-z.

[25] Chao Chen and Daniel Freedman. "Hardness Results for Homology Localization". In: *Discrete & Computational Geometry* 45.3 (Apr. 2011), pp. 425–448. ISSN: 1432-0444. DOI: 10.1007/s00454-010-9322-8. URL: https://doi.org/10.1007/s00454-010-9322-8.

[26] Samir Chowdhury and Facundo Mémoli. *Distances and Isomorphism between Networks and the Stability of Network Invariants*. 2017. arXiv: 1708.04727 [cs.DM].

[27] David Cohen-Steiner, Herbert Edelsbrunner, and John Harer. "Extending persistence using Poincaré and Lefschetz duality". In: *Found. Comput. Math.* 9.1 (2009), pp. 79–103. ISSN: 1615-3375. DOI: 10.1007/s10208-008-9027-z. URL: https://doi.org/10.1007/s10208-008-9027-z.

[28] David Cohen-Steiner, Herbert Edelsbrunner, and John Harer. "Stability of Persistence Diagrams". In: *Discrete & Computational Geometry* 37.1 (2007), pp. 103–120. ISSN: 1432-0444. DOI: 10.1007/s00454-006-1276-5. URL: https://doi.org/10.1007/s00454-006-1276-5.

[29] Vittoria Colizza et al. "The role of the airline transportation network in the prediction and predictability of global epidemics". In: *Proceedings of the National Academy of Sciences* 103.7 (2006), pp. 2015–2020.

[30] Don Coppersmith and Shmuel Winograd. "Matrix multiplication via arithmetic progressions". In: *Journal of Symbolic Computation* 9.3 (1990). Computational algebraic complexity editorial, pp. 251–280. ISSN: 0747-7171. DOI: https://doi.org/10.1016/S0747-7171(08)80013-2. URL: http://www.sciencedirect.com/science/article/pii/S0747717108800132.

[31] Michele D'Amico, Patrizio Frosini, and Claudia Landi. "Using matching distance in size theory: A survey". In: *International Journal of Imaging Systems and Technology* 16.5 (2006), pp. 154–161. ISSN: 08999457. DOI: 10.1002/ima.20076.

[32] Philip I. Davies and Nicholas J. Higham. "Numerically stable generation of correlation matrices and their factors". In: *BIT Numerical Mathematics* 40.4 (2000), pp. 640–651. ISSN: 00063835. DOI: 10.1023/A:1022384216930.

[33] Cecil Jose A. Delfinado and Herbert Edelsbrunner. "An incremental algorithm for Betti numbers of simplicial complexes on the 3-sphere". In: *Computer Aided Geometric Design* 12.7 (1995). Grid Generation, Finite Elements, and Geometric Design, pp. 771–784. ISSN: 0167-8396. DOI: https://doi.org/10.1016/0167-8396(95)00016-Y. URL: http://www.sciencedirect.com/science/article/pii/016783969500016Y.

[34] Cecil Jose A. Delfinado and Herbert Edelsbrunner. "An incremental algorithm for Betti numbers of simplicial complexes on the 3-sphere". In: *Computer Aided Geometric Design* 12.7 (1995). Grid Generation, Finite Elements, and Geometric Design, pp. 771–784. ISSN: 0167-8396. DOI: https://doi.org/10.1016/0167-8396(95)00016-Y.

[35] Tamal K Dey, Tianqi Li, and Yusu Wang. "Efficient algorithms for computing a minimal homology basis". In: *Latin American Symposium on Theoretical Informatics*. Springer. 2018, pp. 376–398.

[36] Tamal Dey, Jian Sun, and Yusu Wang. "Approximating Loops in a Shortest Homology Basis from Point Data". In: *Proceedings of the Annual Symposium on Computational Geometry* (Sept. 2009). DOI: 10.1145/1810959.1810989.

