Simplicial Data Analysis: theory, practice, and algorithms

ScuDo

Scuola di Dottorato ~ Doctoral School

WHAT YOU ARE, TAKES YOU FAR

Doctoral Dissertation

Doctoral Program in Mathematics (29^{th}cycle)

Simplicial Data Analysis

theory, practice and algorithms

By

Alice Patania

******

Supervisor(s):

Prof. Francesco Vaccarino, Supervisor

Dott. Giovanni Petri, Co-Supervisor

Doctoral Examination Committee:

Prof. Ginestra Bianconi , Referee, Queen Mary University of London, U.K.

Prof. Annalisa Marzuoli, Referee, Universitá di Pavia, Italy

Prof. Federica Galluzzi, Universitá di Torino, Italy

Prof. Gianfranco Casnati, Politecnico di Torino, Italy

Prof. Emilio Musso, Politecnico di Torino, Italy

Politecnico di Torino

2017

Declaration

I hereby declare that, the contents and organization 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.

Alice Patania
2017

* This dissertation is presented in partial fulfillment of the requirements for **Ph.D. degree** in the Graduate School of Politecnico di Torino (ScuDo).

A mia nonna Maria,
sei stata la luce nei miei giorni piú bui

Acknowledgements

I strongly believe that there is no way that I would have made it to this point without fantastic people that dedicated their time and expertise to make me the researcher I am today.

Before I start a little disclaimer for all of you that supported me emotionally during these parst 4 years. I will pour my heart out to you at the end of this thesis. For now, allow me to thank the individuals without whose help and dedication this work would have never been written.

First and foremost, I cannot thank enough my supervisors Prof. Vaccarino and Dott. Petri, for their continuous support and constructive critique. They have encouraged me to take chances with my research and gave me countless opportunities which made me grow not only as a scientist, but as a pers on. Thank you, your advice on both research as well as on my career have been priceless. I would also like to thank my collaborators: Jean-Gabriel Young, Prof. Lloyd, Dott. Rebentrost for their invaluable assistance. Applying mathematical methods in so many different fields can be challenging, and their support was fundamental for the success of my work. I would like to thank the I.S.I. Foundation and the project S3: "Steering Socio-technical Systems" from Compagnia di San Paolo who financed my Ph.D. Fellowship and have given me so many fantastic opportunities, enabling me to carry out my research without any financial worry. I am eternally grateful to Prof. Mario Rasetti and the I.S.I. Foundation for the fantastic environment they were able to create, surrounding myself by absolutely fantastic people who have made the last four years completely splendid. I want to thank all the researchers that have been part of the I.S.I. family during my time there. You have all been a tremendous inspiration, and i will always be grateful for all the fun we have had in the

last four years. Lastly, I would like to acknowledge the valuable comments and suggestions of the reviewers, which have improved the quality of this thesis.

Abstract

Simplicial complexes store in discrete form key information on a topological space, and have been used in mathematics to introduce combinatorial and discrete tools in geometry and topology. They represent a topological space as a collection of 'simple elements' (such as vertices, edges, triangles, tetrahedra, and more general simplices) that are glued to each other in a structured manner. In the last 20 years, they have been a basic tool in computer visualization and topological data analysis. Topological data analysis has been used mainly as a qualitative method, the problem being the lack of proper tools to perform effective statistical analysis. Coming from well established techniques in random graph theory, the first models for random simplicial complexes have been introduced in recent years, none of which though can be used effectively in a quantitative analysis of data. We introduce a random model which fixes the size distribution of facets and can be successfully used as a null model. Another challenge is to successfully identify a simplicial complex which can correctly encode the topological space from which the initial data set is sampled from. The most common solution is to build nesting simplicial complexes, and study the evolution of their features. A recent study uncovered that the problem can reside in making wrong assumption on the space of data. We propose a categorical reasoning which enlightens the cause leading to these misconceptions. The construction of the appropriate simplicial complex is not the only obstacle one faces when applying topological methods to real data. Available algorithms for homological features extraction have a memory and time complexity which scales exponentially on the number of simplices, making these techniques not suitable for the analysis of 'big data'. We propose a quantum algorithm which is able to track in logaritmic time the evolution of a quantum version of well known homogical features along a filtration of simplicial complexes.

Contents

Introduction **1**

1 Simplicial Complexes in Data Analysis **6**

1.1 Abstract simplicial complex 7

1.2 Constructing Simplicial Complexes from data 11

1.2.1 Metric case 12

1.2.2 Non-metric case 16

1.3 Random simplicial complexes 18

1.3.1 Generative models 19

1.3.2 Descriptive models 25

2 Simplicial Configuration Model **28**

2.1 Configuration model for pure simplicial complexes 28

2.2 Simplicial configuration model 33

2.2.1 Correctness of the model 34

2.2.2 Empirical results 38

2.3 Generating random simplicial complexes 44

2.3.1 Constraints on the sequences 44

2.4 Future work 49

2.4.1 Existence of cordless cycles 50

3 Weighted graphs and P-Persistent homology 52

3.1 Basic Notions 54

3.1.1 The category of topological spaces 54

3.1.2 The category of simplicial complexes 57

3.1.3 The categories of graphs 58

3.2 P-weighted graphs and P-persistent objects 60

3.2.1 Equivalence 61

3.2.2 Adjunctions 63

3.3 Application to homology: multi-persistent homology 69

3.3.1 Considerations on topological strata 75

3.3.2 Conclusions 77

4 Quantum algorithm for persistent homology 78

4.1 Persistent Homology 80

4.1.1 Expliciting homology maps 81

4.2 Quantum construction of a simplicial complex 84

4.2.1 Quantum notation 84

4.2.2 Simplex quantum state 85

4.3 Quantum algorithm for persistent homology 88

Conclusions 93

References 96

Appendix Computational complexity 103

Introduction

> *[...]a theory that does not lead to the solution of concrete and interesting problems is not worth having. Conversely, any really deep problem tends to stimulate the development of theory for its solution.*
>
> Sir Michael Atiyah, Advice to a Young Mathematician

Throughout history, mathematics has been providing a language capable of making difficult problems understandable and manageable, and for these reasons it has become an efficient source of concepts and tools constituing the backbone of all scientific disciplines. Moreover abstract concepts from logic, algebra, and geometry have found new concrete use with the advent of the computer and the birth of programming. In this thesis we are going to focus on the application to computer science of one of the most versatile algebraic tools of the last centuries: the simplicial complex. Simplicial complexes were first introduced in 1895 by Poincaré in his seminal work "Analysis Situs" [87] as a simplicial decomposition (triangulation) of a manifold, and they are now not only a fundamental construction in combinatorial topology, but also the secret behind every 3D rendering and image recognition software [59, 90].

Simplicial complexes are elementary objects built from such simple polyhedra as points, line segments, triangles, tetrahedra, and their higher dimensional analogues glued together along their faces. Since the late 1800s they have been used to store in discrete form key information on a topological space and

to transform complicated topological problems into more familiar algebraic ones with the introduction of simplicial homology (we refer to Aleksandrov [2] for a beautiful account on the birth of combinatorial topology). Their use in computer science has changed drastically with the advent of Topological Data Analysis [41–43, 38, 21, 19, 20], which uses techniques from computational and algebraic topology to extract information from high-dimension, incomplete and noisy data-sets.

In this work we are going to focus on the theory (chapter 3), practice (chapter 2) and algorithms (chapter 4) of the application of simplicial complexes to data analysis. For each aspect, we are going to introduce original results and insights which are able to shade light on underdeveloped applications for TDA, and further advance the available tool set.

Outline of the thesis

The main intuition of TDA is that data is sampled from a topological space, and the shape of this space is important to better understand the data. To study the shape of the underlying space of data, TDA methods aim to construct a simplicial complex or a filtration of simplicial complexes from the original data, which encodes information on the shape of the underlying space. In Chapter 1 we define the concept of a simplicial complex, and introduce the basic mathematical constructions of simplicial complexes. We then proceed to survey the most suitable methods of construction, distinguishing if the data set can be considered sampled from a metric, or a non-metric space. These topological tools allow for a new type of explorative analysis of data which is able to reveal structures that were unobtainable through other approaches. The

field of topological data analysis has been growing rapidly in the last fifteen years, and its applications have led to discoveries in various fields: genomics [76, 83], sensor analysis [31, 30, 29, 47], brain connectomics [48, 49], fMRI data [84, 65], network science [85, 86], just to name a few.

With the increasing popularity of topological analysis it has become necessary to build sounder statistical foundations. Therefore, the first original contribution in this thesis is to develop a null model[1] of simplicial complexes capable of differentiating between meaningful results and random noise. In recent years, researchers have introduced the first proposals for random simplicial complexes coming from well established techniques in random graph theory: the Erdös-Renyi random graph model [61, 67, 55, 62, 56, 57, 27], preferential attachment [13–15], the exponential random graph model [96], configuration model [28, 94]. Even though these models are good for theoretical studies, they present some shortcomings when used as null models of real data sets, which we present extensively in chapter 1 before introducing in chapter 2 the first original contribution of this thesis: the simplicial configuration model.

The simplicial configuration model builds on the work by Courtney and Bianconi [28] where the authors introduced a configuration model for simplicial complexes, which uses the intuition that the one-mode projection of a bipartite graph can be encoded as a simplicial complex. In their paper, Courtney and Bianconi analyzed in detail the ensemble of the configuration model for simplicial complexes with constant facet size. Our contribution generalizes their approach to general simplicial complexes. Moreover, we show how our

[1] In this context, by null model we mean an instance of a random simplicial complex which matches the original complex in some of its structural properties.

random generative model can be used successfully as a null model for the size distribution of maximal facets in a general simplicial complex.

It is easy to see how the analysis we just introduced are significative if and only if we can safely assume that the starting simplicial complex successfully incorporates the features of the dataset. However, there is seldom a way to unequivocally test whether a simplicial complex correctly encodes the topological space from which the initial data set is sampled from. For this reason, the most common approach is to build nesting simplicial complexes from the data set, and study the evolution of their features across the filtration [42, 43, 38, 19]. This technique is known as persistent homology, and in recent years has become one of the prominent tools in TDA.

Following the example of many researchers [17], that in recent years worked on using category theory to build a stronger foundation for topological data analysis and highlighten its faults, in chapter 3 we start exploring the concept of persistence, and prove the adjuctions and categorical equivalences that dictate the relationships between the categories involved in topological data analysis (topological spaces, graphs, simplicial complexes) [78]. We show how these results dissuade from using the intrinsic metric of graphs (shortest path length metric) for constructing simplicial complexes, backing the empirical results in [86].

In the last chapter of this thesis, we dive into the computational problems that might arise when applying these methods to real data. In fact, the construction of an appropriate simplicial complex is not the only obstacle one faces when applying topological methods to real data. Available algorithms for homological features extraction have a memory and time complexity which

scales exponentially on the number of simplices, making these techniques not suitable for the analysis of 'big data'. With an eye to this problem, we formulated an approach based on quantum computation [81]. Expanding on a method by Lloyd et al. [63], we propose a quantum algorithm which is able to track in logarithmic time the evolution of a quantum version of well known homological features along a filtration of simplicial complexes.

Chapter 1

Simplicial Complexes in Data Analysis

In this chapter we introduce some basic notions from classical algebraic topology that are widely used in topological data analysis. We define the most common types of simplicial complexes (sec. 1.1), and how to construct them from data (sec. 1.2). Finally in section 1.3 we give a thorough introduction to existing models for random simplicial complexes.

Unless otherwise stated, we consider to be working on a field k, that we suppose to be algebraically closed. Moreover, we suppose all the algebras to be associative and all the modules to be left module if not otherwise specified.

1.1 Abstract simplicial complex

Simplicial complexes are one of the most intuitive concepts in mathematics. They are built from such simple polyhedra as points, line segments, triangles, tetrahedra, and their higher dimensional analogues glued together along their faces. Even if their intuition is very geometric, they can easily be generalized to abstract mathematical objects. An **abstract simplicial complex** X is a collection of finite sets such that for every $\sigma \in X$ then for all $\tau \subseteq \sigma$, $\tau \in X$. The sets in X are called simplices, the dimension of a simplex $\sigma \in X$ is $\dim(\sigma) = card(\sigma) - 1$; the dimension of X is the maximum dimension of the simplices it contains.

The proper subsets of a simplex are called its **faces** and, if τ is a proper face of σ, then σ is a proper **coface** of τ. A **facet** is any simplex in a simplicial complex that is not a face of any other simplex. A simplicial complex is called **pure** if all its facets have the same dimension. The vertex set of X is the union of all the simplices it contains, $V = \cup_{\sigma \in X} \sigma$.

Examples of abstract simplicial complexes

We now introduce some concepts related to simplicial complexes which will be useful in the future chapters.

Subcomplex A **subcomplex** X' of X is an abstract simplicial complex such that the vertex set of X' is contained in the vertex set of X, and, for every simplex σ in X', σ belongs to X as well. An important type of subcomplex is the k**-skeleton** $X_{(k)}$ of a simplicial complex X which contains all the simplices

of dimension at most k in X, $X_{(k)} = \{\sigma | \dim \sigma \leq k\}$. In particular, the 1-skeleton of a simplicial complex can be considered as an undirected graph, since it contains only 1-simplices (edges) and 0-simplices (vertices); from this moment onward we will then refer to $X_{(1)}$ as the **underlying graph** of X. It is easy to see how the 1-simplices and 0-simplices contained in any simplex in X, form cliques (complete subgraphs) in $X_{(1)}$. Beware that the opposite it is not necessary true, that is, a clique in the underlying graph of X is not always a representation of a simplex in X. The simplicial complexes for which this property is verified are called **flag complexes**.

Clique complex It is easy to see how to use this definition to construct flag complexes from graphs. Given a graph G, the **clique complex** $\mathrm{Cl}(G)$ is the simplicial complex whose simplices are all the cliques contained in G. A set of vertices $S \in V(G)$ of a graph is said to be independent, if for all $v, w \in S$ the edge $(v, w) \notin E(G)$. It is easy to see that the independent sets of G are the cliques in the graph complement of G, i.e. the graph that has the same vertices as G and all the edges (v, w) such that $(v, w) \notin G$. The **independent complex** $\mathrm{Ind}(G)$ of a graph G is the clique complex of the graph complement of G.

Simplicial complex subdivisions The simplicial complexes we introduced above are used in practice to describe the structural composition of the original simplicial complex. There might be the need in practice to construct a simplicial complex which has the same geometry and topology of the original one, but with a finer resolution. That is, a simplicial complex which contains all the simplices of the original one. A simple example of such a construction is the

stellar subdivision. Let σ be a simplex of X, the **stellar subdivision of** X **at** σ is the abstract simplicial complex $\mathrm{Sd}_X(\sigma)$, where the set of vertices $V(\mathrm{Sd}_X(\sigma)) = V(X) \cup \hat{\sigma}$ where $\hat{\sigma}$ is the new vertex indexed by σ. If σ is already a vertex we have that $\hat{\sigma} = \sigma$ and no new vertex is introduced. Every simplex that does not contain σ as a subset is still a simplex in $\mathrm{Sd}_X(\sigma)$. Otherwise, if a simplex τ in X contains σ as a subset then $\eta \cup \{\hat{\sigma}\} \in \mathrm{Sd}_X(\sigma)$, where η is the difference as sets between τ and σ.

The stellar subdivision is a construction which acts locally on the simplices that contain σ. A global construction of a finer complex is the barycentric subdivision. The **barycentric subdivision** of X is an abstract simplicial complex $\mathrm{Bd}(X)$, where the set of vertices in $\mathrm{Bd}(X)$ is indexed by the non empty simplices in X, and

$$\mathrm{Bd}(X) = \{\{\sigma_1, \ldots, \sigma_t\} | \sigma_1 \supset \cdots \supset \sigma_t,\ \sigma_i \in X,\ t \geq 1\} \cup \{\emptyset\} \tag{1.1.1}$$

It is easy to observe that $\mathrm{Bd}(X)$ is a flag complex.

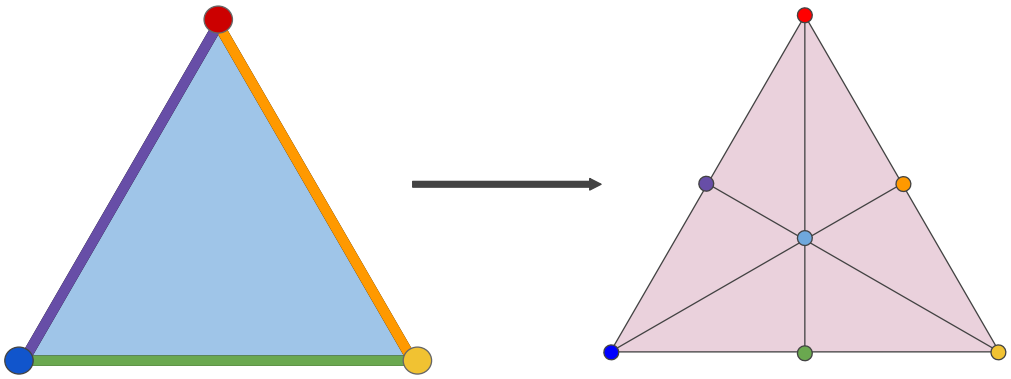

Fig. 1.1 Baricentric subdivision of a 2-simplex.

A typical application of this refinement process is in 3D imagining, when trying to increase the level of details in a picture. It can be proved that

taking the barycentric subdivision can be accomplished by a sequence of stellar subdivisions, which are performed locally and thus provide a computationally more economic construction.

Order Complex A partial ordered set, or poset, is a set P endowed with a binary relation $\leq$ which is reflexive (for all $a \in P$, $a \leq a$), antisymmetric (for all $a, b \in P$, if $a \leq b$ and $b \leq a$ then $a = b$), and transitive (for all $a, b, c \in P$,if $a \leq b$ and $b \leq c$ then $a \leq c$). An abstract simplicial complex X can then be considered a poset, since the inclusion of simplices is a partial order relation on X. One can also construct a simplicial complex from any poset P, considering as simplices all finite chains (i.e. finite totally ordered subsets) of P. The simplicial complex defined in this way is called **order complex** of P. To better clarify the concept, we give some examples of order complexes:

1. The order complex of a totally ordered set A is a simplex $\Delta(A)$.

2. Let $n \in \mathbb{N}$ and let $\mathcal{B}_n$ be the set of all subsets of n partially ordered by inclusion. One can see that the order complex $\Delta(\mathcal{B}_n)$ is isomorphic to the barycentric subdivision of an $(n-1)$-simplex.

3. An abstract simplicial complex X can then be considered a poset, since the inclusion of simplices is a partial order relation on X. Then, the barycentric subdivision of X is the order complex of X considered as a poset.

In chapter 3 we will go in more detail on the key role the order complex plays when analysing data with topological methods.

1.2 Constructing Simplicial Complexes from data

There are two main applications for simplicial complexes in data analysis: the representation of relations, and the discretization of data spaces. In the former, representing relational data, the vertices of the complex are the data points and a $k-$simplex represents a relation between the $k+1$ vertices it contains. In this application, the structure of the simplicial complex comes directly from the dataset itself. In the latter, objects are a topological discretization of the underlying space of data, that is, an object that is topological equivalent to the space from which we sampled the data.

In this section we will show some common simplicial complexes constructed from point clouds, distinguishing the cases in which the dataset is supposed to be sampled from a metric space and those in which it is not.

Before going on with the explanation, we introduce the nerve of an open cover, a construction at the core of the techniques we are going to describe in this section. Let X be a paracompact topological space, that is a topological space in which to every open cover $\mathcal{U}$ one can associate a new cover $\mathcal{V}$ of X with a locally finite index set, such that every set in $\mathcal{V}$ is contained in some set in $\mathcal{U}$. To each open cover $\mathcal{U} = \{U_\alpha\}_{\alpha \in A}$ of X, we can associate an abstract simplicial complex $\mathcal{N}(\mathcal{U})$ called the **nerve** of $\mathcal{U}$. The simplicial complex is constructed in the following way: there is a vertex v_α for each open set U_α in cover. A set of $k+1$ vertices spans a k-simplex whenever the $k+1$ corresponding open sets U_α have non empty intersection. Obviously the simplicial complex thus constructed is determined by the chosen cover.

The following theorem gives the motivation for which the nerve is such a common tool for constructing simplicial complexes from data. Under appropriate hypothesis, the nerve of and open cover has the same homotopy as the underlying topological space, that is, intuitively, it has the same "shape".

Theorem 1.2.1 (Nerve Theorem,[Hatcher, §4G.3])**.** *Let X be a topological space and $\mathcal{U} = \{U_\alpha\}_{\alpha \in A}$ a countable open cover of X.*
If, for every $\emptyset \neq S \subseteq A$, $\cap_{s \in S} U_s$ is contractible or empty then $N(\mathcal{U})$ is homotopically equivalent to X.

1.2.1 Metric case

In applications it is quite common to work with large sets of points sampled from a metric space X. For example, to scan surfaces in 3D one uses time-of-flight cameras which compute the nearest point on the surface from the sensor position along a given direction. A 3D scan may then be composed by a very large set of points corresponding to different directions from the sensor and different sensor positions.

In this section we will consider the data points as sampled from a metric space (X, m), where m is a metric, bestowed with the standard topology where the base $\mathcal{B}$ is made of open balls of radius ε centered in $v \in X$, $\mathcal{B} = \{B_\varepsilon(v) | \varepsilon \in \mathbb{R}^+, v \in X\}$ where $B_\varepsilon(v) = \{u \in X | m(u, v) < \varepsilon\}$.

Čech Complex The Čech Complex is the nerve of an open covering of the data set where the open sets are open balls $B_\varepsilon(v)$ of radius ε centred in $v \in X$. If we denote V as the set of $v \in X$ such that $B_\varepsilon(v) \in \mathcal{U}$ then we can write $X = \cup_{v \in V} B_\varepsilon(v)$. The simplicial complex we obtain through this covering

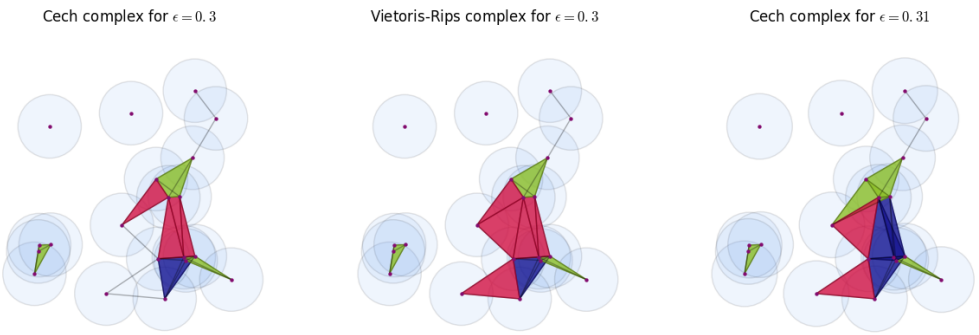

Fig. 1.2 A vizualization of the relation between Čech and Vietoris Rips at different scales described by proposition 1.2.2.

is called **Čech complex** $\check{C}(V, \varepsilon)$. More concretely, the vertices of the Čech complex are the points in V and $k+1$ points spans a k-simplex if all the ε-balls centred in them have non-empty intersection. Since open balls are contractible, from Theorem 1.2.1 follows that the Čech complex captures the topology of the covering.

