Topological Algebra: Tool of Persistent Homology and Its Application In Analysing US School Market

*

Artur Bayramyan †
Advisor: Dr. Steven Sam

June 6, 2021

Abstract

Algebraic topology allows us to use tools from abstract algebra and directly apply them for analyzing topological spaces. This paper aims to discuss the literature and provide a satisfying explanation for the fundamentals of persistent homology (PH). Simplicial complexes and homology are discussed as building blocks for accessing and using the concept of PH that was developed in 2005 by Zomorodian et al., and Carlsson et al,. PH helps to find higher-dimensional topological features (i.e., n dimensional holes) that persist through the filtration process. By implementing some of the techniques discussed, one of the objectives of the paper is to provide a case study on the California school market.

*Acknowledgments: I am very thankful to my advisor Dr. Steven V Sam for his time, comments and for awakening my interests in pure mathematics.

†University of California, San Diego; Email abayramy@ucsd.edu

Contents

"Shape is the global realization of local constraints."

— Dr. Anthony Bak

1 Introduction

In the recent decade, the notion of Big Data and its opportunities and challenges has been of particular interest to specialists within and beyond the field of data science. The main challenge in systematically extracting useful information from big data is not the size of the data per se, but rather its complexity. Datasets representing molecule structure as well as texts and images that are potentially corrupted by noise of incompleteness or have high dimensional components; pose significant complications for conventional methodologies of analysis. Thus, alternative organising principles are required for capturing significant features of data.

"Data has a shape and that shape has a meaning." [1] For example in economics, regression models are used to approximate the shape of the data points using the simplest one dimensional figure such as the line. Basic regressions are built on the principle of least square distance that attempts to fit a line by minimizing the least square distances as depicted in figure 1 below. Such modeling principles give a basic understanding of which variables vary to which way and also allows to make predictions. However, not all the data can be fit with a line. Algebraic methods are particularly unreliable when used for exploring noisy datasets. The simplicity of algebraic methods limits their ability in capturing singular behavior. In this paper, I deviate from capturing individual shapes to modeling: capturing all shapes at once. First, I aim to present satisfying explanation of fundamental tools that are unique for analyzing the shape of the data. I then go onto to discuss their direct application in studying robust features of California school market.

(a) Each point on the graph represents various developing countries such as Tanzania, Albania, Uganda, etc. X - axis shows the average logarithmic amount of liquid assets in the household, while Y - axis represents the percentage of self - employment observed in country. The straight line passing in between the points is the line of best fit according to the least square regression where the algorithms simply aims to minimize the sum of square distances of R shown as a dotted line. *Panel data used for scattering the graph is obtained from various World Bank household surveys.*

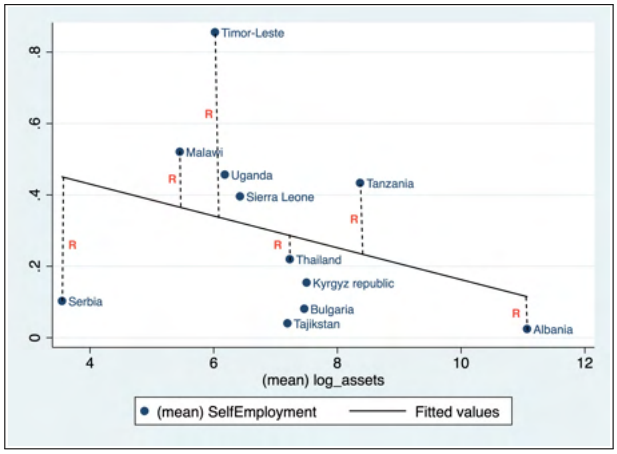


Figure 1: Simple Example of Modeling

The set of tools discussed in this paper stem from topological algebra which studies measuring and representing of shapes. Topological algebra itself is a branch of pure mathematics which was developed in the 1700s. The three main principles of topology that allow for extracting information about shapes are: coordinate freeness, invariance under small deformations and compressed representation of shapes. Topologists like to use the example of transforming a "donut" (i.e., torus) into a cup with a handle as an

Figure 2: It can be seen how the torus in the top left hand corner of the figure is deformed into a cup with a handle. Boundaries are preserved as our initial dough-nut has a hole similar to the handle of the cup represented on the top left corner of the figure. Retrieved: `https://cems.riken.jp/en/laboratory/qmtrt`

example that demonstrates what topology studies. In figure figure 2 below, one can see how the transformation is done while stretching the torus but keeping the boundaries identical. In topology stretching/diminishing objects when keeping structural measures results in a more fundamental knowledge about the shape.