[37] Michel Marie Deza and Elena Deza. *Encyclopedia of distances*. Fourth. Springer, Berlin, 2016, pp. xxii+756. ISBN: 978-3-662-52843-3; 978-3-662-52844-0. DOI: 10.1007/978-3-662-52844-0. URL: https://doi.org/10.1007/978-3-662-52844-0.

[38] O. Dovgoshey and E. Petrov. "Weak similarities of metric and semimetric spaces". In: *Acta Math. Hungar.* 141.4 (2013), pp. 301–319.

[39] William G. Dwyer and Hans-Werner Henn. *Homotopy theoretic methods in group cohomology*. Advanced Courses in Mathematics. CRM Barcelona. Birkhäuser Verlag, Basel, 2001, pp. x+98. ISBN: 3-7643-6605-2. DOI: 10.1007/978-3-0348-8356-6. URL: https://doi.org/10.1007/978-3-0348-8356-6.

[40] Herbert Edelsbrunner and John Harer. "Persistent homology—a survey". In: *Surveys on discrete and computational geometry*. Vol. 453. Contemp. Math. Amer. Math. Soc., Providence, RI, 2008, pp. 257–282. DOI: 10.1090/conm/453/08802. URL: https://doi.org/10.1090/conm/453/08802.

[41] Herbert Edelsbrunner and John Harer. "Persistent homology-a survey". In: *Surveys on discrete and computational geometry*. Vol. 453. Contemp. Math. Amer. Math. Soc., Providence, RI, 2008, pp. 257–282. DOI: 10.1090/conm/453/08802.

[42] Herbert Edelsbrunner and John L. Harer. *Computational topology*. An introduction. American Mathematical Society, Providence, RI, 2010, pp. xii+241. ISBN: 978-0-8218-4925-5. DOI: 10.1090/mbk/069. URL: https://doi.org/10.1090/mbk/069.

[43] Herbert Edelsbrunner, David G. Kirkpatrick, and Raimund Seidel. "On the shape of a set of points in the plane". In: *IEEE Trans. Inform. Theory* 29.4 (1983), pp. 551–559. ISSN: 0018-9448. DOI: 10.1109/TIT.1983.1056714. URL: https://doi.org/10.1109/TIT.1983.1056714.

[44] Alon Efrat, Alon Itai, and Matthew J. Katz. "Geometry helps in bottleneck matching and related problems". In: *Algorithmica* 31 (2001), p. 2001.

[45] P. Frosini and C. Landi. *Size Theory as a Topological Tool for Computer Vision*. Tech. rep. Pattern Recognition and Image Analysis, 1999, pp. 696–603.

[46] A. G. Ganyushkin and V. V. Tsvirkunov. "On the classification of finite metric spaces". In: *Mat. Zametki* 56.4 (1994), pp. 48–58, 157. ISSN: 0025-567X. DOI: 10.1007/BF02362370.

[47] Xiaoyin Ge et al. "Data Skeletonization via Reeb Graphs". In: *Advances in Neural Information Processing Systems 24* (2011). Ed. by J. Shawe-Taylor et al., pp. 837–845.

[48] Michelle Girvan and Mark EJ Newman. "Community structure in social and biological networks". In: *Proceedings of the national academy of sciences* 99.12 (2002), pp. 7821–7826.

[49] Chad Giusti et al. "Clique topology reveals intrinsic geometric structure in neural correlations". In: *Proc. Natl. Acad. Sci. USA* 112.44 (2015), pp. 13455–13460. ISSN: 0027-8424. DOI: 10.1073/pnas.1506407112.

[50] Chad Giusti et al. "Clique topology reveals intrinsic geometric structure in neural correlations". In: *Proceedings of the National Academy of Sciences of the United States of America* 112.44 (2015), pp. 13455–13460.

[51] Mark S Granovetter. "The strength of weak ties". In: *Elsevier*. Social networks (1977), pp. 347–367.

[52] Alessandro De Gregorio et al. *On the notion of weak isometry for finite metric spaces*. 2020. arXiv: 2005.03109 [math.MG].