It is important to notice though that the resulting shape of the covering, and thus that of the Čech complex, depends on the choice of the radius of the open balls that form the covering. When the parameter is very small, smaller than the minimum distance between the points, the corresponding Čech complex is only composed by the points of V. Conversely, when the parameter value is larger than the cloud diameter the corresponding complex contains all the possible subsets of V. The supposition here is that for a parameter $\bar{\varepsilon}$ the open cover of the dataset is also and open cover of the space X underlying the data satisfying 1.2.1. Finding the optimal radius ε for which this happens is very difficult. In recent years, new methods in topological data analysis have been introduced to avoid taking this decision, which we will look at in detail in Chapter 3.

Vietoris-Rips complex The Čech complex is a very good discretization of the space X, but it is rarely used in practice because it is computationally heavy to construct. This is due to the fact that its construction requires the computation of $2^{|A|}$ intersections, where $\mid A \mid$ is the number of open sets in the considered cover, which is equal to the number of vertices. Even though the computational complexity can be reduced with clever algorithms, the process is still a very expensive one. This is why less precise but more computational efficient simplicial complexes were introduced.

The Vietoris-Rips complex is popular in topological analysis thanks to the ease of its construction in every dimension. It is not a nerve as the other previously presented complexes, but it is the clique complex of a particular graph. Let X be a metric space with metric d, a **Vietoris-Rips complex** $VR(X, \varepsilon)$ is the simplicial complex which has as vertex set X and such that $\{x_0, \ldots, x_k\}$ spans a k-simplex if and only if $d(x_i, x_j) \leq \varepsilon$ for all $0 \leq (i-j) \leq k$.

Proposition 1.2.2 ([60]). *Let X be a metric space with metric d, the following inclusions are satisfied :*

$$\check{C}(X, \varepsilon) \subseteq VR(X, 2\varepsilon) \subseteq \check{C}(C, 2\varepsilon) \tag{1.2.1}$$

This proposition justifies the use of the Vietoris-Rips complex as a good-enough substitute of the Čech Complex. Applying this technique solves the computational problems, since it only requires to check if the distances are below a certain threshold for each pair of data points, and there are $\binom{n}{2}$ such matchings.

Witness complexes The methods introduced above produce simplicial complexes whose vertex set has the same size as the underlying set of point cloud data. When working with big data sets, these constructions produce simplicial complexes which are untreatable. In 2004 De Silva and Carlsson, using ideas motivated by the usual Delaunay complex in Euclidean space, introduced a new method [29], the witness complex, which produces topologically equivalent simplicial complexes with a smaller vertex set.

Let X be a metric space, $\varepsilon > 0$ a parameter and lets suppose we have a finite subset $\mathcal{L} \subseteq X$ that we denote as **landmark set**. For every $x \in X$ let m_x be the minimum distance between x and the set $\mathcal{L}$, we shall define the **strong witness complex** as the complex $W^s(X, \mathcal{L}, \varepsilon)$ which has as vertex set $\mathcal{L}$ and $\{l_0, \ldots, l_k\}$ spans a k simplex if and only if there exists $x \in X$ (called witness) such that $d(x, l_i) \leq m_x + \varepsilon \ \forall i$.

This definition is too constraining creating a very small set of strong witnesses, in order to obtain a finer simplicial complex a weaker version of this construction was introduced. Let X be a topological space, point set $\mathcal{L} \subseteq X$, $\Lambda = \{l_0, \ldots, l_k\}$ finite subset of $\mathcal{L}$. Then $x \in X$ is called a **weak witness** for Λ, if for all $i = 0, \ldots, k$, $d(x, l) \geq d(x, l_i)$ for all $l \in \mathcal{L} \backslash \Lambda$. Moreover for $\varepsilon \geq 0$ we will say that x is an ε-weak witness for Λ if $d(x, l) + \varepsilon \geq d(x, l_i)$ for all $i = 0, \ldots, k$ and $l \in \mathcal{L} \backslash \Lambda$.

We can now construct the **weak witness complex** $W^w(X, \mathcal{L}, \varepsilon)$ and we will say that $\Lambda = \{l_0, \ldots, l_k\}$ spans a k-simplex if and only if Λ and all its faces have a weakness ε. This complex depends on the choice of the landmark set. There is no preferred way to choose an optimal landmark set. It is common practice to work with different set of landmarks and see if the results are

replicable. It is however one of the most popular constructions when working with large data sets since it contains a less simplices than the Vietoris-Rips complex, but it is as reliable in approximating the topology of the space.

1.2.2 Non-metric case

Depending on the data set we are working on, it is not always straight forward to know what metric the underlying space has, or whether a metric exists at all. In applications, we rarely have the certainty that the underlying space is a metric space. This is the reason why we introduce now two methods for constructing simplicial complexes which do not require the existence of a metric.

Dowker Complex The **Dowker complex** was first introduced in [24] and named after C. H. Dowker [34] who compared two simplicial complexes constructed from a binary relation. It is defined as follows: let L, W be two sets and $\Lambda : L \times W \to \mathbb{R}$ be a function. For $a \in \mathbb{R}$ consider the simplicial complex $\mathrm{Dow}(\Lambda, a)$ with vertex set L and simplices σ determined by:

$$\exists\, w \in W \text{ such that } \Lambda(l, w) \leq a \text{ for all } l \in \sigma \tag{1.2.2}$$

Remark 1.2.2.1. The simplicial complexes introduced in Subsection 1.2.1 can all be considered as examples of Dowker Complex where as function Λ is considered the metric of the metric space to which the data belongs to, and the sets L, W are chosen accordingly.

Remark 1.2.2.2. The Dowker complex can be seen as the nerve of the covering $\mathcal{U} = \{U_l\}_{l \in L}$ where $U_l = \{w \in W | \Lambda(l, w) \leq a\}$ [24].

Chazal et al. show in [24] that Dowker's theorem implies that for every $a \in \mathbb{R}$, $\mathrm{Dow}(\Lambda, a)$ and $\mathrm{Dow}(\Lambda^T, a)$ have the same homotopy type, where $\Lambda^T : W \times L;\ (w, l) \mapsto \Lambda(l, w)$. On the stability of Dowker complexes we refer the reader to [24].

Mapper Algorithm Mapper was first introduced by Singh, Mémoli, and Carlsson in [89, 88] as part of an algorithm for 3D Object Recognition. Since then it has become one of the most used topological analysis method, and it is at the core of all the software products developed by Ayasdi (www.ayasdi.com).

Mapper is a computational method for extracting simplicial complexes from high-dimensional data sets, it does so combining the notion of the nerve complex with a partial clustering of the data guided by a set of functions. The power of this method comes from the fact that is not dependent on any particular clustering algorithm. Let X and Y be two topological spaces, $f : X \to Y$ be a continuous map. Consider a covering $\mathcal{U} = \{U_\alpha\}_{\alpha \in A}$ be a finite open cover of Y. The **Mapper** construction arising from these data is defined to be the nerve simplicial complex of the pullback cover: $M(\mathcal{U}, f) = \mathcal{N}(\{f^{-1}(U_\alpha)\})$. This construction is quite general. It encompasses both the Reeb graph and merge trees at once [89]. In the past year a number of theoretical improvements have been achieved: the stability of the mapper was proved in late 2015 [22], and a multiscale version was introduced early this year [32].

1.3 Random simplicial complexes

Leveraging the constructions introduced in the previous section, topological analysis can give qualitative information about data sets, which is not readily available by other means. In classic data analysis, the information gathered from explorative methods is used to develop hypotheses and tests that can interpret these data in a more rigorous manner. This step is usually achieved through the construction of random models able to model a specific feature of the data, that can be used to construct characteristic null hypothesis. In recent years, many researchers have tried developing such a random model for simplicial complexes and develop a statistical framework in the context of topological data analysis [55–57, 15, 13, 14, 28, 67, 61, 62, 22, 32].

In this section we review the existing models of random simplicial complexes. All these models use ideas from random graph theory, but do this coming from two different perspectives which we divide as generative or descriptive.

Generative models are algorithms which describe how to generate a network using some probabilistic rules for connecting the nodes. These models are also called growing network models, because the algorithm can be devided in steps in which a node or an edge is added to the existing network. The simplest and most studied example is the Erdös-Rényi random graph(ER), or standard random graph: given n nodes, edges are added to the graph with probability p. Another prominent example is the preferential attachment model: a node is added to the graph at time t and connected to one of the existing nodes with a probability dependent on the node degree. These ER models are the inspiration for the two categories of random simplicial complexes model

which we will describe in section 1.3.1. Generative models can help understand the fundamental organizing principles behind real networks and explain their qualitative behaviour, because they provide a mechanistic rule to build the network.

A descriptive model is explicitly defined as an ensemble $(\mathcal{G}, \mathbb{P}_\theta)$, where $\mathcal{G}$ is a set of graphs and $\mathbb{P}_\theta$ is the joint probability distribution on $\mathcal{G}$ parametrized by a vector of parameters θ, inferred from the observed network data. Any generative model gives rise to an ensemble $(\mathcal{G}, \mathbb{P})$, where $\mathcal{G}$ is the set of all the graphs the model can generate, and $\mathbb{P}$ is the probability distribution on $\mathcal{G}$; it is usually very difficult to find a closed-form expression for it, and so the ensemble is then sampled using the network generating algorithm. A descriptive model gives a closed-form expression for $\mathbb{P}_\theta$ which can be used for further statistical inference. The most studied descriptive model in the network science community is the exponential random graph, or $p^\star$ model. In 1.3.2 we will describe the only descriptive model available for the study of simplicial complexes, the ERSC.

1.3.1 Generative models

Standard random models

We define as standard random models, the random simplicial complex which tried to extend to higher dimension the concepts behind the Erdös-Rényi graph, also known as the standard random graph model; these includes the random d-complexes by Linial and Meshulam [62, 61, 67], the random clique comple by Kahle [57, 56, 55], and the multi-parameter model by Costa and Farber [27].

Random d-complexes Linial and Meshulam initiated the topological study of random simplicial complexes in [61], introducing a method to construct random pure simplicial complexes of dimension 2. Each random simplicial complex constructed with the model has a complete graph of size n as underlying graph. Then each of the possible 2-simplices is included independently with probability p.

The Linial-Meshulam model can be generalized to d-dimensional pure simplicial complexes [67]. In this model, they start with the simplicial complex that has the complete graph as underlying graph, and every d-clique is a facet of the simplicial complex, then d-cofaces are added independently with probability p.

Random clique complexes By random clique complexes we intend the study of clique complexes constructed from random graphs. This kind of approach has been very popular in recent years [55–57]. The most common random graph used as 1-skeleton is the Erdös-Rényi graph, first introduced in [55]. This approach improves on the Linial-Meshulam model, since the simplicial complex generated this way has no constriction on the dimension of its facet. However, using clique complexes to model real-world relational data can be misleading, as it is not always true that a k-clique in a network represents a k-order relation in the data set. Moreover, the randomness of the simplicial complex is induced completely from the underlying graph, the Erdös-Rényi random graph, whose degree distribution is well approximated by the Poisson distribution, which is very unlikely to come across in real networks [73]. These facts make this model a good theoretical tool, but not very interesting in practice.

Multi-parameter random simplicial complex There is a natural multi-parameter model which generalizes all of the models discussed so far which was first studied in [27]. For every every $d = 1, ...$ let $p_i \in [0, 1] \subset \mathbb{R}$. Then define the multi-parameter random complex as follows. Start with n vertices. Insert every edge with probability p_1, producing an Erdös-Renyi random graph $G(n, p_1)$. Then for every 3-clique in the graph, insert a 2-face with probability p_2, and so on.

This random model is more general and more flexible than the ones introduced above, since in general it does not produce neither a clique, nor a pure simplicial complex. Moreover, it is easy to see that the previous models can be interpreted as particular cases of the multi-parameter model. However, the randomness of this model is induced by the underlying Erdös-Renyi graph. Therefore, as for the case of random clique complex, the resulting degree distribution is still unrealistic, making this model unsuitable for modeling real-world simplicial complexes.

Preferential attachment models

The are a lot of networks that have a scale-free structure. In the late 1990s there was a lot of studies in understanding why. An undirect explanation is that scale-free networks are very robust to link/node deletion. The Barabasi-Albert model [3], inspired from preferential attachment, is the first model able to reproduce this characteristic in random networks.

As in the previous paragraph, we call preferential attachment models those models which use the concept of preferential attachment to generate random simplicial complexes. These models were first introduced in [93] and then

extended and used by Bianconi and Rahmede to describe the evolution of quantum network states [13–15].

Bianconi-Rahmede model The **Bianconi-Rahmede model** is a growing model which constructs pure simplicial complexes of dimension d adding simplices of dimension d to $(d-1)$-simplex already in the complex. The simplicial complex thus created displays non-trivial geometric properties which were studied rigorously in [13]. In the paper, the authors introduce the notion of **saturated simplex** as a simplex of dimension $d-1$ which is face of m d-simplices, where m is a parameter of the network which can be either a natural number or infinite. In the latter case no $(d-1)$-simplex can ever become saturated.
In [15] Bianconi and Rahmede introduce the concept of **generalized degree** of a δ-simplex σ in a simplicial complex X, $k_\delta(\sigma)$ is the number of co-faces of δ of dimension d.

The growing process is initialized at time $t = 1$ from a simplicial complex containing only one d-simplex. At each time a d-simplex is added to an unsaturated $(d-1)$-simplex σ in the simplicial complex with probability p_σ given by:

$$p_\sigma = \frac{a_\sigma \xi_\sigma (1 + n_\sigma)}{Z} \tag{1.3.1}$$

where $a_\sigma = 1$ if σ is a $(d-1)$-simplex already in the complex, and 0 otherwise; $\xi_\sigma = 1$ if σ in unsaturated, and 0 otherwise; $n_\sigma = k_\delta(\sigma) - 1$. The linking probability depends on n_σ, unsaturated simplices with a higher number of co-faces or that are closer to becoming saturated are more likely to be selected

than the others. In the simplicial complexes produced this way, the number of facets scales as the number of nodes.

Bianconi-Rahmede model with flavor In [15] the authors proposed an extension to the Bianconi-Rahmede model were they introduced the *flavor* variable $s = 1, 0, -1$ of the model.

As before the process is initialized at time $t = 1$ simplicial complex is formed by a single d-simplex. At time $t > 1$ d-simplex is added to an existing $(d-1)$-simplex σ in the simplicial complex with probability p_σ given by:

$$p_\sigma^{[s]} = \frac{(1 + s\ n_\sigma)}{Z^{[s]}(t)} \tag{1.3.2}$$

where $Z(t)$ is the normalization factor at time t.

This model generates discrete manifolds with $s = -1$, because $p_\sigma^{[-1]} \geq 0$ and therefore $n_\sigma = 0, 1$, this implies that we can glue a new simplex only to faces that has degree 0. The model generates more general simplicial complexes for the other two flavors. For $s = 0$, $p_\sigma^{[0]} = \frac{1}{Z^{[0]}(t)}$ where $Z^{[0]}(t)$ is the number of d-simplices at step t, this will produce a uniform attachment model. For $s = 1$, $1 + n_\mu = k_{d,d-1}(\mu)$, i.e. the generalized degree of the face, therefore producing a preferential attachment according to the generalized degree. For further information on this process and on the study of the associated generalized degree distributions, we advice reading [15].

Bianconi-Rahmede model with link energy In [13] the authors introduce an extension to the BR model inspire by the Bianconi-Barabasi model [12] which allows for a weight or energy influencing the evolution of the network.

They assign to each node i an energy w_i. The energy of the node is assigned when the node is first added to the complex from a distribution $g(w)$, and does not change during the evolution of the network. An energy ϵ_σ is assigned to each $(d-1)$-face of the simplicial complex given by the sum of the energy of the nodes that belong to σ.

$$\epsilon_\sigma = \sum_{i \in \sigma} w_i \tag{1.3.3}$$

The process is defined as for the BR model with flavor, at time $t = 1$ simplicial complex is formed by a single d-simplex. At time $t > 1$ d-simplex is added to an existing $(d-1)$-simplex σ in the simplicial complex with probability p_σ given by:

$$p_\sigma = \frac{e^{-\beta \epsilon_\sigma}(1 + n_\sigma)}{Z} \tag{1.3.4}$$

Following the approach on networks in [12, 58, 71, 72], each network evolution can be considered as a possible quantum network state. In [14] the authors showed, for the case of discrete manifolds $s = -1$, that the average of the generalized degrees of the δ-faces with energy ϵ follows different statistics (Fermi-Dirac, Boltzmann or Bose-Einstein statistics) depending on the dimensionality δ of the faces and on the dimensionality d of the simplicial complex.

Even though this model has a more realistic generalized degree distribution, it generates only pure simplicial complexes, which can be sometimes limiting. For example in the case of a collaboration data set, where each paper can be described by a simplex and its authors as vertices, restricting one-self to only d-dimensional simplices would mean to limit one-self to only paper with 3 authors. We will now introduce a more general model for random simplicial complexes.

1.3.2 Descriptive models

Exponential random simplicial complexes

Exponential random simplicial complexes are a generalization of exponential random graph models first introduced in [96].

Exponential random graph Let $\mathcal{G}_n$ be the set of graphs with n nodes, $x_1, \ldots, x_r$ be functions on $\mathcal{G}_n$ called the graph observables. Let $\bar{x}_1, \ldots, \bar{x}_r$ be the values of the observables for a network of interest $\bar{G} \in \mathcal{G}_n$.

$$\mathbb{P}_\theta(G) = \frac{\exp^{H_\theta(G)}}{Z(\theta)} \qquad \text{with} \qquad H_\theta(G) = \sum_{i=1}^{r} \theta_i x_i(G) \tag{1.3.5}$$

$H_\theta(G)$ is the hamiltonian of the graph, and $Z(\theta)$ the partition function (the normalization function), and $\theta = (\theta_i, \ldots, \theta_r)$ is a vector of model parameters which satisfy: $\bar{x}_i = -\frac{\partial \ln Z}{\partial \theta_i}$.

Exponential random simplicial complexes Let $\mathcal{C}_n$ be the set of all simplicial complexes on n vertices which can be represented as a tensor product:

$$\mathcal{C}_n = \bigotimes_{d=1}^{n} \mathbf{a}_d \tag{1.3.6}$$

where $\mathbf{a}_d$ is a boolean symmetric tensor of order d with zeros on all its diagonals. These condition requires that $a_{i_1,\ldots,i_d}$ is constant for any permutation of subindices i. The only requirement on $\otimes_{d=1}^{n} \mathbf{a}_d$ is the following compatibility conditions with $\mathcal{C}_n$:

$$a_{\mathbf{i}_d} = 1 \Rightarrow b_{\mathbf{i}_d} = \prod_{k=1}^{d} a_{\mathbf{i}_d^{\hat{k}}} = 1 \tag{1.3.7}$$

where $\mathbf{i}_d^{\hat{k}}$ is the $(d-1)$-long multi-index obtained from $\mathbf{i}_d$ by omitting index i_k. For a simplicial complex $C \in \mathcal{C}_n$ the previous condition define $\mathbf{a}_d$ as what Zuev et al. call an adjacency tensor, where $a_{\mathbf{i}_d} = 1$ if $\{\mathbf{i}_d\} \in C$ and zero otherwise. Let $\mathcal{S} \subset \mathcal{C}_n$ a subset of $\mathcal{C}_n$, $\{x_1, \ldots, x_r\}$ a set of real valued functions on $\mathcal{S}$, and $\{\hat{x}_1, \ldots, \hat{x}_r\}$ a set of real numbers. An **exponential random simplicial complex** $(\mathcal{S}, \{x_i\}, \{\hat{x}_i\})$ is a maximum-entropy ensemble that requires the observables x_i to have expected values $\hat{x}_i$ in the ensemble, i.e. a pair $(\mathcal{S}, \mathbb{P})$, where $\mathbb{P}$ is the probability distribution that maximizes the entropy $S(\mathbb{P}) = -\sum_{C \in \mathcal{S}} \mathbb{P}(C) \ln \mathbb{P}(C)$, and such that:

$$\mathbb{E}_{\mathbb{P}}[x_i] = \sum_{C_{\mathcal{S}}} x_i(C)\ \mathbb{P}(C) = \hat{x}_i \tag{1.3.8a}$$

$$\sum_{C \in \mathcal{S}} \mathbb{P}(C) = 1 \tag{1.3.8b}$$

This model has as special cases the models introduced before in this chapter. Even if the formalism for ERSC is well developed, its application to the production of general simplicial complexes with statistically independent simplices appears to be intractable. For a thorough discussion on the matter and a more detailed introduction to the model please refer to [96].

Conclusions

In this chapter we introduced the concept of abstract simplicial complex. After a brief presentation on the most common simplicial complexes in mathematics, we illustrated how to successfully approximate the topology of the space underlying a data set using simplicial complexes. According to the nature of the space, we defined different methods available for the construction of simplicial complexes

from different data sets. In the last section we focused on random simplicial complexes and their importance to fully develop a topological analysis of data. We showed how the models currently available either restrict themselves to construct particular kind of simplicial complexes (pure complexes [62, 61, 13–15], clique complexes [55]), or their application to general simplicial complexes is intractable [96], or generates structures [27] difficult to encounter in reality.

None the less, the need for a functional null model for simplicial complexes has become more pressing in recent years. To fill this gap, in the next chapter we introduce a new random generative model which constructs simplicial complexes with fixed size distribution. The simplicial configuration model generalizes the configuration models for simplicial complexes by Courtney and Bianconi Courtney and Bianconi [28], and we will show empirically that it can be used successfully to model real world simplicial complexes.

Chapter 2

Simplicial Configuration Model

2.1 Configuration model for pure simplicial complexes

As seen in the previous chapter, one of the reasons why the Erdös-Renyi graph generates unrealistic graphs is the degree distribution which is Poisson distributed when the graph is sparse. The preferential-attachment produces graphs with a scale-free degree distribution which is power law distributed. While it has been shown many times how degree distributions in real world networks are scale-free, the same cannot be said for real-world simplicial complexes and their generalized degree sequences. For this reason Courtney and Bianconi [28] , using the configuration model, developed a method which could generate a simplicial complex with a fixed general degree sequence.

Configuration model The configuration model [69, 11] is a generative model that creates a random graph with a fixed degree sequence, that is, the exact

degree of each vertex in the graph is fixed. This implies that the number of nodes n and the number of edges in the network $m = \frac{1}{2}\sum_i k_i$ are fixed. Suppose to have n vertices with fixed degrees k_i for $i = 1, \ldots, n$, the random graph is constructed in the following way. Each vertex i is provided with k_i edge 'stubs', there are therefore $\sum_i k_i = 2\ m$ stubs. Uniformly at random two stubs are chosen and an edge is created connecting the two of them, until no free stubs are left in the graph. The end result is a graph whose every vertex has the desired degree. The model thus generates a matching between stubs. Each matching can be created with equal probability.

The issue with this model is that the created graph might contain multiple edges or self-loops, or both. Indeed nothing in the generative process prevents two stubs from the same vertex to be paired together, or a pairing of stubs to be chosen more than once. The average number of self-edges and multiedges in the configuration model is a constant as the number of vertices increases, which means that their density tends to zero in the large size limit, we refer the interested reader to Newman [73, §13.2] for a more detailed introduction to the model.

We are now going to introduce the concepts of bipartite graph and show how simplicial complexes can be encoded as "one-mode" projections of bipartite graphs.

Bipartite graph A graph is called bipartite if its vertex set can be partitioned into two disjoint sets F, V such that no two vertices within the same set are adjacent in the graph. Some important properties to recognize if a graph is bipartite In many cases, bipartite graphs are actually studied by projecting them

down onto one set of vertices or the other, called **"one-mode" projections**. In such a projection, two nodes are considered connected if they are second neighbours in the bipartite graph. This construction simplifies the study of the relationships involved in the data, at the cost of discarding some of the information contained in the original bipartite graph. First, each neighbourhood of a node that is removed during the projection forms a clique in the new graph, but the projection graph does not hold any information on which node it represents. Second, it is not always true that a clique in the projection graph is representing a node that was removed from the original bipartite graph. To retain this information we can associate to every "one-mode" projection a simplicial complex.

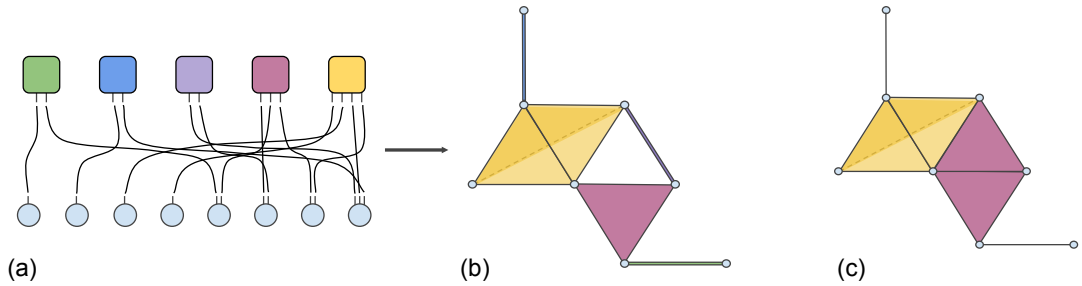

Fig. 2.1 We can see how projecting the bipartite graph in figure (a), we obtain the simplicial complex in figure (b). This is not a flag complex since the 3-clique $[4, 5, 6]$ is not a simplex of the complex. In figure (c) we can see the clique complex of the underlying graph.

Theorem 2.1.1. *Let G be a bipartite graph with vertex sets $\{F, V\}$, G_V its one-mode projections onto the vertex set V. Then it exists a simplicial complex Σ whose underlying graph is G_V.*

Proof. Each neighbourhood of a node that is removed during the projection forms a clique in the new graph, each removed node can be represented as a simplex. The one-mode projection can be seen as substituting one set of

vertices with the simplices that each of them spans, constructing a simplicial complex. □

Note that this process does not necessary produce a flag complex since there can be cliques in the one-mode projection whose vertices are not the neighbourhood of a removed vertex, as it is shown in the example in Figure 2.1. Moreover, this process can be inverted, i.e. any simplicial complex can be seen as the one-mode projection of a bipartite graph.

Theorem 2.1.2. *For every simplicial complex Σ exists a bipartite graph G such that one of its two one-mode projections G_V is the underlying graph of Σ. Moreover, the facet size sequence of Σ is equal to the degree sequence of F.*

Proof. Consider a graph G with vertex set $V \cup F$ where V is the vertex set of Σ and cardinality of F is equal to the number of facets in Σ. For each facet $\sigma \in \Sigma$, $\sigma = [v_0, \ldots, v_k]$, we associate to it a node $f_\sigma \in F$, and connect f_σ to the nodes $v_0, \ldots, v_k$. By construction the projection of G onto the vertex set V will give the desired graph. □

Courtney-Bianconi model Courtney and Bianconi [28] introduced a configuration model for pure simplicial complexes generalizing the approach on hypergraphs introduced by [45]. Their algorithm generates a pure d-simplicial complex with fixed generalized degree sequence $\{k_r\}_{r \leq N}$, where $k_r = k_{d,0}(r)$ is the number of d-simplices incident on node r, and $F = \frac{1}{d+1} \sum_{r=1}^{N} k_r$ is the number of d-simplices or facets of the pure complex (Figure 2.2). The main idea behind their approach is to introduce a set of F auxiliary nodes representing the d-faces of the simplicial complex as seen in the proof of Theorem 2.1.2.

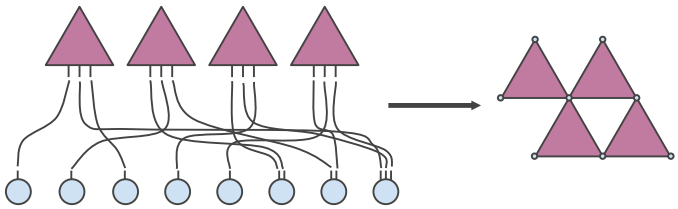

Fig. 2.2 A simple example of the construction of a simplicial complex according to the Courtney-Bianconi model. Each auxiliary node has the same number of stubs which are randomly matched with the stubs in the original node set. One can then obtain a regular simplicial complex by projecting the resulting bipartite graph onto the original node set.

Their algorithm then proceeds defining a configuration model for bipartite graphs as follows:

1. k_r stubs are placed on each node $r = 1, \ldots, V$, and $d+1$ stubs are placed in each auxiliary node $\mu = 1, \ldots, F$. At this step each stub is unmatched.

2. a set of $d+1$ unmatched random stubs of the nodes is chosen with uniform probability. Without loss of generality we assume that the stubs belong to the set of nodes $(r_0, \ldots, r_d)$.

3. if the nodes $(r_0, \ldots, r_d)$ are all distinct, and no auxiliary node μ is matched with the same set of nodes, then with uniform probability an unmatched auxiliary node $\bar{\mu}$ is chosen and matched with the nodes $(r_0, \ldots, r_d)$. Otherwise the process is re-initialized.

4. if all stubs are matched, then a simplicial complex is constructed projecting the auxiliary nodes onto the original node set.

The rejection procedure step executed at step 3 of the algorithm guarantees that there are no spurious correlations in the structure of the simplicial complex.

In [28] the authors treated in detail the configuration model and the canonical ensemble of simplicial complexes, following the approach used on exponential random simplicial complexes, computing analytically the entropy of the ensembles [96]. We refer the reader to [28] for a more detailed study of the statistical mechanics feature of these ensembles.

2.2 Simplicial configuration model

In this section we introduce the simplicial configuration model (SCM) as the maximally random ensemble that generates simplicial complexes with a fixed sequence of maximal clique sizes $\vec{s} = \{s_i\}_{i=1,..,F}$ and nodes total degrees $\vec{d} = \{d_i\}_{i=1,..,N}$; by a node total degree we mean the number of maximal cliques that contain that node.

We now show that the random bipartite ensemble of [75] can be re-interpreted as generating simplicial complexes with high probability when $N \to \infty$. The general idea is to generate a bipartite graph with a vertex set $\mathcal{F} \cup \mathcal{V}$ where $\mathcal{F} = \{f_1, ..., f_F\}$ represents the set of maximal cliques (or facets) and where $\mathcal{V} = \{v_1, ..., v_N\}$ represents the vertex set of the simplicial complex. We then assign stubs (half-edges) to each face and vertex according to $\vec{s}$ and $\vec{d}$. A random matching of the stubs can then be often interpreted as a simplicial complex. That is, it will contain multi-edges with vanishing probability. By multi-edge, we mean that there is two edges or more connecting a node-vertex $v_i \in \mathcal{V}$ to a node-face $f_j \in \mathcal{F}$. Moreover it is not always true that the facets size distribution of the generated simplicial complex is the same as the initial degree

distribution $\vec{f}$. That is, it will contain fully contained neighbourhoods with vanishing probability, once some amendment to the construction procedure are applied. By fully contained neighbourhoods, we mean that the neighbourhood $\mathcal{N}(f_i)$ of a node-face f_i is completely included in the neighbourhood $\mathcal{N}(f_j)$ of node-face f_j.

The stub matching scheme can be implemented as follows:

1.a Generate a list of length $m = \sum_{i=1}^{N} d_i$ where v_i appears d_i times, for each $i = 1, .., V$;

1.b Generate a list of length $m = \sum_{i=1}^{F} s_i$ where f_i appears s_i times, for each $i = 1, .., F$;

2 Generate two random permutations, X^v and X^f, of each list;

3 Connect X_i^v to X_i^f for $i = 1, ..., m$;

4 If both the inclusion and multi edges constraints are satisfied, accept the graph, otherwise go back to step 2.

The resulting bipartite graph $G(\mathcal{V}, \mathcal{F}; E)$ is then interpreted as a simplicial complex: The neighbours $\mathcal{N}(f_i)$ of f_i are the vertices that form the maximal simplex f_i, for each i, or equivalently, the neighbours $\mathcal{N}(v_i)$ of vertex v_i are the facets in which node v_i appears.

2.2.1 Correctness of the model

We will now show that with high probability the simplicial complex constructed by our model has facet size distribution $\vec{s}$, and total degree $\vec{d}$.

We will start proving that with high probability, the simplicial complex will not contain multi-edges following the work by Newman on the configuration model of bipartite graphs [74]. This is a standard calculation that will serve to illustrate the principles that we will apply in the more involved analysis of the next sections.

Theorem 2.2.1. *The simplicial complex constructed with the simplicial configuration model will not contain multi-edges.*

Proof. The probability that there exist an edge (f_i, v_j) in E is

$$\Pr[(f_i, v_j) \in E] = \frac{s_i d_j}{m}, \tag{2.2.1}$$

since there is a uniform probability d_j/m of finding vertex v_j at any position in X^v, and there is s_i occurrences of f_i in X^f. More generally, there is a probability

$$\Pr[(f_i, v_j) \in E | (f_i, v_j)^\ell \in E] = \frac{(s_i - \ell)(d_j - \ell)}{m - \ell}, \qquad \ell < \min\{s_i, d_j\} \tag{2.2.2}$$

of having the edge $(f_i, v_j) \in E$, provided that is has been already observed ℓ times. The probability that (f_i, v_j) appears ℓ times in E is therefore

$$\Pr[(f_i, v_j)^\ell \in E] = \prod_{\lambda=0}^{\ell-1} \Pr[(f_i, v_j) \in E | (f_i, v_j)^\lambda \in E] \tag{2.2.3}$$

For instance for $\ell = 2$,

$$\Pr[(f_i, v_j)^2 \in E] = \frac{s_i(s_i - 1)\ d_j(d_j - 1)}{m(m-1)} \tag{2.2.4}$$

meaning that the probability that *any* edge appears two times is

$$\Pr[\exists\ \ell = 2\ \forall(i,j)] = \sum_{i=1}^{F}\sum_{j=1}^{N}\Pr[(f_i,v_j)^2 \in E] = \sum_{i,j}\frac{s_i(s_i-1)\ d_j(d_j-1)}{m(m-1)} = \\ = \frac{(\mathbb{E}[s^2]-\mathbb{E}[s])(\mathbb{E}[d^2]-\mathbb{E}[d])}{(m-1)}\ . \tag{2.2.5}$$

This goes to zero as $1/m$ with $m \to \infty$. Since the ensemble is sparse in the infinite limit (fixed average degrees as $N \to \infty$), N must scale linearly in m. The probability above there goes to zero as $1/N$ with $N \to \infty$. Moreover, since $\Pr[(f_i,v_j)^{\ell+1} \in E] \leq \Pr[(f_i,v_j)^{\ell} \in E]$ (from Equation (2.2.3)), then triple (or quadruple, etc.) edges are even less likely than double edges, and will vanish at least as rapidly as them. □

For the constructed simplicial complex to have facet size distribution $\vec{s}$. This means that the cliques corresponding to the facet-nodes, in the one-mode projection onto the vertex set V, must not be contained into one another. We show now that with high probability this will not happen.

Lemma 2.2.2. *The probability of inclusion between two facets of dimension* 2 *in a random configuration goes to zero with* $m \to \infty$.

Proof. The probability of constructing a $k-$size simplex $\sigma_k = [v_1,\ldots,v_k]$ is

$$\Pr[\{f_\sigma,v_1),\ldots,(f_\sigma,v_k)\} \in E] = \frac{k!\ \prod_{i=1}^{k} d_i}{m(m-1)\ldots(m-k)} \tag{2.2.6}$$

From the calculation above we can compute the probability that 2 different facets of size 2 are connected to the same nodes.

$$\Pr[\{f_a, v_i), (f_a, v_j)\} \in E] = \frac{2\ d_i(d_j)}{m(m-1)} \tag{2.2.7}$$

$$\Pr[\{f_b, v_i), (f_b, v_j)\} \in E | \{f_a, v_i), (f_a, v_j)\} \in E] = \frac{2\ (d_i-1)(d_j-1)}{(m-2)(m-3)} \tag{2.2.8}$$

The probability that b will be included in a is the following:

$$\begin{aligned}
\Pr[b \subseteq a] &= \sum_{i,j} \frac{4\ d_i(d_i-1)d_j(d_j-1)}{m(m-1)(m-2)(m-3)} \\
&= \frac{4}{m(m-1)(m-2)(m-3)} \left(\frac{1}{2}[\sum_i d_i(d_i-1)][\sum_j d_j(d_j-1)] - \sum_i d_i^2(d_i-1)^2 \right) \\
&= \frac{4}{m(m-1)(m-2)(m-3)} \left(\frac{1}{2}(\mathbb{E}[d^2] - \mathbb{E}[d])^2 - \mathbb{E}[(d^2-d)^2] \right)
\end{aligned} \tag{2.2.9}$$

This goes to zero as m^{-4} with $m \to \infty$. This probability upper bounds the probability of inclusion in a random configuration, which then will also go to zero with $m \to \infty$. □

Theorem 2.2.3. *For every σ, τ maximal simplices of size s_σ, s_τ respectively, with $s_\sigma \leq s_\tau$; we have that*

$$\Pr[\sigma \subseteq \tau] \tag{2.2.10}$$

goes to zero as $m \to \infty$.

Proof. it follows from the lemma above. □

2.2.2 Empirical results

Sampling the from SCM

The generalized line-graph representation of simplicial complexes will be the most useful for discussing the sampling algorithm. In this representation, one associates a vertex $v_i \in \mathcal{V}$ to each vertex of the complex, as well as a vertex $f_j \in \mathcal{F}$ to each of its maximal facets; an edge connects v_i and f_j if $v_i \in f_j$. Each v_i in this line-graph has degree d_i (the number of facets in which it partakes), and each vertex representing a facet has degree s_i (the facet's size). To each of these degrees, one may associate labeled stubs, i.e., distinguishable half-edges stemming from the associated node. We have defined the support of the SCM as any *simplicial* matching of these labeled stubs, i.e., a matching that yields no multiple memberships of a node to a facet, and no inclusion (a facet containing all the vertices of another facet). For incidence degree and size sequences of finite, a random matching of stubs will often contain at least one inclusion or multiple memberships. An efficient sampler is thus necessary to avoid these culprit. We now show how to sample efficiently from this support with the Metropolis-Hasting algorithm.

Metropolis-Hasting algorithm

The Metropolis-Hasting allows the construction of an ergodic Markov chain over the support of the SCM. One can therefore sample from this chain at regular interval in lieu of sampling constructing random instances of the model from scratch. To ensure ergodicity, a move from a matching X to another

matching X' must be accepted with probability

$$a = \min\left\{1, \frac{g(X \to X')}{g(X' \to X)} \frac{\mathbb{P}(X'; \vec{d}, \vec{s})}{\mathbb{P}(X; \vec{d}, \vec{s})}\right\} \tag{2.2.11}$$

where $g(X \to X')$ is the probability of proposing a move from matching X to matching X', and $\mathbb{P}(X; \vec{d}, \vec{s})$ is the likelihood of matching X under the SCM of degree and size sequences $\vec{d}$ and $\vec{s}$.

Our proposal distribution is the following: We pick two random edges from the set of m edges, say, (v_i, f_j) and (v_k, f_ℓ) and replace them by edges (v_i, f_ℓ) and (v_k, f_j). However, if the matching leads to a non-simplicial configuration, then we give this particular proposal a probability of zero. This means that

$$g(X \to X') = \frac{1}{L(X)}\,, \tag{2.2.12}$$

where $L(X)$ is the number of "legal" configuration in the neighborhood of matching X. Thus, a random move will always be accepted with probability 1, and this move consists of reconnecting two stubs such that the resulting configuration is simplicial. The resulting chain is, again, ergodic by construction.

It is somewhat costly to verify that a matching is simplicial as a whole. One must check that no pair of facet is included, and even clever comparison method will have complexity of the order of $\mathcal{O}(f)$. It is, however, much simpler to check that a move does indeed lead to a simplicial matching, provided that the base matching is itself simplicial. Indeed, the new matching will only differ in two places, such that one only has to check the facets in which vertices v_i and v_j are involved. More specifically, if vertex v_i is disconnected from facet f_k and reconnected to facet f_ℓ (and v_j to f_k), then one needs to check that none

of the d_i facets $\{f(v_i)\}$ of v_i will lead to an inclusion of f_ℓ. If $|f_\ell| \geq |f(v_i)|$, then an inclusion will occur if

$$f(v_i) - f_\ell = \{v_i\} \ , \tag{2.2.13}$$

where the minus sign denotes the set difference.

If $|f_\ell| \leq |f(v_i)|$, then an inclusion will occur if

$$f_\ell - f(v_i) = \{v_j\} \ . \tag{2.2.14}$$

A similar condition obviously holds for the facets of v_j. Since computing the set difference is a linear operation, the condition is testable in $\mathcal{O}(\mathbb{E}[d]\mathbb{E}[s])$ time, which is *much* more efficient, especially in sparse complexes.

The data sets

We applied the simplicial configuration model to the randomization of two data sets depicting the corporate leaderships in Chicago [5], and in Minneapolis-St.Paul [44].

The first example data set we consider is the affiliation data set of corporate directors from 1962 in the Chicago area studied by Barnes and Burkett [5]. This data set contains the affiliation between 24 companies and 20 people in a leadership position in those companies. To construct the simplicial complex we considered as vertices the companies and each facet represents a person in a leadership position in the companies represented by the vertices.

As another example, we considered the affiliation data set of club and board memberships of corporate executive officers studied by Galaskiewicz [44] as part of his research on the urban grants economy in Minneapolis-St.Paul. We followed the approach adopted by Faust [39] and focused on a subset of 26 CEOs and 15 clubs/boards from Galaskiewicz's data. We then constructed a second simplicial complex in as done for the Barnes-Burkett data set. The simplicial complexes from these data sets are represented in figure 2.3.

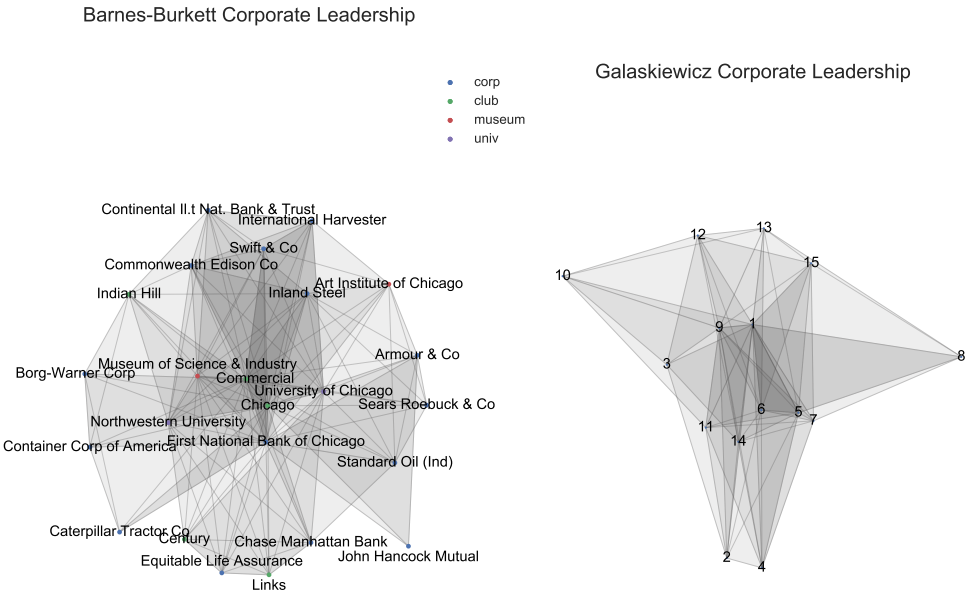

Fig. 2.3 Visualization of the simplicial complex generated from the Barnes-Burkett data set (left), and the Galaskiewicz data set (right). Each simplex (in gray) represents a person in a leadership position in the companies represented by the vertices of the complex.

To further study the structure of the simplicial complexes we constructed from data, we computed its homological cycles. We will introduce in detail the concept of homology in the next chapter (sec. 3.3). We will give now a practical idea of the concept.

Homology of dimension k is a functor that assigns to each simplicial complex a vector space H_k. The generating elements of the vector space H_k are called the homological k-cycles. In low dimensions the homological cycles can be interpreted easily as particular features of the simplicial complex: the 0-cycles represent the connected components, the 1-cycles are cordless cycles not closed by triangles, the 2-cycles are voids closed by a triangle tessellation. These structure can be meaningful in understanding particular features of a data set. These assumptions can then be validated comparing it with the empirical probability distribution of the ensemble generated by our model.

For each constructed simplicial complex we computed its homological cycles (javaPlex library [91]). In Figure 2.4 we show the 1 and 2 dimensional cycles of the simplicial complex constructed from the Barnes-Burkett data sets. The 1-dimensional cycle can be interpreted a set of institutions or corporations for which corporate interlock is not as tightly bound as in the rest of the data set. the only two universities in the data set are present in this cycle. Furthermore, we detected three 2-dimensional cycles in the simplicial complexes. These voids can be interpreted as a set of institutions or corporations for which there is no single person in a leadership position in all of them. It is interesting to notice that these voids are connected to each other through and edge or a triangular face, as shown in figure 2.4.

To validate these results we sampled the ensemble generated by the simplicial configuration model with facet size and incidence degree sequences fixed by the Barnes-Burkett, and the Galaskiewicz data sets. We sampled the two ensembles with the algorithm described above, and computed the homology of each sampled simplicial complex.