[53] Misha Gromov. *Metric structures for Riemannian and non-Riemannian spaces*. English. Birkhäuser Boston, Inc., Boston, MA, 2007, pp. xx+585. ISBN: 978-0-8176-4582-3; 0-8176-4582-9.

[54] Marco Guerra and Alessandro De Gregorio. *GitHub repository* MinScaffold. Version 1.0. Available at https://github.com/marcoguerra192/MinScaffold. 2019.

[55] Marco Guerra et al. "Homological scaffold via minimal homology bases". In: *Scientific Reports* 11.1 (2021), p. 5355. DOI: 10.1038/s41598-021-84486-1. URL: https://doi.org/10.1038/s41598-021-84486-1.

[56] Mitsugu Hirasaka and Masashi Shinohara. *Characterization of finite metric space by their isometric sequences*. 2018. arXiv: 1802.06097 [math.CO].

[57] Danijela Horak, Slobodan Maletić, and Milan Rajković. "Persistent homology of complex networks". In: *Journal of Statistical Mechanics: Theory and Experiment* 2009.03 (Mar. 2009), P03034. DOI: 10.1088/1742-5468/2009/03/p03034.

[58] J. Horton. "A Polynomial-Time Algorithm to Find the Shortest Cycle Basis of a Graph". In: *SIAM Journal on Computing* 16.2 (1987), pp. 358–366. DOI: 10.1137/0216026. URL: https://doi.org/10.1137/0216026.

[59] Benoit Hudson et al. "Topological Inference via Meshing". In: *Proceedings of the Twenty-Sixth Annual Symposium on Computational Geometry*. SoCG '10. Snowbird, Utah, USA: Association for Computing Machinery, 2010, pp. 277–286. ISBN: 9781450300162. DOI: 10.1145/1810959.1811006. URL: https://doi.org/10.1145/1810959.1811006.

[60] Esther Ibáñez-Marcelo et al. "Spectral and topological analyses of the cortical representation of the head position: Does hypnotizability matter?" In: *Brain and behavior* 9.6 (2019), e01277.

[61] Esther Ibáñez-Marcelo et al. "Topology highlights mesoscopic functional equivalence between imagery and perception: The case of hypnotizability". In: *NeuroImage* 200 (2019), pp. 437–449.

[62] Nathan Jacobson. *Basic algebra. I*. Second. W. H. Freeman and Company, New York, 1985, pp. xviii+499. ISBN: 0-7167-1480-9.

[63] S. Kalisnik, V. Kurlin, and D. Lesnik. "A higher-dimensional Homologically Persistent Skeleton". In: *Advances in Applied Mathematics* 102 (2019), pp. 113–142.

[64] Telikepalli Kavitha et al. "A Faster Algorithm for Minimum Cycle Basis of Graphs". In: *Automata, Languages and Programming*. Ed. by Josep Díaz et al. Berlin, Heidelberg: Springer Berlin Heidelberg, 2004, pp. 846–857. ISBN: 978-3-540-27836-8.

[65] K. Keller and E. Petrov. "Ordinal spaces". In: *Acta Math. Hungar.* 160.1 (2020), pp. 119–152. ISSN: 0236-5294. DOI: 10.1007/s10474-019-00972-z. URL: https://doi.org/10.1007/s10474-019-00972-z.

[66] Matthäus Kleindessner and Ulrike von Luxburg. "Lens Depth Function and k-Relative Neighborhood Graph: Versatile Tools for Ordinal Data Analysis". In: *Journal of Machine Learning Research* 18.58 (2017), pp. 1–52. URL: http://jmlr.org/papers/v18/16-061.html.

[67] Anna Katharina Kuhlen, Carsten Allefeld, and John-Dylan Haynes. "Content-specific coordination of listeners' to speakers' EEG during communication". In: *Frontiers in human neuroscience* 6 (2012), p. 266.