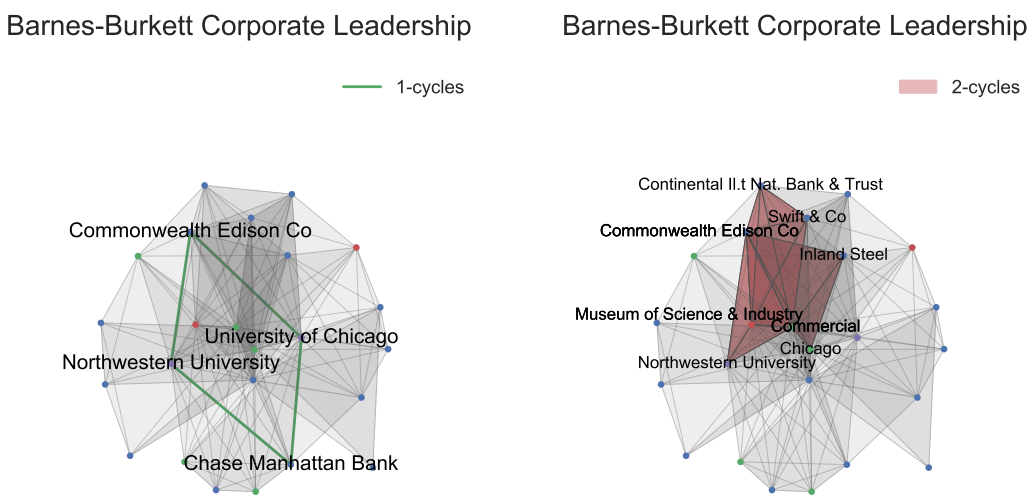

Fig. 2.4 Visualization of the 1-dimensional cycle (in green) and of the 2-dimensional cycles (in red) of the simplicial complex generated from the Barnes-Burkett Corporate Leadership data set.

In figure 2.5 we show the sampling distribution of the number of cycles in a simplicial complex, called Betti number, for the two ensembles. We can see that in both cases the probability to generate a simplicial complex with only one connected component is 1, which might depend on the small sizes of the data sets we considered. Moreover, we can notice how unlikely is the emergence of 1 and 2- dimensional cycles in the configuration generated from the Barne-Burkett data set, validating our findings. On the other hand, from the sampling distribution obtained on the Garlaskewicz data's ensemble we can deduce that the absence of homology in the real simplicial complex is quite probable. Meaning that, in this case, homology might not be the best tool to analyse the data.

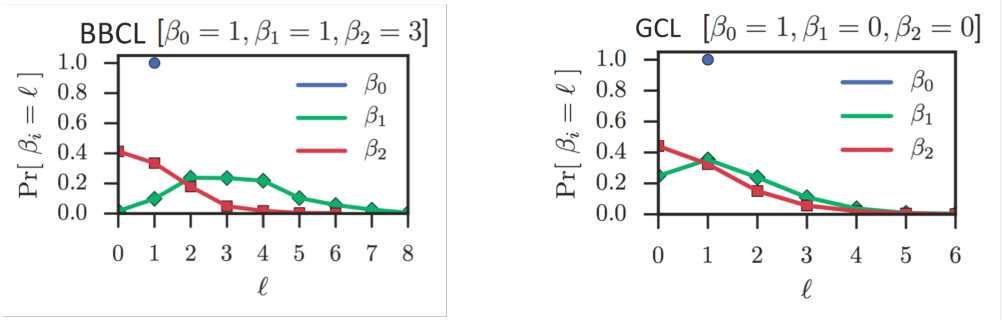

Fig. 2.5 Sampling distribution for the Betti number (dimensions 0,1, and 2) of two different ensembles of simplicial complexes generated with the simplicial configuration model. On the left, the ensemble with fixed size distribution and incidence degree distribution from the Barnes-Burkett Corporate Leadership data set, whose real Betti numbers are $\beta_0 = 1, \beta_1 = 1, \beta_2 = 3$. On the right, the ensemble with fixed size distribution and incidence degree distribution from the Galaskiewicz data set, whose real Betti numbers are $\beta_0 = 1, \beta_1 = 0, \beta_2 = 0$.

2.3 Generating random simplicial complexes

The simplicial configuration model can be used to generate random simplicial complexes with fixed size and incidence degree picked randomly. For the results shown in the previous section to hold, some constraints have to enforced to at least one of the randomly chosen sequences. Here we introduce the constraint necessary to be satisfied by the incidence degree distribution $\vec{d}$, given a facet size sequence $\vec{s}$. From these constraints one can easily deduce the inverse case, when given an incidence degree distribution, one wants to randomly match a facet size sequence.

2.3.1 Constraints on the sequences

This ensemble will be defined as long as the sequences $\vec{d}$ and $\vec{s}$ satisfy a number of constraints, analogous to that of [75].

First we are introducing a number of constraints that ensure that the sequences are compatible with each other, that is for the matching of stubs to be possible the nodes V and the nodes F must have the same number of stubs. That is, if the sequences $\vec{d}$ and $\vec{s}$ are specified directly, they must verify the following equation:

$$\sum_{i=1}^{F} s_i = \sum_{i=1}^{N} d_i \ . \tag{2.3.1}$$

Alternatively, if $\vec{s}$ and $\vec{d}$ are drawn from distributions $\mathcal{P}_s$ and $\mathcal{P}_d$ of expectations $\mathbb{E}[s]$ and $\mathbb{E}[d]$, then we must have

$$F\mathbb{E}[s] = N\mathbb{E}[d] \ , \tag{2.3.2}$$

which can also be written as

$$\alpha\mathbb{E}[s] = (1-\alpha)\mathbb{E}[d] \ , \tag{2.3.3}$$

if we define the degrees to faces ratio $\alpha = F/(N+F)$. This ensures that sequences drawn from $\mathcal{P}_s$ and $\mathcal{P}_d$ will be compatible, on average.

Now we are going to introduce a number of constraints depending on the nature of the sequences, that is, the fact that the node-facets must be maximal for inclusion in the resulting simplicial complex.

Proposition 2.3.1 (Maximum number of vertices)**.** *For any given size sequence $\vec{s}$, the degree sequence $\vec{d}$ which allows the maximal number of vertices V_{max} in the simplicial complex is the one where $d_i = 1$ for all $i = 1, \ldots, V_{max}$, and $V_{max} = m$.*

Up to isomorphisms, the only allowed configuration will be a simplicial complex

with F connected components, one per maximal simplex. The Betti number of the generate simplicial complex will be $\beta_0 = F$, and $\beta_k = 0$ for all $k > 0$.

Proof. Obvious. □

Lemma 2.3.2. *If the size sequence $\vec{s} = \{s\}$ has only one element, then the only configuration allowed is an $s-1$-simplex, and $\vec{d} = \{d_i = 1 | i = 1, \ldots, s\}$.*

Proposition 2.3.3 (Minimum number of vertices)**.** *For any given size sequence $\vec{s}$ with $F > 1$. The minimal number of vertices V_{min} that can be in a simplicial complex with that specific facet size sequence, must satisfy the following inequalities:*

$$\max(\vec{s}) + 1 \leq V_{min} \leq \max(\vec{s}) + F - 1 \tag{2.3.4}$$

Proof. We will prove the inequalities separately.

- $V_{min} \leq max(\vec{s}) + F - 1$. First, we need to prove that for every size sequence $\vec{s}$, there always exists a simplicial complex with $max(\vec{s}) + F$ vertices with size sequence $\vec{s}$.

  Let $\vec{s}$ be a size sequence, and $V = w \cup f$ a vertex set where $card(w) = \max(\vec{s}) - 1$ and $card(w) = F$. For every facet σ of size k, $card(\sigma \cap w) = k - 1$ and there exists a unique $f_\sigma \in f$ such that $f_\sigma \in \sigma$ and $f_\sigma \notin \tau$, where τ is any facet in the simplicial complex not in σ.

  Finally, we need to prove that there exists a size sequence $\vec{s}$, for which $max(\vec{s})+F$ is the minimum number of vertices one would need to construct a simplicial complex with size sequence $\vec{s}$. The equality is verified when $s_i = k$ for all $i = 1, \ldots, F$ with $F > k$.

- $V_{min} \geq max(\vec{s}) + 1$. We notice first that the configuration model needs to have enough vertices to construct the facet of maximal size.

$$V_{\min} \geq \max(\vec{s}) \tag{2.3.5}$$

Since $F > 1$, there are at least two facet-nodes σ, τ with sizes $s_\sigma \geq s_\tau = \max(\vec{s})$. If $V = \max(\vec{s})$, this would imply that $\sigma \subseteq \tau$ against the hypothesis of maximality under inclusion of facets. Therefore, there must exist a vertex $\bar{v} \notin \tau$ such that $\bar{v} \notin \sigma$, which gives us a new lower boundary for $V_{\min}$:

$$V_{\min} \geq \max(\vec{s}) + 1 \tag{2.3.6}$$

The equality is going to be verified only for those sequences $\vec{s} = \{s_k \text{ for } k = 1, \ldots, F\}$ such that a set of $\max(\vec{s})$ elements can have $F - 1$ non overlapping sets of sizes $\{s_k - 1| \text{ for } k = 1, \ldots, F \text{ and } s_k \neq \max(\vec{s})\}$. For the other cases $V_{\min}$ can be computed solving the set of equation 2.3.7.

Let $n_k = \sum_i \delta_{s_i=k}$ be the number of facets of size k. The equality is verified under the following condition:

$$\max(\vec{s}) \geq \sum_{k \neq \max(\vec{s})} x_k \tag{2.3.7}$$

where $x_k = \min\{x|\binom{x}{k-1} \geq n_k\}$ that is, the minimum number of vertices that have at least n_k subsets of cardinality $k - 1$. The simplicial complex satisfying 2.3.7, will have $\max(\vec{s})$ which form the facet of maximal size τ and a vertex $\hat{v} \notin \tau$. For every facet $\sigma \neq \tau$ of size k, $\hat{v} \in \sigma$ and $card(\sigma \cap \tau) = k - 1$.

□

Proposition 2.3.4 (maximal degree constraints)**.** *There can be at most* $min(\vec{s})-1$ *vertices with degree* F*.*

Proof. If there were $min(\vec{s})$ with degree F, this would imply that there were $min(\vec{s})$ in common with all facets. Then the maximality of facets of size $min(\vec{s})$ would not be verified anymore. □

Corollary 2.3.5. *If* $\min(\vec{s}) = 1$ *then* $\max \vec{d} < F$*.*

Proof. Follows from Theorem 2.3.4. □

Remark 2.3.5.1. It is useful to notice that 1-cliques can only be maximal if they contain a single, disconnected vertex. This implies that we must have at least the same number vertices of degree one as maximal 1-cliques. It implies that the sequences must satisfy the following inequality:

$$\sum_{i=1}^{F} s_i \delta_{s_i,1} \leq \sum_{i=1}^{N} d_i \delta_{d_i,1} \ . \tag{2.3.8}$$

Therefore they can be matched beforehand. Therefore, without loss of generality we can assume that :

$$\min(\vec{s}) > 1 \ . \tag{2.3.9}$$

These results enable us to construct random sequences which are simplicial, and justify the use of the simplicial configuration model as a random simplicial complexes generator. Moreover, these results can be used to facilitate the execution of the algorithm introduced in 2.2, performing the following simplification to the input sequences we can avoid the cases that force only one type

of matching, i.e. where $max(d) < F$ and $min(s) > 1$. Theorem 2.3.4 and and remark 2.3.5.1 imply that these additional steps are not going to compromise the ensemble. Therefore, we can suppose the following step to be run before the SCM on the input sequences.

1. Connect all facet-nodes of degree 1 with random chosen vertices with degree 1. Remove the respective elements from the sequences $\vec{s}$, and $\vec{d}$.

2. Connect all vertices with degree F with all the facets. Remove the respective elements from $\vec{d}$, and for every element removed in $\vec{d}$, and correct the number of stubs in $\vec{s}$.

3. If there are facet-nodes of degree 1 in the updated $\vec{s}$, repeat the procedure from 1. Otherwise proceed with the algorithm.

2.4 Future work

In the previous sections, we introduced the Simplicial Configuration Model and proved its correctness. We tested our model on real datasets and we showed empirically how it can be used to validate the existence of homological cycles. Homology is an important tool in topoological data analysis since it can discern the shape of the data set. For this reason, it would be useful to be able to account for the probability of occurance of homological cycles in the ensemble. Regrettably the algebraic nature of the definition of homological cycles, make any analytical computation of their number quite ardous. Therefore, we decided to work on extracting an upper boundary on the number of 1-dimensional homological cycles in the ensamble. To achieve this we intend to compute the

probability to have N cordless cycles in the simplicial complexes generated in the ensemble. We introduce now some preliminary results we obtained on this project.

2.4.1 Existence of cordless cycles

A cordless cycle of length 2ℓ in the bipartite graph is going to be a representative of a cycle of length ℓ in the simplicial complex. The probability $\mathbb{P}(c_\ell)$ of having a cycle c of length ℓ in the simplicial complex is equal to the probability to have a cordless cycle of length 2ℓ in the bipartite graph. This probability is given by $\mathbb{P}(c_\ell) = \mathbb{P}(\text{len}(c) = 2\ell) \cup \mathbb{P}(c \text{ has no chords})$, the probability to have a

.

The probability to have a cycle c is the probability that each node in the cycle is connected to only two others in c:

$$\mathbb{P}(\text{len}(c) = 2\ell) = \prod_{i=1}^{\ell} d_i(d_i - 1) \prod_{k=1}^{\ell} s_k(s_k - 1) \left(\frac{(m - 2\ell)!}{m!} \right)^2 \tag{2.4.1}$$

The probability that a cycle does not have a chord is equal to the probability that the remaining $d - 2$ stubs of a node (chosen with probability $p(d)$) are connected with $d - 2$ stubs randomly chosen from all the nodes not in c, and same goes for the nodes of type s, which gives:

$$\begin{aligned} &\mathbb{P}(c \text{ has no chords} | \text{len}(c) = 2\ell) = \\ &\quad = \left[\binom{m - \sigma_d}{\sigma_s - 2l} \binom{m - 2l}{\sigma_s - 2l}^{-1} \right] \left[\binom{m - \sigma_s}{\sigma_d - 2l} \binom{m - 2l}{\sigma_d - 2l}^{-1} \right] \end{aligned} \tag{2.4.2}$$

where $\sigma_s = \sum_{k=1}^{\ell} s_k$, and $\sigma_d = \sum_{i=1}^{\ell} d_i$.

Therefore the final probability is:

$$\begin{aligned}\mathbb{P}(c_\ell) &= \mathbb{P}(\text{len}(c) = 2\ell | c \text{ has no chords}) = \\ &= \sum_{I,K} \prod_{i=1}^{\ell} d_i(d_i - 1) \prod_{k=1}^{\ell} s_k(s_k - 1) \left[\frac{(m - \sigma_s)!(m - \sigma_d)!}{m!(m - \sigma_s - \sigma_d + 2\ell)!} \right]^2 \end{aligned} \tag{2.4.3}$$

If $m >> \sigma_s$ and $m >> \sigma_d$ then, using Stirling approximation, the probability scales as $\left(\frac{e^4 \ d^2 \ s^2}{m^4}\right)^\ell$ for $m \to \infty$.

We are now able to derive the probability for a simplicial complex in the ensemble to have a cordless cycle of length l. We believe this is a promising result, and we intend to further develop this study to give a more complete description of the occurrence of cordless cycles in the ensemble.

In the next chapter we are going to study in details the relationships between the different categories involved in the topological analysis of weighted networks. We are then going to use our results to give correct guidelines for the construction of appropriate simplicial complexes.

Chapter 3

Weighted graphs and P-Persistent homology

In the previous chapters we focused on how simplicial complexes can properly represent the **shape** of data, but, as we noted in Section 1.2, most of the techniques available for the construction of simplicial complex are highly dependent on the choice of one, or more parameters (e.g. the ball radius in the Vietoris-Rips complexes). In order to study how the parameter choice influences the **shape** of the simplicial complexes, Edelsbrunner et al. [38], Cagliari et al. [18], Carlsson [19] independently introduced the concept of P-persistence.

In this chapter we use the abstract framework of category theory to get a closer look at the key ideas behind P-persistent homology expanding on the work by Bubenik and Scott [17], Chazal et al. [23] in order to obtain a clear understanding of the mathematical structure behind the observation done in Petri et al. [86] that embedding a weighted network into a metric space

generally obfuscates most of its interesting structures, which become evident when one focuses on the weighted connectivity structures without enforcing a metric.

Our aim is to discover an equivalence of categories which highlights the correct approach to apply topological methods to weighted networks. In section 3.1 we introduce the categories involved in this process (topological spaces, simplicial complexes, and graphs) and how they relate to one another. By the end of this section the reader will have a categorical view, summarized in diagram 3.1.4, of how the information of underlying topological spaces is encoded in simplicial complexes and graphs.

$$\mathcal{S} \xrightarrow{\pi} \mathcal{P} \simeq \mathcal{T}_f^0 \xrightarrow{\mathcal{O}} \mathcal{F} \underset{k_1}{\overset{\mathrm{Cl}}{\underset{\cong}{\rightleftarrows}}} \mathcal{G} \qquad \mathcal{O}\circ\pi \tag{3.1.4}$$

In section 3.2 we define the categories that are the main focus of our research: the category of weighted graphs $\mathcal{G}_P$, and that of P-persistent graphs $\mathcal{G}^P$. We then prove the equivalence between the sub-categories of weighted graphs whose morphisms preserve the poset structure of the weights $\overline{\mathcal{G}}_P$, and that of one-critical P-persistent graphs $\mathcal{G}_1^P$.

Furthermore, we show that there exist adjoint functors that describe the relation between the sub-categories involved in the equivalence and the other sup-categories.

Finally in the last section we introduce the concept of homology and use the equivalence found in the previous section to give an explanation of the observations done by Petri et al. in [86].

3.1 Basic Notions

A category consists of a collections of objects and a collection of morphisms. Every morphism has a source object and a target object. If f is a morphism with x as its source and y as its target, we write $f : x \to y$. In a category, we can compose two morphisms $f : x \to y$, and $g : y \to z$ in order to obtain a third morphism of the category $f \circ g : x \to z$. In a category composition is an associative operation and satisfies the left and right unit laws. Moreover, source and target are respected by composition and by the identities.

A functor is a mapping between categories which associates to each object x in $\mathcal{C}$ an object $F(x)$ in $\mathcal{D}$, and to each morphism in $\mathcal{C}$ a morphism in D such that the following conditions hold:

$$F(\mathrm{id}_x) = \mathrm{id}_{F(x)} \tag{3.1.1a}$$

$$F(g \circ f) = F(g) \circ F(f) F(g \circ f) = F(g) \circ F(f) \tag{3.1.1b}$$

for every object x in $\mathcal{C}$, for all morphisms $f : x \to y\, g : y \to z$ in $\mathcal{C}$.

3.1.1 The category of topological spaces

We start with a few considerations on finite topological spaces i.e. topological spaces with a finite number of elements, which we imagine to be given as some sampling taken from a dataset. Finiteness is not a constraint for our purposes, since every application will have a finite data space.

Finite topological spaces form a subcategory, denoted by $\mathcal{T}_f$, of the category $\mathcal{T}$ of topological spaces and continuous maps.

A T_0 space is a topological space such that for any two different points x and y there is an open set which contains one of these points and not the other. Two such points will be called topologically distinguishable. It is clear that this property is highly desirable in order to be able to extract meaningful information from a topological space.

In this paper we will denote by $\mathcal{T}_f^0$ the category of finite T_0-spaces.

From here on when we write topological space we will intend finite topological space if not elsewhere stated.

It may happen that a space we are working with is not T_0, but this difficulty is easily overcome as shown by the following known proposition:

Proposition 3.1.1 ([4, §1.3]). *Let X be a finite space not T_0. Let $X/\sim$ be the Kolmogorov quotient of X defined by $x \sim y$ if it does not exists an open set which contains one of these points and not the other. Then, $X/\sim$ is T_0 and the quotient map $q : X \to x_0$ is a homotopy equivalence.*
The Kolmogorov quotient $X \to X/\sim$ induces a functor from the category of topological spaces to the category of T_0-spaces.

Since homology is defined up to weak homotopy equivalence, the Kolmogorov quotient allows us to restrict our analysis from general topological spaces to T_0-spaces without any loss of information.

Finite T_0-spaces are posets

A partially ordered set, or poset, is a pair $P = (P, \leq)$, where P is a set and $\leq$ is an order relation on it, i.e. a reflexive, antisymmetric, and transitive relation on P. Posets form a category, denoted by $\mathcal{P}$, where morphisms are the order

preserving functions. We will restrict ourselves to study only to finite posets, since in real applications we will always be working with finite sets.

Every poset P is a category on its own, where the objects are the elements of P, and there is a (unique) morphism $x \to y$ if and only if $x \leq y$, for all $x, y \in P$.

Theorem 3.1.2. *There is an isomorphism of categories:*

$$\mathcal{T}_f^0 \cong \mathcal{P}$$

Proof. Let $X \in \mathcal{T}_f^0$, for $x \in X$ let U_x be the intersection of all the closed sets in X that contain x. Then we can give in X an order relation in the following way:

$$x \leq y \leftrightarrow U_x \subseteq U_y \tag{3.1.2}$$

Since X is T_0 this relation is a partial order. In this way we have a correspondence $X \mapsto (X, \leq)$ which induces a functor $\mathcal{T}_f^0 \to \mathcal{P}$.
On the other end, a poset $P \in \mathcal{P}$ is also a topological space via the **Alexandrov topology**. In this topology the closed sets are the **lower sets**: $\Gamma \subset P$ such that $\forall x, y \in P$ with $x \in \Gamma$ and $y \leq x$ implies that $y \in \Gamma$. A poset endowed with this topology satisfies the T_0 condition. The assignment of this topology on P induces a functor $\mathcal{P} \to \mathcal{T}_f^0$ which is left and right inverse of the previous one, that is:

$$\mathcal{T}_f^0 \cong \mathcal{P} \tag{3.1.3}$$

We refer the reader to [4, Ch. 1] for details. □

From now on, we will identify any $X \in \mathcal{T}_f$ with the poset associated to its Kolmogorov quotient i.e., by abuse of notation, we will write

$$X = (X, \leq) = (X/\sim, \leq)$$

where $\leq$ is the order relation given in (3.1.2).

3.1.2 The category of simplicial complexes

Let us consider now abstract simplicial complexes, introduced in Chapter 1. Simplicial complexes form a category, $\mathcal{S}$, where a morphism of simplicial complex is called **simplicial map** and is given by a map on vertices such that the image of a face is again a face. We are going to remind some well known relations between simplicial complexes, topological spaces and posets which will be useful to have a general idea of what are the categorical relation that we exploit when we analyse data through simplicial complexes.

Proposition 3.1.3. *There exists a functor $\mathcal{O} : \mathcal{P} \to \mathcal{S}$ which associates to every poset P a simplicial complex, called the order complex.*

Proof. For every $P \in \mathcal{P}$ we can construct a simplicial complex as follows:

$[x_0, \ldots, x_k] \in \mathcal{O}(P)$ if and only if $x_0 < x_1 < \cdots < x_k$, for all $x_j \in P$.

$\mathcal{O}(P)$ is called the **order complex of** P. □

Every simplicial complex can be made into a topological space by considering it a poset, i.e. $\Gamma \subseteq \Sigma$ is closed if and only if Γ is a simplicial complex. This gives a functor $\pi : \mathcal{S} \to \mathcal{P} \simeq \mathcal{T}_f^0$ by $\pi(\Sigma) = (\Sigma, \subseteq)$ the poset with elements the

simplices in Σ and as partial order the inclusion of simplices.
Given a simplicial complex Σ we write $\mathcal{O}(\Sigma) := \mathcal{O}(\pi(\Sigma))$. The simplicial complex $\mathcal{O}(\Sigma)$ is the barycentric subdivision of Σ.
By abuse of notation we will also write $\mathcal{O}(X) := \mathcal{O}(X/\sim, \leq)$ for all $X \in \mathcal{T}_f$ via the isomorphism in Theorem 3.1.2.

It is well known that Σ and $\mathcal{O}(\Sigma)$, endowed with the Alexandrov topology, are weakly homotopy equivalent. We refer the interested reader to [4] for further details.

3.1.3 The categories of graphs

A reflexive graph is a pair $G = (V, E)$, where V is a finite set whose elements are called vertices, and has an edge (v, v), called **self-loop**, for every vertex $v \in V$, specifically the set E is composed by $E = \Delta_{V\times V} \cup E'$ with $E' \subseteq \binom{V}{2}$. Equivalently, reflexive graphs can be seen as one dimensional simplicial complexes identifying self-loops and vertices with 0-simplices and edges with 1-simplices. We will denote by $\mathcal{G}$ the category with objects reflexive graphs and morphisms the simplicial maps defined via the given identification with one dimensional simplicial complexes.

It should be clear that $\mathcal{G}$ is isomorphic to the full subcategory of $\mathcal{S}$ whose objects are the one dimensional simplicial complexes. Moreover, it is useful to notice that the null graph $G_\emptyset = (\emptyset, \emptyset)$ is an object in $\mathcal{G}$, since graphs in $\mathcal{G}$ are defined as $G = (V, E)$.

Given a graph $G \in \mathcal{G}$ there is a covariant functor, $\mathrm{Cl} : \mathcal{G} \to \mathcal{S}$, called the **clique functor** given by $[v_0, \ldots, v_k] \in \mathrm{Cl}(G)$ if and only if $(v_i, v_j) \in E$ for all $0 \leq i \neq j \leq k$.. This functor is well defined because $(v, v) \mapsto [v]$, for all $v \in V$.

Viceversa there is a functor $k_1 : \mathcal{S} \to \mathcal{G}$ where, given a simplicial complex Σ, $k_1(\Sigma)$ is the (reflexive) graph corresponding to the $1-$skeleton of Σ.

It is important to notice that in general $\Sigma \neq \mathrm{Cl}(k_1(\Sigma))$. For example, if we consider the simplicial complex $\Sigma = \{[a], [b], [c], [a, b], [a, c], [b, c]\}$, the clique complex of its underlying graph is $\mathrm{Cl}(k_1(\Sigma)) = \Sigma \cup \{[a, b, c]\}$.

Following this formalism, we can redefine a flag complex as a simplicial complex Σ for which $\Sigma = \mathrm{Cl}(k_1(\Sigma))$. Flag complexes form a subcategory of $\mathcal{S}$ denoted by $\mathcal{F}$.

Remark 3.1.3.1. It is easy to see that the order complex $\mathcal{O}(X)$ is a flag complex for all $X \in \mathcal{T}_f$. In particular this implies that, for all $\Sigma \in \mathcal{S}$, the barycentric subdivision $\mathcal{O}(\Sigma)$ is a flag complex.

Proposition 3.1.4. *The functors* $\mathrm{Cl} : \mathcal{G} \to \mathcal{F}$ *and* $k_{1|\mathcal{F}} : \mathcal{F} \to \mathcal{G}$ *give an isomorphism* $\mathcal{G} \simeq \mathcal{F}$.

Proof. Obvious. □

Summarizing:

$$\mathcal{S} \xrightarrow{\pi} \mathcal{P} \simeq \mathcal{T}_f^0 \xrightarrow{\mathcal{O}} \mathcal{F} \underset{k_1}{\overset{\mathrm{Cl}}{\cong}} \mathcal{G} \qquad \mathcal{O}\circ\pi : \mathcal{S} \to \mathcal{F} \tag{3.1.4}$$

3.2 P-weighted graphs and P-persistent objects: equivalences and adjunctions

Let $P \in \mathcal{P}$ be a poset and $G = (V, E) \in \mathcal{G}$ a (reflexive) graph, let us denote by $G \in \mathcal{S}$ the corresponding one dimensional simplicial complex. A $P-$*weighted graph* is a pair (G, ω), where $\omega : (G, \subseteq) \to P$ is a morphism of posets, that is a function $G \to P$ continuous in the Alexandrov topology. We define as $\mathcal{G}_P$ the **category of** $P-$**weighted graphs**, having objects $P-$weighted graphs and whose morphisms $\alpha : (G, \omega) \to (H, \theta)$ are induced by a simplicial map $\rho : G \to H$, such that $\alpha(G_v) \subseteq H_v$, where, for any $v \in P$, $G_v = \{x \in G \,|\, \omega(x) \leq v\}$.

Following [19, Section 2.3], we introduce a P**-persistent object** in $\mathcal{A}$ as a functor $\varphi : P \to \mathcal{A}$, where P be a poset and $\mathcal{A}$ an arbitrary category. P-persistent objects in $\mathcal{A}$ with their natural transformations form a category, which we will denote, as usual, by $\mathcal{A}^P$. Given two categories $\mathcal{A}$ and $\mathcal{B}$, to any functor $\phi : \mathcal{A} \to \mathcal{B}$ it corresponds a functor $\mathcal{A}^P \to \mathcal{B}^P$. It is given by $\varphi \in \mathcal{A}^P \mapsto \phi \circ \varphi$. It will be denoted by ϕ^P.

We define two functors that relate to each other the category of weighted graphs $\mathcal{G}_P$ and that of P-persistence weighted graphs $\mathcal{G}^P$.

Proposition 3.2.1. *For all $P \in \mathcal{P}$, there is a functor $\Phi_P : \mathcal{G}_P \to \mathcal{G}^P$.*

Proof. Let $(G, \omega) \in \mathcal{G}_P$. From the definition of G_v we have that $G_u \subseteq G_v$ for every $u \leq v$.
We can associate to $(G, \omega) \in \mathcal{G}_P$ a $P-$persistent object in $\varphi_G \in \mathcal{G}^P$, namely $\varphi_G(v) = G_v$ with the inclusions maps $\varphi_G(u \leq v) : G_u \hookrightarrow G_v$ for all $v \in P$,

$u \in P_v$. It is easy to check that the correspondence $(G, \omega) \to \varphi_G$ is natural in G. Therefore Φ_P is a functor between the two categories. □

Proposition 3.2.2. *For all $P \in \mathcal{P}$ there exists a functor $\Psi_P : \mathcal{G}^P \to \mathcal{G}_P$.*

Proof. Choose $\varphi \in \mathcal{G}^P$, and, for every $v \in P$, set $\varphi_v := \varphi(v) \in \mathcal{G}$.
Let $\omega^\varphi : \coprod_{v \in P} \varphi_v \to P$ be given by $\omega^\varphi|_{\varphi_v} = v$. It is easy to check that the correspondence $\varphi \mapsto (\coprod_{v \in P} \varphi_v, \omega^\varphi)$ is natural in φ, thus giving a functor $\Psi_P : \mathcal{G}^P \to \mathcal{G}_P$. □

3.2.1 Equivalence

Let $\bar{\mathcal{G}}_P$ be the subcategory of $\mathcal{G}_P$ with the same objects, and morphisms the maps $\alpha : (C, \omega) \to (D, \omega')$ such that for every $x \in G$, $\omega'(\alpha(x)) = \omega(x)$. We set $\bar{\Phi}_P$ as the restriction of Φ_P to $\bar{\mathcal{G}}_P$.

Remark 3.2.2.1. It is useful to notice that, actually, $\Psi_P : \mathcal{G}^P \to \bar{\mathcal{G}}_P$. Since $\Psi_P(\varphi) \in Ob(\mathcal{G}_P) = Ob(\bar{\mathcal{G}}_P)$ for all $\varphi \in \mathcal{G}^P$, we just show that $\Psi_P(\mu)$ preserves weights for every $\mu : \varphi \to \tau \in \mathcal{G}^P$. Indeed from the definition of the weights $\omega^\varphi, \omega^\tau$ we have that $(\omega^\varphi)^{-1}(u) = \varphi_u$ then $\Psi_P(\mu)(\varphi_u) \subseteq \tau_u = (\omega^\tau)^{-1}(u)$.

Let $\mathcal{G}_\iota^P$ be the subcategory of $\mathcal{G}^P$ whose objects are $\varphi \in \mathcal{G}^P$ such that the morphisms $\varphi(u \leq v) : \varphi(u) \to \varphi(v)$ are inclusions.
We set Ψ_P^ι as the restriction of Ψ_P to $\mathcal{G}_\iota^P$.

Following Carlsson and Zomorodian [20] we introduce the concept of **one critical** $P-$persistent object. Let $\varphi \in \mathcal{G}_\iota^P$. φ is said to be one-critical if for all $v \in P$, for all $(x, y) \in E_{\varphi(a)}$

$$\exists!\ m_{xy} = \min\{u \in P | \varphi_{uv}(x, y) = (x, y)\} \tag{3.2.1}$$

The one-critical P-persistent objects form a subcategory of $\mathcal{G}_\iota^P$, which will be denoted by $\mathcal{G}_{\mathbf{1}}^P$. We set $\Psi_P^{\mathbf{1}}$ as the restriction of Φ_P to $\mathcal{G}_{\mathbf{1}}^P$.

Remark 3.2.2.2. It is useful to notice that $\Phi_P : \mathcal{G}_P \to \mathcal{G}_{\mathbf{1}}^P$. Indeed consider $(G, \omega) \in \mathcal{G}_P$. By definition of Φ_P, it is clear that $\varphi_G \in \mathcal{G}_{\mathbf{1}}^P$, with $m_{xy} = \omega(x, y)$.

Theorem 3.2.3. *The categories $\mathcal{G}_{\mathbf{1}}^P$ and $\bar{\mathcal{G}}_P$ are equivalent.*

Proof. It is a well known fact in category theory that a functor is an equivalence if and only if it is full, faithful and essentially surjective. To prove the equivalence of category we then need to verify that Φ_P has these three properties.

Consider $(G, \omega), (H, \theta) \in \bar{\mathcal{G}}_P$, and $\alpha \in \hom_{\bar{\mathcal{G}}_P}((G, \omega), (H, \theta))$. The functor Φ_P is essentially surjective if it is surjective on objects up to isomorphism. Let $\varphi \in \mathcal{G}_{\mathbf{1}}^P$, then we can construct $(G, \omega) \in \bar{\mathcal{G}}_P$ by $G := \bigcup_{a \in P} \varphi(a)$ and $\omega((x, y)) = m_{xy}$ (see 3.2.1). It follows that $\Phi_P((G, \omega))$ is such that, for all $u \in P$, one has $\varphi_G(u) = \{x \in G |\ \omega(x) \leq u\} = \bigcup_{a \in P; a \leq u} \varphi(a) \cong \varphi(u)$ by definition of φ.

The functor Φ_P is full if the map

$$\Phi_P((G, \omega), (H, \theta)) : \hom_{\bar{\mathcal{G}}_P}((G, \omega), (H, \theta)) \to \hom_{\mathcal{G}_{\mathbf{1}}^P}(\varphi_G, \varphi_H)$$

is surjective for all $(G, \omega), (H, \theta) \in \bar{\mathcal{G}}_P$.
Consider a morphism $\rho : \varphi_G \to \varphi_H$ in $\mathcal{G}_{\mathbf{1}}^P$. Let $\alpha : E_G \to E_H$ be given by $\alpha((x, y)) = \rho_{\omega(x,y)}(x, y)$, for every $(x, y) \in E_G$. Then α is a morphism $(G, \omega) \to (H, \theta)$ in $\bar{\mathcal{G}}_P$, because from $\rho_{\omega(x,y)} : \varphi_G(\omega(x, y)) \to \varphi_H(\omega(x, y))$ we have that $\theta(\alpha((x, y))) = \omega(x, y)$. It is clear that $\Phi_P(\alpha)$ is ρ by the definitions of Φ_P and α.

As last step, we prove that Φ_P is said faithful, i.e. that the map

$$\Phi_P((G,\omega),(H,\theta)) : \hom_{\bar{\mathcal{G}}_P}((G,\omega),(H,\theta)) \to \hom_{\mathcal{G}_1^P}(\varphi_G,\varphi_H)$$

is injective for all $(G,\omega),(H,\theta) \in \bar{\mathcal{G}}_P$.
Consider $\alpha,\beta \in \hom_{\bar{\mathcal{G}}_P}((G,\omega),(H,\theta))$ such that $\Phi(\alpha)=\Phi(\beta)$. This means that $\Phi(\alpha)_v = \Phi(\beta)_v$ for all $v \in P$, but this implies that $\alpha|_{G_v} = \beta|_{G_v}$ for all $v \in P$, then $\alpha = \beta$. □

3.2.2 Adjunctions

Beside the equivalence in Th.3.2.3, there are also some results on the relationships between the other categories involved.

Theorem 3.2.4. *$\bar{\Phi}_P$ is left adjoint of Ψ_P, that is*

$$\hom_{\bar{\mathcal{G}}_P}((X,\omega),\Psi_P(\varphi)) \cong \hom_{\mathcal{G}^P}(\bar{\Phi}_P((X,\omega)),\varphi)$$

Proof. Let $\pi : (X,\omega) \to (\coprod_P X_u, \omega^{\varphi_X})$ given by $\pi(x) = (x,\omega(x)) \in X_{\omega(x)}$, for every $x \in (X,\omega)$. This map is well defined and is actually a morphism in the category $\bar{\mathcal{G}}_P$ since $\omega^{\varphi_X}(\pi(x)) = \omega^{\varphi_X}(\,(x,\omega(x))\,) = \omega(x)$.
To prove the assumption we will show that, for every $\alpha \in \hom_{\bar{\mathcal{G}}_P}((X,\omega),\Psi_P(\varphi))$, there is a unique morphism in $\mathcal{G}^P$, $\bar{\alpha} : \bar{\Phi}_P((X,\omega)) \to \varphi$ such that the following diagram commutes

$$\begin{array}{ccc} (X,\omega) & \xrightarrow{\pi} & \Psi_P(\bar{\Phi}_P((X,\omega))) \\ {\scriptstyle\alpha}\downarrow & \swarrow_{\Psi_P(\bar{\alpha})} & \\ \Psi_P(\varphi) & & \end{array} \tag{3.2.2}$$

Let $\alpha \in \hom_{\bar{\mathcal{G}}_P}((X,\omega),(\coprod_P \varphi(u),\omega^\varphi))$, then α will be such that $\omega^\varphi(\alpha(x)) = \omega(x)$, for every $x \in (X,\omega)$. By construction of ω^φ, we will have that $\alpha(x) \in \varphi(\omega(x))$, and, with a little abuse of notation, we will write $\alpha : x \mapsto (\alpha(x),\omega(x))$.

Let $\bar{\alpha} : \bar{\Phi}_P((X,\omega)) \to \varphi$ be the morphism in $\mathcal{G}^P$ defined through $\bar{\alpha}|_{X_u} : X_u \to \varphi(u)$ with $\bar{\alpha}|_{X_u}(x) = \varphi_{\omega(x)u}(\alpha(x),\omega(x))$, where $\varphi_{\omega(x)u} = \varphi(\omega(x) \leq u) : \varphi(\omega(x)) \to \varphi(u)$.

We still have to show that diagram 3.2.2 commutes, i.e. $\Psi_P(\bar{\alpha}) \circ \pi = \alpha$. Let x be an element of (X,ω) with weight $\omega(x)$, then $\pi(x) = (x,\omega(x)) \in X_{\omega(x)}$, so $\bar{\alpha}(\pi(x)) = \varphi_{\omega(x)\omega(x)}(\alpha(x),\omega(x)) = (\alpha(x),\omega(x))$. It follows that the diagram commutes and this proves the adjuction.

□

There is another adjunction.

Theorem 3.2.5. *Ψ^ι_P is left adjoint of Φ_P.*

In order to prove this theorem we need some technical lemmata.

Lemma 3.2.6. *There is a natural transformation $\epsilon : \Psi^\iota_P \Phi_P \longrightarrow 1_{\mathcal{G}_P}$.*

Proof. Consider $(G,\omega) \in \mathcal{G}_P$, then

$$\mathcal{G}_P \xrightarrow{\Phi_P} \mathcal{G}^P_\iota \xrightarrow{\Psi^\iota_P} \mathcal{G}_P \tag{3.2.3}$$

$$(G,\omega) \longmapsto \{G_v\} \longmapsto (\coprod_{v\in P} G_v, \omega^{\varphi_G})$$

Define now $\varepsilon_{(G,\omega)}$ as follows:

$$\begin{array}{rcl} \varepsilon_{(G,\omega)} : (\coprod_{v\in P} G_v, \omega^{\varphi_G}) & \longrightarrow & (G,\omega) \\ (x,u) & \mapsto & x \end{array} \tag{3.2.4}$$

This map is well defined since $\omega^{\varphi_G}((x,u)) = u \geq \omega(x) = \omega(\varepsilon_{(G,\omega)}(x,u))$.

Consider now (F,τ), and $\alpha : (G,\omega) \to (F,\tau)$ in $\mathcal{G}_P$, trivially the following diagram commutes:

$$\begin{array}{ccccc} (G,\omega) & \xrightarrow{\Psi^\iota_P\circ\Phi_P} & (\coprod_{v\in P} G_v, \omega^{\varphi_G}) & \xrightarrow{\varepsilon_{(G,\omega)}} & (G,\omega) \\ \downarrow{\scriptstyle \alpha} & & \downarrow{\scriptstyle \Psi^\iota_P(\Phi_P(\alpha))} & & \downarrow{\scriptstyle \alpha} \\ (F,\tau) & \xrightarrow[\Psi^\iota_P\circ\Phi_P]{} & (\coprod_{v\in P} F_v, \omega^{\varphi_F}) & \xrightarrow[\varepsilon_{(F,\tau)}]{} & (F,\tau) \end{array} \tag{3.2.5}$$

ε is the natural transformation we were searching for. □

Lemma 3.2.7. *There is a natural trasformation $\eta : \Psi^\iota_P\Phi_P \longrightarrow 1_{\mathcal{G}^P_\iota}$.*

Proof. Consider $\varphi \in \mathcal{G}^P_\iota$, $\Phi\circ\Psi(\varphi) = (\coprod_{u\leq v}\phi(u), \subseteq)$.

$$\mathcal{G}^P_\iota \xrightarrow{\Psi^\iota_P} \mathcal{G}_P \xrightarrow{\Phi_P} \mathcal{G}^P_\iota \tag{3.2.6}$$

$$\{\varphi(v), \subseteq\}_{v\in P} \longmapsto (\coprod_{v\in P}\varphi(v), \omega^\varphi) \longmapsto \{\coprod_{u\leq v}\varphi(u), \subseteq\}$$

Define now η_φ as follows:

$$\begin{array}{rcl} \eta_\varphi : \varphi(v) & \to & \coprod_{u\leq v}\varphi(u) \\ x & \mapsto & (x,v) \end{array} \tag{3.2.7}$$

Consider now θ, and $\alpha : \varphi \to \theta$ in $\mathcal{G}_\iota^P$, the following diagram commutes:

$$\begin{array}{ccc} \varphi(v) & \xrightarrow{\eta_\varphi} & \coprod_{u \leq v} \varphi(u) \\ \Big\downarrow{\scriptstyle \alpha_v} & & \Big\downarrow{\scriptstyle \coprod \alpha_u} \\ \theta(v) & \xrightarrow{\eta_\theta} & \coprod_{u \leq v} \theta(u) \end{array} \tag{3.2.8}$$

where for every $(x, w) \in \coprod_{u \leq v} \varphi(u)$, $\coprod \alpha_u((x, w)) = (\alpha(x), w)$.
η is the natural transformation we were searching for. □

Proof of Theorem 3.2.5. We prove the unit-counit adjunction, with ε and η the natural transformations defined in Lemma 3.2.6, and 3.2.7.