[68] V. Kurlin. "A one-dimensional homologically persistent skeleton of an unstructured point cloud in any metric space". In: *Computer Graphics Forum* 34.5 (2015), pp. 253–262. DOI: 10.1111/cgf.12713.

[69] François Le Gall. "Powers of Tensors and Fast Matrix Multiplication". In: *Proceedings of the 39th International Symposium on Symbolic and Algebraic Computation*. ISSAC '14. Kobe, Japan: ACM, 2014, pp. 296–303. ISBN: 978-1-4503-2501-1. DOI: 10.1145/2608628.2608664. URL: http://doi.acm.org/10.1145/2608628.2608664.

[70] Hyekyoung Lee et al. "Discriminative persistent homology of brain networks". In: *2011 IEEE International Symposium on Biomedical Imaging: From Nano to Macro*. IEEE. 2011, pp. 841–844.

[71] Michael Lesnick. "The theory of the interleaving distance on multidimensional persistence modules". In: *Found. Comput. Math.* 15.3 (2015), pp. 613–650. ISSN: 1615-3375. DOI: 10.1007/s10208-015-9255-y. URL: https://doi.org/10.1007/s10208-015-9255-y.

[72] Jacob Leygonie, Steve Oudot, and Ulrike Tillmann. *A Framework for Differential Calculus on Persistence Barcodes*. 2019. arXiv: 1910.00960 [math.AT].

[73] Louis-David Lord et al. "Insights into brain architectures from the homological scaffolds of functional connectivity networks". In: *Frontiers in systems neuroscience* 10 (2016), p. 85.

[74] Pek Y Lum et al. "Extracting insights from the shape of complex data using topology". In: *Scientific reports* 3 (2013), p. 1236.

[75] A. Jaillard M. Termenon C. Delon-Martin and S. Achard. "Reliability of graph analysis of resting state fMRI using test-retest dataset from the human connectome project". In: *Neuroimage* 142.15 (2016), pp. 172–187. DOI: 10.1016/j.neuroimage.2016.05.062.

[76] Rossana Mastrandrea et al. "Organization and hierarchy of the human functional brain network lead to a chain-like core". In: *Scientific Reports* 7.1 (2017), pp. 1–13. ISSN: 20452322. DOI: 10.1038/s41598-017-04716-3.

[77] Facundo Mémoli. "Some Properties of Gromov–Hausdorff Distances". In: *Discrete Comput. Geom.* 48.2 (Sept. 2012), pp. 416–440. ISSN: 1432-0444. DOI: 10.1007/s00454-012-9406-8.

[78] James R. Munkres. *Elements of algebraic topology*. Addison-Wesley Publishing Company, Menlo Park, CA, 1984, pp. ix+454. ISBN: 0-201-04586-9.

[79] Alessandro Muscoloni and Carlo Vittorio Cannistraci. "A nonuniform popularity-similarity optimization (nPSO) model to efficiently generate realistic complex networks with communities". In: *New Journal of Physics* 20.5 (2018), p. 052002.

[80] Mark EJ Newman. "The structure and function of complex networks". In: *SIAM review* 45.2 (2003), pp. 167–256.

[81] Ippei Obayashi. "Volume-Optimal Cycle: Tightest Representative Cycle of a Generator in Persistent Homology". In: *SIAM Journal on Applied Algebra and Geometry* 2.4 (2018), pp. 508–534.

[82] Steve Y. Oudot. *Persistence theory: from quiver representations to data analysis*. Vol. 209. Mathematical Surveys and Monographs. American Mathematical Society, Providence, RI, 2015, pp. viii+218. ISBN: 978-1-4704-2545-6. DOI: 10.1090/surv/209. URL: https://doi.org/10.1090/surv/209.

[83] Romualdo Pastor-Satorras et al. "Epidemic processes in complex networks". In: *Reviews of modern physics* 87.3 (2015), p. 925.