To prove the adjunction we verify that the following compositions are the identity transformation of the respective categories.

$$\begin{array}{ccccc} \Phi_P & \xrightarrow{\eta\Phi} & \Phi_P\Psi_P^\iota\Phi_P & \xrightarrow{\Phi\varepsilon} & \Phi_P \\ \Big| & & & & \Big\uparrow \\ \Big| & \multicolumn{3}{c}{\xrightarrow{\quad id_{\mathcal{G}_P} \quad}} & \Big\uparrow \end{array} \qquad \begin{array}{ccccc} \Psi_P^\iota & \xrightarrow{\Psi\eta} & \Psi_P^\iota\Phi_P\Psi_P^\iota & \xrightarrow{\varepsilon\Psi} & \Psi_P^\iota \\ \Big| & & & & \Big\uparrow \\ \Big| & \multicolumn{3}{c}{\xrightarrow{\quad id_{\mathcal{G}_\iota^P} \quad}} & \Big\uparrow \end{array} \tag{3.2.9}$$

which means that for each (G, ω) in $\mathcal{G}_P$ and each φ in $\mathcal{G}^P$,

$$1_{\Psi_P^\iota(\varphi)} = \varepsilon_{\Psi_P^\iota(\varphi)} \circ \Psi_P^\iota(\eta_\varphi) \tag{3.2.10}$$

$$1_{\Phi_P((G,\omega))} = \Phi_P(\varepsilon_{(G,\omega)}) \circ \eta_{\Phi_P((G,\omega))} \tag{3.2.11}$$

We start by verifying equation 3.2.10. Let $\varphi \in \mathcal{G}_\iota^P$, we know that $\eta_\varphi : \varphi \longrightarrow \Phi_P \circ \Psi_P^\iota(\varphi)$ is a natural transformation defined for every $v \in P$ by

$$\begin{array}{rccl} \eta_\varphi(v): & \varphi(v) & \longrightarrow & \Phi_P(\Psi_P^\iota(\varphi))(v) = \coprod_{u \leq v} \varphi(u) \\ & x & \mapsto & (x, v) \end{array} \tag{3.2.12}$$

Then

$$\Psi_P^\iota(\eta_\varphi) : \Psi_P^\iota(\varphi) \to (\coprod_{v \in P} \Phi_P(\Psi_P^\iota(\varphi))(v), \omega^{\Phi_P \Psi_P^\iota(\varphi)}) = (\coprod_{v \in P} \coprod_{u \leq v} \varphi(u), \omega^{\Phi_P \Psi_P^\iota(\varphi)}),$$

where $\omega^{\Phi_P \Psi_P^\iota(\varphi)}|_{\coprod_{u \leq v} \varphi(u)} = v$. From the definition of ε we gave in Lemma 3.2.6, we deduce that

$$\varepsilon_{\Psi_P^\iota(\varphi)} : \Psi_P^\iota \circ \Phi_P(\Psi_P^\iota(\varphi)) \longrightarrow \Psi_P^\iota(\varphi).$$

One has that $\Psi_P^\iota \circ \Phi_P(\Psi_P^\iota(\varphi))$ is the weighted graph $(\coprod_{v \in P} \Psi_P^\iota(\varphi)_v \, , \omega^{\Phi_P \Psi_P^\iota(\varphi)})$, where $\Psi_P^\iota(\varphi)_v = \{x \in \Psi_P^\iota(\varphi) | \omega^\varphi(x) \leq v\} = \coprod_{u \leq v} \varphi(u)$, and $\omega^{\Phi_P \Psi_P^\iota(\varphi)}|_{\coprod_{u \leq v} \varphi(u)} = v$.

$$\begin{array}{rccl} \varepsilon_{\Psi_P^\iota(\varphi)}: & (\coprod_{v \in P} \coprod_{u \leq v} \varphi(u), \omega^{\Phi_P \Psi_P^\iota(\varphi)}) & \longrightarrow & \Psi_P^\iota(\varphi) \\ & ((x, u), v) & \mapsto & (x, u) \end{array} \tag{3.2.13}$$

where $\Psi_P^\iota(\varphi) = (\coprod_{v \in P} \varphi(v), \omega^\varphi)$, with $\omega^\varphi|_{\varphi(v)} = v$.

Then $\varepsilon_{\Psi_P^\iota(\varphi)} \circ \Psi_P^\iota(\eta_\varphi) = 1_{\Psi_P^\iota(\varphi)}$ as the following shows:

$$\begin{array}{ccccc} (\coprod_{v \in P} \varphi(v), \omega^\varphi) & \xrightarrow{\Psi(\eta_\varphi)} & (\coprod_{v \in P} \coprod_{u \leq v} \varphi(u), \omega^{\Phi_P \Psi_P^\iota(\varphi)}) & \xrightarrow{\varepsilon_{\Psi_P^\iota(\varphi)}} & (\coprod_{v \in P} \varphi(v), \omega^\varphi) \\ (x, v) & \mapsto & ((x, v), v) & \mapsto & (x, v) \end{array} \tag{3.2.14}$$

We verify now identity 3.2.11. Consider $(G,\omega) \in \mathcal{G}_P$, we have that

$$\eta_{\Phi_P((G,\omega))} : \Phi_P((G,\omega)) \to \Phi_P \circ \Psi_P^{\iota}(\Phi((G,\omega)))$$

where $\Phi_P((G,\omega))(v) = G_v$, with $G_v = \{x \in G | \omega(x) \leq v\}$. For every $v \in P$, we find that $\eta_{\Phi_P((G,\omega))}$ is determined by:

$$\begin{array}{rccc} \eta_{\Phi_P((G,\omega))}(v): & G_v & \longrightarrow & \coprod_{u \leq v} G_u \\ & x & \mapsto & (x,v) \end{array} \tag{3.2.15}$$

Considering that $\varepsilon_{(G,\omega)} : (\coprod_{v \in P} G_v, \omega^{\varphi_G}) \mapsto (G,\omega)$, where $\omega^{\varphi_G}|_{G_v} = v$. We have that $\Phi_P(\varepsilon_{(G,\omega)})$ is defined for every $v \in P$:

$$\Phi_P(\varepsilon_{(G,\omega)})(v): \quad \Phi_P(\Psi_P^{\iota} \circ \Phi_P((G,\omega)))(v) \quad \longrightarrow \quad \Phi_P((G,\omega))(v) \tag{3.2.16}$$

where

$$\Phi_P(\Psi_P^{\iota} \circ \Phi_P((G,\omega)))(v) = \{(x,u) \in \coprod_{v \in P} G_v \text{ s.t. } \omega^{\varphi_G}((x,u)) = u \leq v\} = \coprod_{u \leq v} G_u$$

and $\Phi_P((G,\omega))(v) = G_v$. This gives the following natural transformation:

$$\begin{array}{ccccc} G_v & \xrightarrow{\eta_{\Phi_P((G,\omega))}(v)} & \coprod_{u \leq v} G_u & \xrightarrow{\Phi_P(\varepsilon_{(G,\omega)})(v)} & G_v \\ x & \mapsto & (x,v) & \mapsto & x \end{array} \tag{3.2.17}$$

which proves that $\Phi_P(\varepsilon_{(G,\omega)}) \circ \eta_{\Phi_P((G,\omega))} = 1_{\Phi_P((G,\omega))}$. □

3.3 Application to homology: multi-persistent homology

Algebraic topology constructs appropriate algebraic objects to apply on topological spaces in order to discern their properties. Homology theory does so by introducing functors from the category of topological spaces (or some related category) and continuous maps to the category of modules over a commutative base ring, such that these modules are topological invariants.

In this section we will show how homology is affected by the results we found in the previous section. To do so, we will first introduce homology over simplicial complexes, which are our main setting, and we will then proceed to define it over general topological spaces.

Simplicial homology

Fixed a field k, in the following, by vector space we intend a $k-$vector space. Given a simplicial complex Σ of dimension d, for $0 \leq n \leq d$ consider the vector spaces $\mathrm{C}_n := \mathrm{C}_n(\Sigma)$ with basis the set of n-faces in Σ. Elements in C_n are called **n-chains**.

The linear maps sending a n-face to the alternate sum of it's $(n-1)$-faces are called **boundaries** and share the property $\partial_{n-1} \circ \partial_n = 0$.

$$\begin{array}{rcl} \partial_n : \mathrm{C}_n & \longrightarrow & \mathrm{C}_{n-1} \\ [p_0, \ldots, p_n] & \to & \displaystyle\sum_{i=0}^{n} (-1)^i [p_0, \ldots, p_{i-1}, p_{i+1}, \ldots, p_n]. \end{array}$$

The subspace $\ker \partial_n$ of C_n is called the vector space of n-**cycles** and denoted by $\mathrm{Z}_n := \mathrm{Z}_n(\Sigma)$. The subspace $\mathrm{Im}\, \partial_{n+1}$ of C_n, is called the vector space of n-**boundaries** and denoted by $\mathrm{B}_n := \mathrm{B}_n(\Sigma)$.

Remark 3.3.0.1. From $\partial_{n-1} \circ \partial_n = 0$ it follows that $\mathrm{B}_n \subseteq \mathrm{Z}_n$ for all n.

The $n-$th simplicial homology space of Σ, with coefficients in k, is the vector space $\mathrm{H}_n := \mathrm{H}_n(\Sigma) := \mathrm{Z}_n/\mathrm{B}_n$. We denote by $\beta_n := \beta_n(\Sigma)$ the rank of H_n : it is usually called the n-th Betti number of Σ.

The first Betti numbers of Σ have an easy intuitive meaning: the 0-th Betti number is the number of connected components of Σ, the first Betti number is the number of two dimensional (poligonal) holes, the third Betti number is the number of three dimensional holes (convex polyhedron).

Remark 3.3.0.2. It easy to check that $\mathrm{C}_n, \mathrm{Z}_n, \mathrm{B}_n$ and, therefore, H_n are all functors $\mathcal{S} \to \mathrm{Vect}_k$, where Vect_k denotes the category of vector spaces and linear mappings.

There is plenty of literature on homology and in particular on simplicial homology, we refer the interested reader to [70]. In particular, one can easily prove the following proposition.

Proposition 3.3.1. *The functors* H_i *are invariants by homeomorphism and homotopy type.*

Let $G \in \mathcal{G}$ be a graph. We now define as the homology space of G,

$$H_i(G) := H_i(\mathrm{Cl}(G))$$

.

Proposition 3.3.2. *Let Σ be a simplicial complex. Then, there exists a graph $G \in \mathcal{G}$ such that $H_i(\Sigma) = H_i(G)$.*

Proof. The proof is a consequence of Remark 3.1.3 and 3.1.3.1. It is sufficient to consider as $G = k_1(\mathcal{O}(\pi(\Sigma)))$, the 1-skeleton of the barycentric subdivision of Σ, which is a flag complex. □

Singular homology

Simplicial homology has an analogous for general topological spaces, namely **singular homology**, whose definition and properties we briefly recall now. Although we confine ourselves into the category of finite topological spaces, the following definition remains valid for arbitrary topological spaces. We address the interested reader to [Hatcher, 70] for a thorough treatise on these topics.

Let $X \in \mathcal{T}_f$ be a topological space, the chain spaces C_n are in this case replaced by the vector spaces C_n^S freely generated by the set of all continuous functions from the geometric realization of the standard n-simplex Δ^n to X. (C_n^S, ∂_n^S) is a chain complex whose boundaries are defined in the following way. Let σ be a generator of C_n, i.e. a continuous function from $\Delta^n \to X$. Then

the **boundary homomorphism** ∂_n^S can be constructed in the following way:

$$\partial_n^S(\sigma) = \sum_i^n \sigma|_{[v_0,\ldots,v_{i-1},v_{i+1},\ldots,v_n]}$$

where $\sigma|_{[v_0,\ldots,v_{i-1},v_{i+1},\ldots,v_n]}$ is the restriction of σ to $[v_0,\ldots,v_{i-1},v_{i+1},\ldots,v_n]$. It is easy to verify that $\partial_n^S \circ \partial_{n+1}^S = 0$, thus we can define the homology spaces as we did for simplicial homology. We will denote the i^{th} **singular homology space** by $\mathrm{H}_i^{\mathrm{S}}(X)$. For general nonsense it is easy to check that H_i^S gives a functor $\mathcal{T}_f \to \mathrm{Vect}_k$.

Theorem 3.3.3 ([Hatcher, Theorem 2.27])**.** *For any simplicial complex Σ, the singular homology groups are isomorphic to the simplicial homology groups.*

$$\forall i \in \mathbb{N} \qquad \mathrm{H}_i^{\mathrm{S}}(\Sigma) \cong \mathrm{H}_i(\Sigma)$$

Let $X, Y \in \mathcal{T}_f$, and let $\pi_n(X,x)$ denote the homotopy group of the space X at base point $x \in X$.
A map $f : X \to Y$ is a **weak homotopy equivalence** if the following conditions are verified:

1. f induces an isomorphism of the connected components of X and Y

$$\Pi_0(f) : \Pi_0(X) \to \Pi_0(Y)$$

2. for all $x \in X$, and $n \geq 1$ is an isomorphism on the homotopy groups

$$\pi_n(f) : \pi_n(X,x) \to \pi_n(Y, f(x))$$

There is the following result.

Theorem 3.3.4 (McCord, [66]). *Let $X \in \mathcal{T}_f$ with $X/\sim$ its Kolmogorov quotient, then $\mathcal{O}(X/\sim) \in \mathcal{F}$ is weak homotopy equivalent to X.*

We refer the interested reader to [4, Ch. 1.4]. In view of this result it makes sense to set $\mathcal{O}(X) := \mathcal{O}(X/\sim)$ for all $X \in \mathcal{T}_f$.

We bring together these new definitions on homology, with the one introduced in 3.1 and summarized in diagram 3.1.4. From theorem 3.3.3 and 3.3.4 we deduce that $\mathrm{H}_i^{\mathrm{S}}(X) \cong \mathrm{H}_i(\mathcal{O}(X))$ for all $X \in \mathcal{T}_f$. Moreover, since $\mathcal{O}(X)$ is a flag complex $\mathrm{H}_i(\mathcal{O}(X)) = \mathrm{H}_i(\mathrm{Cl}(k_1(\mathcal{O}(X))))$, that is the graph homology of the graph which is the 1-skeleton of $\mathcal{O}(X)$. Thus we can restrict ourselves to the study of the graph homology of $k_1(\mathcal{O}(X))$. We can then sum up these information in the folliwng commutative diagram:

$$\begin{array}{ccccc} \mathcal{T}_f & \longrightarrow & \mathcal{T}_f^0 & \xrightarrow{\mathrm{H}_i^{\mathrm{S}}} & \mathrm{Vect}_k \\ & & \mathcal{O}\downarrow & \nearrow_{\mathrm{H}_i} & \uparrow{\scriptstyle \mathrm{H}_i} \\ & & \mathcal{F} & \underset{\mathrm{Cl}}{\overset{k_1}{\rightleftarrows}} & \mathcal{G} \end{array} \tag{3.3.1}$$

The main objective in topological data analysis is to compute the singular homology of the finite topological space underlying our data. If the data that is available to us is a weighted graph, if we suppose it to be the graph underlying the order complex of the unknown space, from the previous diagram we can deduce that computing the simplicial homology of its clique complex is equivalent to computing the singular homology of the underlying space.

P-persistence homology

Although these observations are interesting per se, they become much more significant if we consider not only the homological structure of a data space but also its P-persistent properties.

Let $\tau \in \mathcal{T}_f^P$, the composition $\mathrm{H}_i^{\mathrm{S}} \circ \tau \in \mathrm{Vect}_k^P$ will be called the i^{th} P-persistent homology of $\tau \in \mathcal{T}_f^P$.

Taking in consideration the concepts defined in the previous section, we have the following result which states that for any P-persistent finite topological space, there is a one-critical P-persistent graph having the same P-persistent homology.

Proposition 3.3.5. *Let $\tau \in \mathcal{T}_f^P$, then there is $\theta \in \mathcal{G}_{\mathbf{1}}^P$ such that*

$$\mathrm{H}_i^S \circ \tau \cong \mathrm{H}_i \circ \mathrm{Cl} \circ \theta \tag{3.3.2}$$

as functors.

Proof. The commutativity of diagram 3.3.1 implies that the following diagram is commutative:

$$\begin{array}{ccccc} \mathcal{T}_f^P & \longrightarrow & (\mathcal{T}_f^0)^P & \xrightarrow{(\mathrm{H}_i^{\mathrm{S}})^P} & \mathrm{Vect}_k^P \\ & & \downarrow{\scriptstyle (k_1 \circ \mathcal{O})^P} & \nearrow{\scriptstyle (\mathrm{H}_i \circ \mathrm{Cl})^P} & \\ & & \mathcal{G}^P & & \end{array} \tag{3.3.3}$$

Therefore the statement holds with $\theta = k_1 \circ \mathcal{O} \circ \tau$. □

The above result implies that P-persistent singular homology of finite spaces can be computed as P-persistent homology of graphs, and translates into the following existence theorem.

Theorem 3.3.6. *Let $\tau \in \mathcal{T}_f^P$ be a P-filtration of topological spaces such that $\tau_{ab} : X_a \to X_b$ is injective for all $a, b \in P$ with $a \leq b$,*
Then exists a weighted graph $(G, \omega) \in \bar{\mathcal{G}}_P$ such that $\mathrm{H}_i^S \circ \tau \cong \mathrm{H}_i(\mathrm{Cl}(\Phi_P(G, \omega)))$.

Proof. It follows from Prop.3.3.5 and Th.3.2.3. □

3.3.1 Considerations on topological strata

In [86] the authors adopted different techniques to build filtrations of simplicial complexes from weighted networks. We are going to focus on two of these that are qualitatively different: a metrical, and a non-metrical filtration. The metrical filtration was obtained constructing a sequence of Vietoris-Rips complexes by studying the change in the overlap of ϵ-neighbourhoods of vertices while varying their radius ϵ, considering as metric of the underlying space the inverse-weighted shortest path. The non-metrical one relied instead on associating clique complexes to a series of binary networks obtained from a progressively less restrictive thresholding on the edge weights. The comparison highlighted a clear difference between the diagrams of the two filtrations: in the metric case, most generators had short persistence and were thus distributed along the diagonal; in the non-metric, generators displayed a range of persistences, including some very large ones, and thus poitned to the presence of interesting heterogeneities in the network structure which were not noticeable via the metric filtration.

This result can be reinterpreted in the light of Theorem 3.3.6. To do so, we need to define the two filtrations in terms of category theory.

Let $\mathbb{R}$ denote, as usual, the set of real numbers, which in this setting is considered as a (totally) ordered set via its usual ordering. Let us consider a (finite) metric space $X = (X, d)$ where $d : X \times X \to \mathbb{R}$. There is a weighted graph (G_X, ω_d) associated to X, namely its distance graph, which is the complete graph on the (vertex) set X, whose edges are weighted by the distance between their extrema. The Vietoris-Rips filtration associated to X is then $\mathrm{Cl}(\Phi_{\mathbb{R}}(G_X, \omega_d))$. It should be clear that $\Phi_{\mathbb{R}}(G_X, \omega_d) \in \mathcal{G}^{\mathbb{R}}_{inc}$.

Let us consider now a weighted graph (G, w) having weights in $\mathbb{R}$. We can bestow the set of vertices V of G with various distances induced by the weights and graph structure. One classic example is the weighted shortest path metric, that is $\tilde{d}(x, y) = \min\{\sum_{(u,v)\in p} w(u, v)\}_p$ where p is a path between the vertices x and y. As we have just explained above, associated to $(V, \tilde{d})$ there is then an element of $\mathcal{G}^{\mathbb{R}}_{inc}$, namely $(G_V, \omega_{\tilde{d}})$. After this step one can compose with the clique functor and with the homology functor to obtain persistent homology. The metric filtration introduced in [86] was computed in this way and can be written in categorical form as $\mathrm{Cl}(\Phi_{\mathbb{R}}(G_V, \omega_{\tilde{d}}))$, while the non-metric filtration is $\mathrm{Cl}(\Phi_{\mathbb{R}}(G, w))$. Theorem 3.3.6 tells us that, while it is possible to reconstruct the original data structure from the non-metrical filtration, this is not possible when we obfuscate the data by adopting a metrical lens, confirming the empirical results found in [85] and the validity of that approach, which also finds further empirical support in [86] and [84].

3.3.2 Conclusions

In this chapter we proved the categorical equivalence between a subcategory of weighted graphs and that of corresponding P-persistent objects. Moreover, we showed how these results influence topological data analysis in the construction of suitable filtrations of simplicial complexes, which are able to encode the correct persistent homology. Finally we also showed how these results give a formal window into why it can be unwise to use metric tools to construct simplicial complexes when datasets are not necessary sampled from a metric space.

In the next chapter we are going to look into the physical limitations of computing persistent homology. Expanding on the work of Lloyd et al. [63], we will provide a solution using quantum computation.

Chapter 4

Quantum algorithm for persistent homology

In the previous chapters we have seen how topological methods for the analysis of data require the construction and storage of a simplicial complex when computing topological features. The application of these techniques to large simplicial complexes is still limited since the most efficient classical algorithms for estimating topological invariants for persistent homology such as Betti numbers scale as $O(m^\epsilon) \sim O(m^3)$, where m is the number of simplices in the simplicial complex [6–9, 25, 26, 40, 95, 68, 10].

The difficulty of the implementation comes from the fact that most construction methods build a filtration of simplicial complexes where every element is nested into an n-simplex, where n is the number of data points in the data set. In this case the number of simplices in the complex is $m = 2^n$, which implies that the algorithm for computing Betti numbers scales exponentially in the number of points of the data set.

Recently S. Lloyd, S. Garnone, and P. Zanardi solved this problem introducing a quantum machine learning algorithm for performing topological analysis (QTA) of large datasets [63], which leverages the improved parallelism of quantum computation to provide an exponential speedup over the corresponding classical algorithms. The QTA algorithm encodes the simplicial complex and its elements into quantum mechanical states, and later it identifies the topological invariants by performing linear operation on those states. The algorithm then yields at each step ϵ of the filtration an estimate of the Betti number for all orders, to accuracy δ in time $O(n^5/\delta)$. Even though the quantum framework introduced in the paper is constructed to include the entire filtration of simplicial complexes, the algorithm proposed in [63] limits itself to the computation of homological features at each step of the filtration.

In this chapter we improve on QTA algorithm providing new insight which takes in consideration the evolving topology of the simplicial complex. To achieve this, we study in depth the homology maps induced by the filtration of simplicial complexes in order to track the progression of the Betti numbers, together with their relation to the combinatorial laplacian. This new mathematical insight that enables tracking the evolution of topological features along a filtration of simplicial complexes.

Our method, though not very practical in a classical framework, can be used to yield a more informative topological invariant, which takes into account the effects of the simplicial complex growth through the filtration.

The algorithms given here are related to quantum matrix inversion algorithms [51, 64, 1]. The original matrix inversion algorithm yielded as solution a quantum state, and left open the question of how to extract useful information

from that state [51]. The algorithms yield as output not quantum states but rather topological invariants and do so in time exponentially faster than the best existing classical algorithms. In section (sec. 4.1) we will present the concept of persistent homology, and provide the necessary mathematical background for the quantum algorithm. We will then give a way to encode the information about simplicial complexes in a quantum state (sec. 4.2), and introduce the algorithm (sec. 4.3).

4.1 Persistent Homology

Let us consider a filtration of simplicial complexes, that is a family of simplicial complexes $\{X^\epsilon\}_\epsilon$ such that $X^0 = \emptyset \subset X^1 \subset \cdots \subset X^m = X$. A filtration of simplicial complexes induces injective morphism between the corresponding chain complexes by inclusion of the canonical bases:

$$0 \cdots \hookrightarrow C_\bullet^\epsilon(X^\epsilon, F) \hookrightarrow C_\bullet^{\epsilon+1}(X^{\epsilon+1}, F) \hookrightarrow \dots C_\bullet^m = C_\bullet(X, F) \tag{4.1.1}$$

where every $f : C_\bullet^{\epsilon-1}(X^{\epsilon-1}, F) \hookrightarrow C_\bullet^\epsilon(X^\epsilon, F)$ is a family of maps $\{f_k^\epsilon : C_k^\epsilon(X^\epsilon, F) \to C_k^{\epsilon+1}(X^{\epsilon+1}, F)\}_k$ that commute with the boundary maps of the chain complexes introduced in 3.3, that is, $f_{k-1}^\epsilon \circ \partial_k^\epsilon = \partial_k^{\epsilon+1} \circ f_k^\epsilon$. The boundary maps satisfy $\partial_{k-1}^* \circ \partial_k^* = 0$ for all $k \in \mathbb{Z}$, this implies that $\mathrm{B}_k^* = Im(\partial_k^*(C_{k+1}^*) \subseteq \mathrm{Z}_k^* = Ker(\partial_k^*(C_k^*)$, which justifies the following definition.