[84] Alice Patania, Giovanni Petri, and Francesco Vaccarino. "The shape of collaborations". In: *EPJ Data Science* 6.1 (Aug. 2017), p. 18. ISSN: 2193-1127. DOI: 10.1140/epjds/s13688-017-0114-8.

[85] G. Petri et al. "Homological scaffolds of brain functional networks". In: *Journal of The Royal Society Interface* 11.101 (2014), p. 20140873. DOI: 10.1098/rsif.2014.0873.

[86] J. C. de Pina. "Applications of shortest path methods". In: *PhD Thesis University of Amsterdam* 1 (1995).

[87] B. Rieck et al. "Clique Community Persistence: A Topological Visual Analysis Approach for Complex Networks". In: *IEEE Transactions on Visualization and Computer Graphics* 24.1 (Jan. 2018), pp. 822–831. ISSN: 2160-9306. DOI: 10.1109/TVCG.2017.2744321.

[88] V. Robins. "Towards computing homology from finite approximations". In: *Proceedings of the 14th Summer Conference on General Topology and its Applications (Brookville, NY, 1999)*. Vol. 24. Summer. 1999, 503–532 (2001).

[89] Martina Scolamiero et al. "Multidimensional persistence and noise". In: *Foundations of Computational Mathematics* 17.6 (2017), pp. 1367–1406. ISSN: 1615-3375. DOI: 10.1007/s10208-016-9323-y.

[90] Ann E Sizemore et al. "Cliques and cavities in the human connectome". In: *Journal of computational neuroscience* 44.1 (2018), pp. 115–145.

[91] Ann Sizemore, Chad Giusti, and Danielle S Bassett. "Classification of weighted networks through mesoscale homological features". In: *Journal of Complex Networks* 5.2 (2017), pp. 245–273.

[92] David I. Spivak. *Category Theory for the Sciences*. The MIT Press, 2014. ISBN: 0262028131.

[93] Bosiljka Tadić, Miroslav Andjelković, and Milovan Šuvakov. "Origin of Hyperbolicity in Brain-to-Brain Coordination Networks". In: *Frontiers in Physics* 6 (2018), p. 7. ISSN: 2296-424X. DOI: 10.3389/fphy.2018.00007. URL: https://www.frontiersin.org/article/10.3389/fphy.2018.00007.

[94] Andrew Tausz, Mikael Vejdemo-Johansson, and Henry Adams. "JavaPlex: A research software package for persistent (co)homology". In: *Proceedings of ICMS 2014*. Ed. by Han Hong and Chee Yap. Lecture Notes in Computer Science 8592. 2014, pp. 129–136.

[95] Maïté Termenon et al. "Reliability of graph analysis of resting state fMRI using test-retest dataset from the Human Connectome Project". In: *Neuroimage* 142 (2016), pp. 172–187.

[96] R. W. Thomason. "Homotopy colimits in the category of small categories". In: *Math. Proc. Cambridge Philos. Soc.* 85.1 (1979), pp. 91–109. ISSN: 0305-0041. DOI: 10.1017/S0305004100055535. URL: https://doi.org/10.1017/S0305004100055535.

[97] Katharine Turner. "Rips filtrations for quasimetric spaces and asymmetric functions with stability results". In: *Algebr. Geom. Topol.* 19.3 (2019), pp. 1135–1170. ISSN: 1472-2747. DOI: 10.2140/agt.2019.19.1135. URL: https://doi.org/10.2140/agt.2019.19.1135.

[98] Fernando Vega-Redondo. "Complex social networks". In: *Cambridge University Press* (2007).

[99] Afra Zomorodian and Gunnar Carlsson. "Computing persistent homology". In: *Proceedings of the Annual Symposium on Computational Geometry* (2004), pp. 347–356. DOI: 10.1145/997817.997870.

This Ph.D. thesis has been typeset by means of the TEX-system facilities. The typesetting engine was pdfLATEX. The document class was `toptesi`, by Claudio Beccari, with option `tipotesi=scudo`. This class is available in every up-to-date and complete TEX-system installation.