Definition 4.1.1. The kth-**persistent homology** group of X for the interval (b, d), for $b < d$ in the filtration, is given by:

$$\mathrm{H}_k^{b,d}(X, F) = \frac{\mathrm{Z}_k^b(X)}{\mathrm{B}_k^d(X) \cap \mathrm{Z}_k^b(X)}$$

The dimension of the persistent homology group is called persistence Betti number $\beta_k^{b,d} = \dim(\mathrm{H}_k^{b,d}(X, F))$. The kth Betti number can be interpreted as the number of k-dimensional cycles that are in the simplicial complex at scale b of the filtration and that become boundaries in the simplicial complex at step $d > b$. The Betti number is a topological invariant that is able to characterize the topology of the simplicial complex in a very clear and intuitive way. This is why it has become a very important tool in Topological Data Analysis [19].

In chapter 1 we introduced some techniques to contruct simplicial complexes from data sets. In particular some of these methods, like the Vietoris-Rips complex, are dependent on the choice of one parameter. In the case of the Vietoris-Rips complex, the parameter is the radius ϵ of the balls. As the radius increases, the method creates nested simplicial complexes as can be seen in fig. 4.1. For a detailed account on persistent homology and its application, we refer the interest reader to [37].

4.1.1 Expliciting homology maps

We will now study how to better describe algebraically the evolution of persistent homology of a simplicial complex through different scales of the filtration.

Let $C(X, F)$ be a chain complex of the simplicial complex $X = (V, \Sigma)$. Then the k**th combinatorial laplacian** $\mathcal{L}_k : C_k \longrightarrow C_k$ is defined as follows:

$$\mathcal{L}_k = \partial_k^* \partial_k + \partial_{k+1} \partial_{k+1}^* \tag{4.1.2}$$

where $\partial_k : C_k \to C_{k-1}$ is the boundary map $\partial_k^* : C_{k-1} \to C_k$ is the coboundary map.

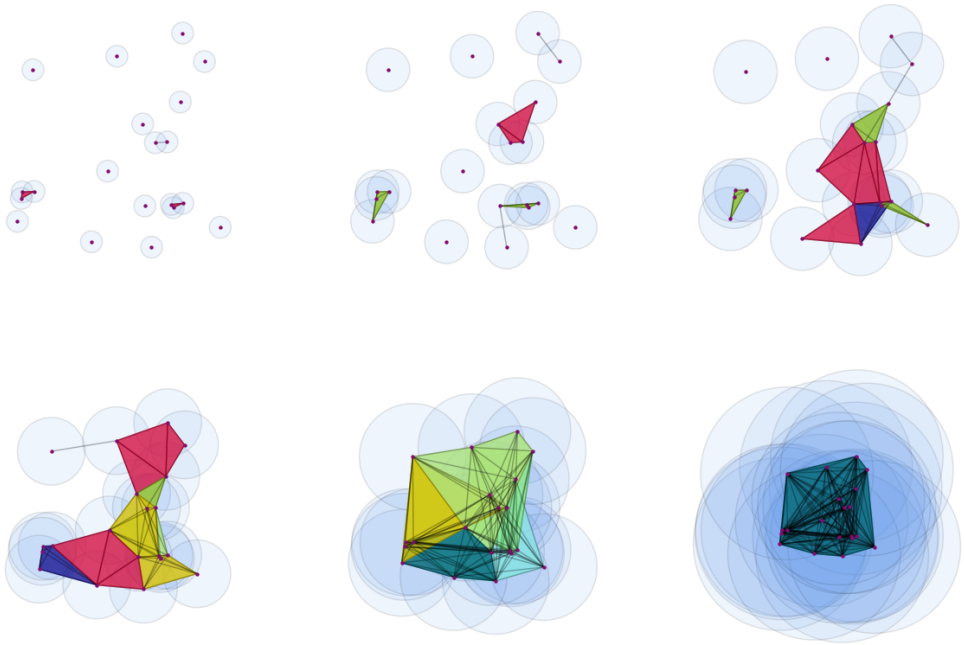

Fig. 4.1 Vietoris Rips complex at different scales. Each facet of the complex is colored according to their dimension. It is easy to see how, as the balls' radius increases, the Vietoris-Rips construction method creates a sequence of nested simplicial complexes.

The graph Laplacian was generalized to simplicial complexes by Eckmann [36], who formulated and proved the discrete version of the Hodge theorem. We will now enunciate this and other well known results, which introduce a connection between the combinatorial Laplacian and the study of homology.

Proposition 4.1.2 (Hodge [53]). *Let $C(X, F)$ be given then C_k, for all $k \in \mathbb{Z}$ decomposes as*

$$C_k = \ker(\mathcal{L}_k) \oplus \mathrm{Im}(\partial_{k+1}) \oplus \mathrm{Im}(\partial_k^*).$$

Moreover $Z_k = \ker(\mathcal{L}_k) \oplus \mathrm{Im}(\partial_{k+1})$.

Theorem 4.1.3 (Eckmann [36]). *NLet $C(X, F)$ be a chain complex. Then* $\ker(\mathcal{L}_k)$ *and* $H_k(X)$ *are isomorphic.*

Therefore, as a consequence of Theorem 4.1.3, when computing the kth-homology group we can restrict ourselves to compute the kernel of the combi-

natorial laplacian of order k, as it wwas done in the QTA algorithm proposed in [63]. To use this method also for persistent homology groups one must explicitly give the morphism induced by the inclusion maps of the filtration $F : C_k(X^\epsilon) \hookrightarrow C_k(X^{\epsilon+1})$.

It is useful to notice that there exists a quotient map (surjective and open) $Q_k^\epsilon : Z_k^\epsilon \to Z_k^\epsilon / B_k^\epsilon = H_k^\epsilon$. Moreover from proposition 4.1.2 we know that $\ker(\mathcal{L}_k^\epsilon) \subseteq Z_k^\epsilon$. Therefore we can redefine our problem as wanting to find an explicit description of $\Phi : \ker(\mathcal{L}_k^\epsilon) \to \ker(\mathcal{L}_k^{\epsilon+1})$ such that the following diagram commutes.

$$\begin{array}{ccc} Z_k^\epsilon & \xhookrightarrow{F|_{Z_k^\epsilon}} & Z_k^{\epsilon+1} \\ \uparrow \downarrow & & \downarrow \\ \ker(\mathcal{L}_k^\epsilon) & \xrightarrow{\Phi} & \ker(\mathcal{L}_k^{\epsilon+1}) \end{array} \tag{4.1.3}$$

Let us consider now $\gamma \in \ker(\mathcal{L}_k^\epsilon)$. Since $\ker(\mathcal{L}_k^\epsilon) \subset Z_k^\epsilon$ then $F(\gamma) \in Z_k^{\epsilon+1}$. Therefore $F(\gamma)$ can be decomposed as $\gamma' + b$ with $\gamma' \in \ker(\mathcal{L}_k^{\epsilon+1})$, $b \in \mathrm{Im}(\partial_{k+1}^{\epsilon+1})$. Then $F(\gamma)$ will belong to the equivalence class $\overline{\gamma'} \in H_k^{\epsilon+1}$. Finally we have that:

$$\Phi : \ker(\mathcal{L}_k^\epsilon) \xhookrightarrow{F|_{Z_k^\epsilon}} Z_k^{\epsilon+1} = \ker(\mathcal{L}_k^{\epsilon+1}) \oplus \mathrm{Im}(\partial_{k+1}^{\epsilon+1}) \xrightarrow{P} \ker(\mathcal{L}_k^{\epsilon+1}) \tag{4.1.4}$$

where P is the projection of $Z_k^{\epsilon+1}$ onto the kernel of the laplacian $\ker(\mathcal{L}_k^{\epsilon+1})$. The morphism Φ is not surjective nor injective anymore, since an element in $\ker(\mathcal{L}_k^\epsilon)$ can become a boundary later in the filtration, and new elements added to the simplicial complex can create cycles that were not present before, thus increasing the dimension of the kernel of the Laplacian.

The same process can be extended for steps ϵ, ℓ in the filtration with $\epsilon < \ell$. In

this case Φ will be:

$$\Phi : \ker(\mathcal{L}_k^{\epsilon}) \xhookrightarrow{F|_{Z_k^{\epsilon}}} Z_k^{\epsilon+1} \xhookrightarrow{F|_{Z_k^{\epsilon+1}}} \cdots \xhookrightarrow{F|_{Z_k^{\ell-1}}} Z_k^{\ell} \xrightarrow{P} \ker(\mathcal{L}_k^{\ell}) \tag{4.1.5}$$

In the following sections we will illustrate how to use this new result to further develop the QTA algorithm.

4.2 Quantum construction of a simplicial complex

4.2.1 Quantum notation

Before describing how to construct the appropriate quantum states needed for our aims, we will give some basic notations and principles which are well used in quantum mechanics and quantum computation. Quantum states are defined in an Hilbert space, that is, a finite-dimensional complex vector space $\mathbb{C}^n$ with an inner product $\langle ., . \rangle$. A vector in $\mathbb{C}^n$ is called a **ket** vector and is denoted as:

$$|x\rangle = \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix} \qquad \text{for } x_i \in \mathbb{C} \tag{4.2.1}$$

while a vector in the dual space $\mathbb{C}^{n*}$ is called a **bra** vector $\langle \alpha| = (a_1, \ldots, a_n)$, where $a_i \in \mathbb{C}$. The inner product of $|x\rangle$ and $\langle \alpha|$ is $\langle \alpha|x\rangle = \sum_{i=1}^{n} a_i x_i$. This inner product naturally introduces a correspondence between $|x\rangle = (x_1, \ldots, x_n)^T$ and $\langle x| = (x_1^*, \ldots, x_n^*)$, then the norm is defined as $|| \, |x\rangle \, || = \sqrt{\langle x|x\rangle}$. The tensor

product of two vectors $|x\rangle$ and $|y\rangle$ is given by:

$$|x\rangle \otimes |y\rangle = (x_1y_1, ..., x_1y_q, x_2y_1, ..., x_2y_q, ..., x_py_1, ...x_py_q)^T \tag{4.2.2}$$

The tensor product $|x\rangle \otimes |y\rangle$ is often abbreviated as $|x\rangle\,|y\rangle$. If two states $|\psi_1\rangle$ and $|\psi_2\rangle$ are physical states of the system, their linear superposition $c_1\,|\psi_1\rangle + c_2\,|\psi_1\rangle$ is also a possible state of the same system (superposition principle), with $c_k \in \mathbb{C}$ and $\sum_{i=1}^{2} |c_k|^2 = 1$.

We cannot say definitely in which state a quantum system is in, in other words the system might be in the state $|\psi_i\rangle$ with a probability p_i. Such a system is said to be in a **mixed state**, while a system whose vector is uniquely specified is in a **pure state**, i.e. $p_i = 1$ for some i and $p_j = 0$ for $j \neq i$.

Let us introduce the density matrix by $\rho = \sum_{i=1}^{N} p_i\,|\psi_i\rangle\,\langle\psi_i|$. A density matrix is a positive-semidefinite Hermitian operator with $\mathrm{tr}(\rho) = 1$ because $\sum \rho_i = 1$. Then a pure state $|\psi\rangle$ is a special case in which the corresponding density matrix is the projection operator onto the state, $\rho = |\psi_i\rangle\,\langle\psi_i|$. At the other extreme a maximally mixed state is the quantum state representing a totally **uniform mixture** of states in the quantum system.

For the reader interested to an introduction to quantum computing from a linear algebra point of view we recommend the overview by Nakahara and Ohmi [72], Nakahara [71].

4.2.2 Simplex quantum state

Big quantum data analysis works by mapping each data point $\overrightarrow{v}$ (a d-dimensional vector over complex numbers) to a quantum state $|v\rangle \in \mathbb{C}^d$, and the entire

dataset to a quantum state $\frac{1}{\sqrt{n}}\sum_j |j\rangle |v_j\rangle \in \mathbb{C}^n \times \mathbb{C}^d$ where n is the number of elements in the dataset. When mapping simplicial complexes we have as input a set of points V of cardinality n, with $\binom{n}{2}$ distances or weights between the points, that is $\{w_{ij} = w_{ji} \in F | i, j \in V\}$ where F is a totally ordered set.

Following what has been done by Lloyd et al. [63], we introduce now the quantum framework that encodes the filtration of simplicial complexes $X = \{X^\epsilon\}_\epsilon$, where X^ϵ is the Vietoris-Rips complex constructed from the data with parameter $\epsilon \in F$.

Each simplex $\sigma_k = [v_0, \ldots, v_k] \in X$ is encoded as a quantum state over n-qubit $|\sigma_k\rangle := |010 \ldots 001\rangle \in \mathbb{C}^{2^n}$ the 1s in $|\sigma_k\rangle$ are at the positions corresponding to the vertices $v_i \in \sigma_k$.

Denote by W_k the $\binom{n}{k+1}$-dimensional Hilbert space corresponding to all possible k-simplices in a simplicial complex with n vertices. Let C_k^ϵ be the subspace of $\mathbb{C}^{2^n}$ spanned by $|\sigma_k\rangle$ where $\sigma_k \in X_k^\epsilon$, the set of k-simplices in X^ϵ. The full **k-simplex space** at scale ϵ is defined to be $C^\epsilon = \oplus_k C_k^\epsilon$. In order to construct the simplicial complexes X^ϵ at each scale, we need to evaluate the distances between points which correspond to the application of a projector $P_k^\epsilon := \frac{1}{\sqrt{|S_k^\epsilon|}} \sum_{\sigma_k \in S_k^\epsilon} |\sigma_k\rangle \langle\sigma_k|$ onto C_k^ϵ, where S_k^ϵ is the set of k-dimensional simplices at scale ϵ.

The **k-simplex state** at scale ϵ, $|\psi\rangle_k^\epsilon$ can be constructed using Grover's algorithm (a quantum search algorithm) [50], that is is a quantum algorithm that finds with high probability the unique input to a black box function that produces a particular output value, using just $O(\sqrt{N})$ evaluations of the function, where N is the size of the function's domain. In our case the function used to implement Grover's algorithm is the membership function $f_k^\epsilon(|\sigma_k\rangle) = 1$

if $\sigma_k \in S_k^\epsilon$. The multi-solution version of Grover's algorithm then allows us to construct the uniform superposition of the quantum states corresponding to the k-simplices at step ϵ of the filtration.

$$|\psi\rangle_k^\epsilon := \frac{1}{\sqrt{|S_k^\epsilon|}} \sum_{\sigma_k \in S_k^\epsilon} |\sigma_k\rangle \tag{4.2.3}$$

The construction of the k-simplex state via Grover's algorithm takes time $O\left(\frac{n^2}{\sqrt{(\varsigma_k^\epsilon)}}\right)$, where $\varsigma_k^\epsilon = \frac{|S_k^\epsilon|}{\binom{n}{k+1}} = \frac{\dim C_k^\epsilon}{\dim W_k}$, that is the fraction of simplices that are actually in the complex at scale ϵ. When this fraction is too small the procedure will fail to find the simplices, if only an exponentially small set of possible k-simplices actually lie in the simplicial complex, then the quantum search will fail to find them. Therefore following [63], we fix an accuracy parameter ς so that at each scale ϵ the algorithm will find k-simplices when $\varsigma_k^\epsilon > \varsigma$, and estimate the number of k-simplices to accuracy $\varsigma_k^\epsilon \pm \varsigma$. Then, as ϵ increases, more simplices will be in the simplicial complex, making the quantum search more likely to succeed at different dimensions k. For ϵ larger than the maximum distance between vectors, all possible simplices will be in the complex (see fig. 4.1).

In a quantum computation there is no way to deterministically put bits in a specific prescribed state unless one is given access to bits whose original state is known in advance. Such bits which are known in advance to be in the state are called **ancilla** bits. Then through opportune ancillas we are able to introduce the state $\rho_k^\epsilon = \frac{1}{|S_k^\epsilon|} \sum_{\sigma_k \in S_k^\epsilon} |\sigma_k\rangle \langle\sigma_k|$, which is the state of the uniform mixture of all k-simplices states in the complex at grouping scale ϵ (see method section in [63] for a detailed construction of this state).

4.3 Quantum algorithm for persistent homology

We have now a quantum state representing each simplicial complex X_ϵ in the filtration, we can introduce the tools described in [63] which use quantum information processing to analyse the topological properties of the simplicial complexes. We can identify the Hilbert space C_k^ϵ with the k-chain group (introduced in sec. 4.1) then, as we did in the classic case, we can define the boundary map by mapping each k-simplex state to a sum of $(k-1)$-simplex states in the following way:

$$\partial_k |\sigma_k\rangle = \sum_i (-1)^i |\sigma_k(i)\rangle \tag{4.3.1}$$

where $\sigma_k(i)$ is the $(k-1)$-simplex obtained by σ_k omitting the ith vertex. The boundary operator so defined acts on the space of all simplices W_k, it can be restricted to C_k^ϵ by projecting onto it, and we will denote it by $\partial_k^\epsilon = \partial_k P_k^\epsilon$. The **Dirac operator** B^ϵ is constructed as:

$$B^\epsilon = \begin{pmatrix} 0 & \partial_1^\epsilon & 0 & & & \\ \partial_1^{\epsilon *} & 0 & \partial_2^\epsilon & & \dots & \\ 0 & \partial_2^{\epsilon *} & 0 & & & \\ & \dots & & & \dots & \\ & & & 0 & \partial_{n-1}^\epsilon & 0 \\ & \dots & & {\partial_{n-1}^\epsilon}^* & 0 & \partial_n^\epsilon \\ & & & 0 & \partial_n^{\epsilon *} & 0 \end{pmatrix} \tag{4.3.2}$$

$$(B^\epsilon)^2 = \begin{pmatrix} \partial_1^\epsilon \partial_1^{\epsilon *} & 0 & & \\ 0 & \partial_1^{\epsilon *}\partial_1^\epsilon + \partial_2^\epsilon \partial_2^{\epsilon *} & \dots & \\ & \dots & \dots & \\ & \dots & {\partial_{n-1}^\epsilon}^* \partial_{n-1}^\epsilon + \partial_n^\epsilon \partial_n^{\epsilon *} & 0 \\ & & 0 & \partial_n^{\epsilon *}\partial_n^\epsilon \end{pmatrix} \tag{4.3.3}$$

$(B^\epsilon)^2$ is a block matrix with all the kth-combinatorial Laplacian in the diagonal.

It is useful to notice that $\ker(B^\epsilon) = \ker((B^\epsilon)^2)$ then this latter is equal to $\oplus_k \ker(\mathrm{L}_k^\epsilon)$; as a consequence of Theorem 4.1.3 finding Betti numbers in all dimension is equivalent to computing the dimension of the kernel of the combinatorial Laplacian, which can be done by identifying the singular values and singular vectors of the Dirac operator $\ker(B^\epsilon) \cong \oplus_k H_k(X^\epsilon)$.

The QPA decomposes $|\rho^\epsilon\rangle$ into the eigenvectors of the Laplacian

$$|\rho^\epsilon\rangle = \frac{1}{\sqrt{M}} \sum_k \alpha_k \sum_{j=0}^{M-1} |j\rangle \, (\lambda_k)^j \, |\chi_k\rangle = \frac{1}{\sqrt{M}} \sum_k \alpha_k \, |\chi_k\rangle \sum_{j=0}^{M-1} e^{i w_k j} \, |j\rangle \tag{4.3.4}$$

, where M is the number of index qubits used to store the state ρ^ϵ, $|\chi_k\rangle$ are the eigenvectors of B^ϵ, λ_k are the corresponding eigenvalues, and $\alpha_j = \langle \chi_j | \rho^\epsilon \rangle$. Only the eigenvalues related to the biggest eigenspaces will register as non-zero in the decomposition 4.3.4 [64] A quantum fast Fourier trasform performed on the M index qubits will reveal the phases w_k and thereby the eigenvalues λ_k. Qne is then able to obtain each eigenvalue with probability $|\alpha_k|^2$. We define as quantum Betti number q^ϵ as the magnitude α_k corresponding to the eigenvalue $\lambda_k = 0$.

We can now construct the full decomposition of the simplicial complex X^ϵ in terms of eigenvectors and eigenvalues of the combinatorial Laplacian at each scale ϵ. The aim of our work is to be able to monitor the evolution of the quantum betti number q_k^ϵ through the filtration. That is, to be able to check at step $\ell > \epsilon$ if the decomposition of ρ_k^ϵ as linear combination of eigenvectors of L_k^ℓ still contains non-zero coefficient associated with the zero eigenvalue.

The algorithm we introduced above takes as input the uniform mixture ρ_k^ϵ and the Dirac operator $B^\epsilon = \oplus \partial_k P_k^\epsilon$ and it returns the quantum Betti number at step q_k^ϵ. If we apply the same algorithm to the input ρ_k^ϵ and the Dirac operator $B^\ell = \oplus \partial_k P_k^\ell$, we will obtain the magnitude of the eigenvalue zero in the decomposition of ρ_k^ϵ into the eigenvectors of B^ℓ, that is the probability that performing a measurement on B^ℓ from ρ_k^ϵ yields zero.

This application is mathematically justified by the theoretical results we obtained in sec.4.1. With some abuse of notation we can identify our modification of the quantum algorithm with $F|_{C^{\epsilon,\ell}} B^\epsilon$ where $F|_{C^{\epsilon,\ell}} : C^\epsilon \hookrightarrow C^\ell$ is the inclusion of the space C^ϵ into C^ℓ.

What we obtain is therefore the probability that the state describing the k-simplices at step ϵ can collapse into a state representing an homological cycle of the simplicial complex at step ℓ. We denote this persistent quantum Betti number by $q_k^{\epsilon,\ell}$.

Conclusions

In this chapter we extended the existing quantum machine learning methods for topological analysis to track Betti numbers along a filtration of simplicial

complexes yielding a description analogous to that introduced by Ghrist [46] for the classical case. This result can be achieved by tracking the evolution of the eigenspaces of the combinatorial laplacian through the filtration.

In the future, we intend to use the same insight explained in this chapter to track not only the appearance and disappearance of cycles (barcode [46]), but also the evolution of the components of the shortest representative of each homological cycle. Applying the quantum phase algorithm to the the Hermitian matrix B_ϵ, starting from $|s_k\rangle$ where s_k is a k-simplex in the complex at scale ϵ. We obtain a decomposition of $|s_k\rangle$ according to the eigenvectors of B_ϵ, we can then collapse onto the one corresponding to the null eigenvalue. Repeating this procedure for all simplexes s_k we can hen reconstruct the harmonic representative of the homological cycles (eigenvector of the zero eigenvalue of the laplacian). Analogously as we did for the Betti numbers, we can repeat the process for $B_{\epsilon+\Delta}$ starting from each $|s_k\rangle$ k-simplex in the harmonic cycle found at scale ϵ. In this case we will obtain a decomposition of $|s_k\rangle$ according to the eigenvectors of $B_{\epsilon+\Delta}$. Repeating the process for all simplexes s_k in the harmonic cycle will allow us to record the evolution of the cycle representative throughout the entire filtration, and to track not only the topology, but also the evolving geometry of the simplicial complexes. Even though this method would only give an approximation, this kind of analysis is still not obtainable through classic methods due to its computational complexity.

The method we just described gives an exponential speedup over the best classical algorithms for topological data analysis. Classical algorithms for finding the eigenvalues and eigenvectors of the combinatorial laplacian Δ_k

and their dimension take $\binom{n}{k}^2 \sim O(2^{2n})$ computational steps using Gaussian elimination [35, 33]. On a quantum computer the quantum phase algorithm can project the simplex states $|\psi\rangle_k^\epsilon$ onto the eigenspace of the Dirac operator B^ϵ and find the corresponding eigenvalues to accuracy δ in time $O(\frac{n^5}{\delta\sqrt{\varsigma}})$, where ς is the accuracy we chose to construct our simplex state. The algorithm identifies the dimension of eigenspaces in time $O(\frac{n^5}{\delta\sqrt{\varsigma\eta_i}})$, where η_i is the dimension of the ith eigenspace divided by the cardinality of the k-simplex space $|S_k^\epsilon|$. Given the insight this method provides on the structure of the combinatorial laplacians, we believe this algorithm could be put into use to study other algebraic and combinatorial problems in topological data analysis.

Conclusions

Simplicial complexes have been used since the late 1800s to transform complicated topological problems into more familiar algebraic ones. With the advent of computers, their ability to store in discrete form geometric and topological information, has made them a key tool in image recognition and, more recently, in data analysis to successfully approximate the topology of the space underlying a data set.

In this thesis we examine the role of simplicial complexes in data analysis, and, to enhance the versatility of this approach, we tackle the shortcomings of this application from three different perspectives: practical, theoretical and algorithmic. Our contribution in this account is threefold: we build and test a tool which can validate the significance of the structural features of simplicial complexes, we advance our understanding of the applicability of a simplicial approach to weighted graphs, and lastly we provide a new point of view on persistence theory, providing a quantum tool which can track the evolution of single features along a filtration of nested simplicial complexes.

Our first contribution, the simplicial configuration model, can be readily used in practice. The SCM can generate any kind of simplicial complex. Moreover, its ability to fix either, or both, degree and size distributions make the simplicial

configuration model an invaluable method to study real world data sets, by randomizing their structure while fixing their key features. This property of the method is also one of its limitations. Extracting the facet size distribution can be quite costly, when real data does not represent high-dimensional relations and does not implicitly contain the information.

Unlike the SCM, the other original results presented in this thesis are not directly applicable. In fact, the categorical equivalences and adjunctions presented here, give only a detailed description of the categorical relations between weighted graphs and P-persistent objects, and the quantum contribution can only be theoretical, since large scale quantum computers have yet to be built. Nevertheless, these results are of great importance to topological data analysis, as they aid in the construction of suitable filtrations of simplicial complexes, and open the field to unexplored approaches to the computation of persistent homology.

Several questions remain to be addressed. On the basis of the findings presented in this thesis, further research on the development of persistent homological tools would be of great interest. In particular, in our future research on the simplicial configuration model, we intend to concentrate on the homological aspects of the ensemble, and in expanding the model to the weighted case. Moreover, it would be interesting to use our quantum results to track not only the appearance and disappearance of cycles, but also the evolution of the components of the shortest representative of each homological cycle along the filtration. Even though this method would only give an approximation, this kind of analysis is still impractical through classic methods due to its computational complexity.

In conclusion, simplicial complexes are a very promising tool in data analysis and can open new prospects in analyzing high-order data without loss of information. The topological data analysis has become a useful tool for uncovering qualitative features in data which cannot be recovered in other way. The results presented in this thesis will open new possibilities to the application of topological tool for statistical, quantum, and network science application.

References

[1] Abrams, D. S. and Lloyd, S. (1999). Quantum algorithm providing exponential speed increase for finding eigenvalues and eigenvectors. *Physical Review Letters*, 83(24):5162.

[2] Aleksandrov, P. S. (1972). Poincaré and topology. *Russian Mathematical Surveys*, 27(1):157–168.

[3] Barabási, A.-L. and Albert, R. (1999). Emergence of scaling in random networks. *science*, 286(5439):509–512.

[4] Barmak, J. A. (2011). *Algebraic topology of finite topological spaces and applications*, volume 2032. Springer.

[5] Barnes, R. and Burkett, T. (2010). Structural redundancy and multiplicity in corporate networks. *Connections*, 30(2):4–20.

[6] Basu, S. (1996). On bounding the betti numbers and computing the euler characteristic of semi-algebraic sets. In *Proceedings of the twenty-eighth annual ACM symposium on Theory of computing*, pages 408–417. ACM.

[7] Basu, S. (2003). Different bounds on the different betti numbers of semi-algebraic sets. *Discrete and Computational Geometry*, 30(1):65–85.

[8] Basu, S. (2008). Computing the top betti numbers of semialgebraic sets defined by quadratic inequalities in polynomial time. *Foundations of Computational Mathematics*, 8(1):45–80.

[9] Basu, S. (2014). Algorithms in real algebraic geometry: a survey. *arXiv preprint arXiv:1409.1534*.

[10] Bauer, U., Kerber, M., and Reininghaus, J. (2014). Clear and compress: Computing persistent homology in chunks. In *Topological Methods in Data Analysis and Visualization III*, pages 103–117. Springer.

[11] Bender, E. A. and Canfield, E. R. (1978). The asymptotic number of labeled graphs with given degree sequences. *Journal of Combinatorial Theory, Series A*, 24(3):296–307.

[12] Bianconi, G. and Barabási, A.-L. (2001). Bose-einstein condensation in complex networks. *Physical review letters*, 86(24):5632.

[13] Bianconi, G. and Rahmede, C. (2015). Complex quantum network manifolds in dimension d> 2 are scale-free. *Scientific reports*, 5.

[14] Bianconi, G. and Rahmede, C. (2016a). Emergent hyperbolic geometry of growing simplicial complexes. *arXiv preprint arXiv:1607.05710*.

[15] Bianconi, G. and Rahmede, C. (2016b). Network geometry with flavor: from complexity to quantum geometry. *Physical Review E*, 93(3):032315.

[16] Bron, C. and Kerbosch, J. (1973). Algorithm 457: finding all cliques of an undirected graph. *Communications of the ACM*, 16(9):575–577.

[17] Bubenik, P. and Scott, J. A. (2014). Categorification of persistent homology. *Discrete & Computational Geometry*, 51(3):600–627.

[18] Cagliari, F., Ferri, M., and Pozzi, P. (2001). Size functions from a categorical viewpoint. *Acta Applicandae Mathematica*, 67(3):225–235.

[19] Carlsson, G. (2009). Topology and data. *Bulletin of the American Mathematical Society*, 46(2):255–308.

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

[21] Carlsson, G., Zomorodian, A., Collins, A., and Guibas, L. J. (2005). Persistence barcodes for shapes. *International Journal of Shape Modeling*, 11(02):149–187.

[22] Carrière, M. and Oudot, S. (2015). Structure and stability of the 1-dimensional mapper. *arXiv preprint arXiv:1511.05823*.

[23] Chazal, F., Crawley-Boevey, W., and de Silva, V. (2016). The observable structure of persistence modules. *Homology, Homotopy and Applications*.

[24] Chazal, F., De Silva, V., and Oudot, S. (2014). Persistence stability for geometric complexes. *Geometriae Dedicata*, 173(1):193–214.

[25] Chazal, F. and Lieutier, A. (2007). Stability and computation of topological invariants of solids in $\setminus^n$. *Discrete & Computational Geometry*, 37(4):601–617.

[26] Cohen-Steiner, D., Edelsbrunner, H., and Harer, J. (2007). Stability of persistence diagrams. *Discrete & Computational Geometry*, 37(1):103–120.

[27] Costa, A. and Farber, M. (2016). Random simplicial complexes. In *Configuration Spaces*, pages 129–153. Springer.

[28] Courtney, O. T. and Bianconi, G. (2016). Generalized network structures: The configuration model and the canonical ensemble of simplicial complexes. *arXiv preprint arXiv:1602.04110*.

[29] De Silva, V. and Carlsson, G. (2004). Topological estimation using witness complexes. *Proc. Sympos. Point-Based Graphics*, pages 157–166.

[30] De Silva, V. and Ghrist, R. (2007a). Coverage in sensor networks via persistent homology. *Algebraic & Geometric Topology*, 7(1):339–358.

[31] De Silva, V. and Ghrist, R. (2007b). Homological sensor networks. *Notices of the American mathematical society*, 54(1).

[32] Dey, T. K., Mémoli, F., and Wang, Y. (2016). Multiscale mapper: topological summarization via codomain covers. In *Proceedings of the Twenty-Seventh Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 997–1013. SIAM.

[33] Dixon, J. D. (1982). Exact solution of linear equations usingp-adic expansions. *Numerische Mathematik*, 40(1):137–141.

[34] Dowker, C. H. (1952). Homology groups of relations. *Annals of mathematics*, pages 84–95.

[35] Eberly, W., Giesbrecht, M., Giorgi, P., Storjohann, A., and Villard, G. (2006). Solving sparse rational linear systems. In *Proceedings of the 2006 international symposium on Symbolic and algebraic computation*, pages 63–70. ACM.

[36] Eckmann, B. (1944). Harmonische funktionen und randwertaufgaben in einem komplex. *Commentarii Mathematici Helvetici*, 17(1):240–255.

[37] Edelsbrunner, H. and Harer, J. (2008). Persistent homology-a survey.

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

[39] Faust, K. (1997). Centrality in affiliation networks. *Social networks*, 19(2):157–191.

[40] Friedman, J. (1998). Computing betti numbers via combinatorial laplacians. *Algorithmica*, 21(4):331–346.

[41] Frosini, P. (1990). A distance for similarity classes of submanifolds of a euclidean space. *Bulletin of the Australian Mathematical Society*, 42(03):407–415.

[42] Frosini, P. (1992). Measuring shapes by size functions. In *Intelligent Robots and Computer Vision X: Algorithms and Techniques*, pages 122–133. International Society for Optics and Photonics.

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

[44] Galaskiewicz, J. (1985). Interorganizational relations. *Annual review of sociology*, pages 281–304.

[45] Ghoshal, G., Zlatić, V., Caldarelli, G., and Newman, M. (2009). Random hypergraphs and their applications. *Physical Review E*, 79(6):066118.

[46] Ghrist, R. (2008). Barcodes: the persistent topology of data. *Bulletin of the American Mathematical Society*, 45(1):61–75.

[47] Ghrist, R. and Muhammad, A. (2005). Coverage and hole-detection in sensor networks via homology. In *Proceedings of the 4th international symposium on Information processing in sensor networks*, page 34. IEEE Press.

[48] Giusti, C., Ghrist, R., and Bassett, D. S. (2016). Two's company, three (or more) is a simplex. *Journal of Computational Neuroscience*, pages 1–14.

[49] Giusti, C., Pastalkova, E., Curto, C., and Itskov, V. (2015). Clique topology reveals intrinsic geometric structure in neural correlations. *Proceedings of the National Academy of Sciences*, 112(44):13455–13460.

[50] Grover, L. K. (1996). A fast quantum mechanical algorithm for database search. In *Proceedings of the twenty-eighth annual ACM symposium on Theory of computing*, pages 212–219. ACM.

[51] Harrow, A. W., Hassidim, A., and Lloyd, S. (2009). Quantum algorithm for linear systems of equations. *Physical review letters*, 103(15):150502.

[Hatcher] Hatcher, A. *Algebraic topology. 2002*, volume 606.

[53] Hodge, W. V. D. (1989). *The theory and applications of harmonic integrals*. CUP Archive.

[54] Jonsson, J. (2008). *Simplicial complexes of graphs*. Springer.

[55] Kahle, M. (2009). Topology of random clique complexes. *Discrete Mathematics*, 309(6):1658–1671.

[56] Kahle, M. (2012). Sharp vanishing thresholds for cohomology of random flag complexes. *arXiv preprint arXiv:1207.0149*.

[57] Kahle, M. (2014). Topology of random simplicial complexes: a survey. *AMS Contemp. Math*, 620:201–222.

[58] Konopka, T., Markopoulou, F., and Severini, S. (2008). Quantum graphity: a model of emergent locality. *Physical Review D*, 77(10):104029.

[59] Kovalevsky, V. A. (1989). Finite topology as applied to image analysis. *Computer vision, graphics, and image processing*, 46(2):141–161.

[60] Kozlov, D. (2008). Combinatorial algebraic topology, volume 21 of algorithms and computation in mathematics. *Springer, Berlin*, 4:5.

[61] Linial, N. and Meshulam, R. (2006). Homological connectivity of random 2-complexes. *Combinatorica*, 26(4):475–487.

[62] Linial, N., Meshulam, R., and Rosenthal, M. (2010). Sum complexes—a new family of hypertrees. *Discrete & Computational Geometry*, 44(3):622–636.

[63] Lloyd, S., Garnerone, S., and Zanardi, P. (2016). Quantum algorithms for topological and geometric analysis of data. *Nature communications*, 7.

[64] Lloyd, S., Mohseni, M., and Rebentrost, P. (2014). Quantum principal component analysis. *Nature Physics*, 10(9):631–633.

[65] Lord, L.-D., Expert, P., Fernandes, H. M., Petri, G., Van Hartevelt, T. J., Vaccarino, F., Deco, G., Turkheimer, F., and Kringelbach, M. L. (2016). Insights into brain architectures from the homological scaffolds of functional connectivity networks. *Frontiers in Systems Neuroscience*, 10.

[66] McCord, M. C. et al. (1966). Singular homology groups and homotopy groups of finite topological spaces. *Duke Math. J*, 33(3):465–474.

[67] Meshulam, R. and Wallach, N. (2009). Homological connectivity of random k-dimensional complexes. *Random Structures & Algorithms*, 34(3):408–417.

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

[69] Molloy, M. and Reed, B. (1995). A critical point for random graphs with a given degree sequence. *Random structures & algorithms*, 6(2-3):161–180.

[70] Munkres, J. R. (1984). *Elements of algebraic topology*, volume 2. Addison-Wesley Menlo Park.

[71] Nakahara, M. (2008). Quantum computing: an overview. In *Mathematical Aspects of Quantum Computing 2007*, volume 1, pages 1–53. World Scientific.

[72] Nakahara, M. and Ohmi, T. (2008). *Quantum computing: from linear algebra to physical realizations*. CRC press.

[73] Newman, M. (2010). *Networks: an introduction*. Oxford university press.

[74] Newman, M. E. (2003). The structure and function of complex networks. *SIAM review*, 45(2):167–256.

[75] Newman, M. E., Strogatz, S. H., and Watts, D. J. (2001). Random graphs with arbitrary degree distributions and their applications. *Physical review E*, 64(2):026118.

[76] Nicolau, M., Levine, A. J., and Carlsson, G. (2011). Topology based data analysis identifies a subgroup of breast cancers with a unique mutational profile and excellent survival. *Proceedings of the National Academy of Sciences*, 108(17):7265–7270.

[77] Otter, N., Porter, M. A., Tillmann, U., Grindrod, P., and Harrington, H. A. (2015). A roadmap for the computation of persistent homology. *arXiv preprint arXiv:1506.08903*.

[78] Patania, A., Vaccarino, F., and Giovanni, P. (2017a). P-persistent homology of finite topological spaces. *Rendiconti del Seminario Matematico, Universitá e Politecnico di Torino*.

[79] Patania, A., Vaccarino, F., and Giovanni, P. (2017b). The shape of simplicial collaboration. *Submitted to EPJ Data Science*.

[80] Patania, A., Vaccarino, F., and Giovanni, P. (2017c). Topological analysis of data. *Submitted to EPJ Data Science*.

[81] Patania, A., Vaccarino, F., Zanardi, P., and Lloyd, S. (2017d). Quantum barcode for persistent homology. *In preparation*.

[82] Patania, A., Veronese, M., Selvaggi, P., Turkheimer, F., Vaccarino, F., Expert, P., and Giovanni, P. (2017e). Dopaminergic pathways revealed through topological studies of gene-expression. *In preparation*.

[83] Patania, A., Veronese, M., Selvaggi, P., Turkheimer, F., Vaccarino, F., Expert, P., and Giovanni, P. (2017f). Topological networks for gene-expression of microarray data. *In preparation*.

[84] Petri, G., Expert, P., Turkheimer, F., Carhart-Harris, R., Nutt, D., Hellyer, P., and Vaccarino, F. (2014). Homological scaffolds of brain functional networks. *Journal of The Royal Society Interface*, 11(101):20140873.

[85] Petri, G., Scolamiero, M., Donato, I., and Vaccarino, F. (2013a). Networks and cycles: a persistent homology approach to complex networks. In *Proceedings of the European Conference on Complex Systems 2012*, pages 93–99. Springer.

[86] Petri, G., Scolamiero, M., Donato, I., and Vaccarino, F. (2013b). Topological strata of weighted complex networks. *PloS one*, 8(6):e66506.

[87] Poincaré, H. (1895). Analysis situs. *Journal de l'École Polytechnique*, 1(2):1–123.

[88] Singh, G., Memoli, F., and Carlsson, G. (1991). Mapper: a topological mapping tool for point cloud data. In *Eurographics symposium on point-based graphics*.

[89] Singh, G., Mémoli, F., and Carlsson, G. E. (2007). Topological methods for the analysis of high dimensional data sets and 3d object recognition. In *SPBG*, pages 91–100.

[90] Steenrod, S. E.-N. and Eilenberg, S. (1952). Foundations of algebraic topology.

[91] Tausz, A., Vejdemo-Johansson, M., and Adams, H. (2011). Javaplex: A research software package for persistent (co) homology. *Software available at http://code. google. com/javaplex*.

[92] Timothy Gowers, A. E. J. B.-G. and Leader, I., editors (2008). *The Princeton Companion to Mathematics*. Princeton University Press.

[93] Wu, Z., Menichetti, G., Rahmede, C., and Bianconi, G. (2015). Emergent complex network geometry. *Scientific Reports*, 5:10073.

[94] Young, J.-G. and Patania, A. (2017). Simplicial configuration model. *In preparation*.

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

[96] Zuev, K., Eisenberg, O., and Krioukov, D. (2015). Exponential random simplicial complexes. *Journal of Physics A: Mathematical and Theoretical*, 48(46):465002.

Computational complexity

In this we briefly list the computational cost for performing the different algorithms and methods introduced in this thesis. For a more thorough study on the matter we refer the reader to [77, 54, 89].

Building Simplicial Complexes

The Čech complex construction requires the computation of $2^{|A|}$ intersections, where $|\ A\ |$ is the number of open sets in the considered cover, which is equal to the number of vertices. The Vietoris-Rips complex is popular in topological analysis thanks to the ease of its construction in every dimension. It is not a nerve as the other previously presented complexes, but it is the clique complex of a particular graph. The Dowker complex construction highly depends on the number of points chosen for the landmark set.

Construction method	Computational Complexity
Čech Complex	$O(2^n)$
Vietoris-Rips Complex	$O(\binom{n}{2} + n^k)$
Witness Complex/Dowker Complex	$O(2^{\lvert L \rvert})$
Mapper	$C * O(\prod_i (f_i - 1))$

Table 1 Summary of the computational complexity for each method to construct simplicial complexes where: n is the number of points in the data set, k is the dimension of the Vietoris-Rips complex, $\mid L \mid$ is the number of points in the landmark set, C is the computational complexity related to the chosen clustering method, and f_i is the number of bins for the ith filter.

Classical and quantum computation of persistent homology

Classical algorithms for computing the kth simplicial homology of a complex Σ relies on the reduction of the kth boundary matrix which takes at most $O(n_k^3)$ operations, where n_k is the dimension of the k-chain space, i.e. the number of simplices of dimension k in Σ [35, 33]. On a quantum computer the quantum phase algorithm can project the simplex states $|\psi\rangle_k^\epsilon$ onto the eigenspace of the Dirac operator B^ϵ and find the corresponding eigenvalues to accuracy δ in time $O(\frac{n^5}{\delta\sqrt{\varsigma}})$, where ς is the accuracy we chose to construct our simplex state. The algorithm identifies the dimension of eigenspaces in time $O(\frac{n^5}{\delta\sqrt{\varsigma\eta_i}})$, where η_i is the dimension of the ith eigenspace divided by the k-simplex space $|S_k^\epsilon|$.

Ringraziamenti

Nelle prossime righe cercheró di ringraziare tutti coloro che, in un modo o in un altro, mi hanno aiutato ad essere qui oggi. Cercheró di ringraziare tutti, ma se mi dimenticassi di qualcuno non se ne abbia a male.

In primis, vorrei ringraziare la mia famiglia – i miei genitori, le mie sorelle e i miei nonni, Lina e Gaetano – per avermi sempre supportato (e sopportato), perché anche quando vi sentivate trascurati mi siete sempre stati accanto, e non mi avete mai permesso di mollare. Se riesco ad affrontare la vita a testa alta lo devo principalmente a voi.

The most heartfelt thank you goes to Jean-Gabriel Young. He has not been there through the entire journey, but he was definitely there for the hardest part. Thank you for your encouragement, support and for being your positive self every time I fell into the abyss. I will always be grateful.

I want to thank all the researchers that have been part of the I.S.I. family during my time there. You have all been a tremendous inspiration, and i will always be grateful for all the fun we have had in the last four years. A special thanks goes to Giovanni, for offering me a drink every time I needed one, for introducing me to the first sport I could actually enjoy and the wonderful people behind it, but mostly for throwing me off the cliff over and over again, and

being there to help when I could not climb back up on my own. In particular, I would like to thank my Ph.D. comrades: Anna and Giovanna Chiara, we have been through a lot together, and if I made it through it's mostly because you were with me every step of the way. To the other members of the research group: Riccardo, Esther and Andrea. Throughout my Ph.D. they have always been there to support me, and to double-check everything I ever wrote or proved. I also would like to thank Luca Rossi, Kyriaki Kalimeri for the wine, the foosball, and our long conversations on the upsides and downsides of our profession. Thank you for being there when I needed to rant. Among these wonderful nerds a special thanks goes to my scattered nerds: Pietro Coletti, Lorenzo Argante, Zsolt Bertalan and Jacopo Jacopini. During the most difficult times they gave me the moral support i needed, together with a good beer, great boardgames and – when life divided us – long skype calls. And for all of you, that where there when I collapsed, when everything seemed dark and hopeless – Davide, Fran, Indaco, Bernardo, Ubi, Valeria, Giulia, Zoey, Paolo, Corrado – THANK YOU. You all have made Torino my home away from home.

Last but not the least, I would like to thank my dear friends Michele, Alessandro and Davide, who have been with me since the beginning of my studies in Torino, and even when far, gave me the courage to follow my dreams.