To learn about the shape of a data set, first it needs to be reduced in a comprehensive way. Simplicial complexes, which will be described in chapter (2) approximate the structure of the data and allow numerical computation of its features. Simplices further expand on representing data in the form of points, line segments, triangles and other n - dimensional realizations of simplices. In applied work, metric data sets are simply transformed into simplicial complexes by building filtered complex on top of the data.

Moreover, homology, a well established qualitative principle in abstract algebra, serves the purpose of decomposing higher dimensional features of simplices into its subsequent lower dimensional versions. Homology aims to identify the relationship between n and $n-1$ dimensional components of simplices. This is done through finding boundary homomorphisms of the chain groups by discarding n - dimensional cycles that are also boundaries of the simplicial complex. Such cycles, which are not boundaries, are called

(a) Looking at the torus discussed above, we can notice that it has two types of holes. One dimensional hole which is represented as a 2-cycle is represented with a red contour. Besides slicing the torus through its x - axis, one can try slicing it using the y - axis which will allow to identify another type of hole. Furthermore, the blue contour describes 2 dimensional hole which are voids. *Retrieved from:* `https://www.wikiwand.com/en/Torus#/Topology`

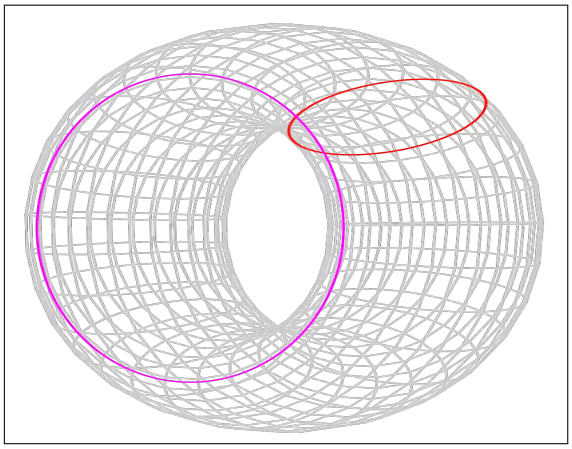
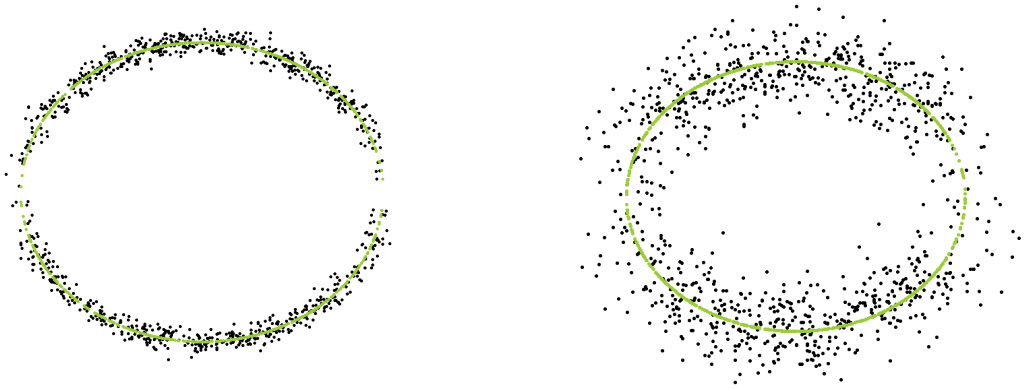


Figure 3: Torus an example of a figure containing both 1 and 2 dimensional holes



(a) noise: std=0.2

(b) noise: std=0.8

Figure 4: Uniformly Distributed Point Cloud Data Set (Randomly chosen 1000 points) of a Circle with Varying Noise

to be "holes". For example, 1-dimensional holes such as loops and 2 dimensional holes such as voids are depicted in Figure 1.2 below.

The artificially created grid of point clouds drawn in figure 4 above both resemble the structure of a one dimensional hole as the points are accumulated around a circle, drawn with a green contour. The figure 4a varies from the figure 4b with the noise that is introduced in the set of data. Unlike conventional methods of analysis, the tools and techniques that are discussed in this paper are not prone to failure when data is colluded with such noise. Another example of a 2 dimensional hole is depicted in figure 5 where points are accumulated on the outline of a void with a radius of one.

Originally, persistent homology and particularly the persistence algorithm was introduced in 2002 with a strong emphasis on the data analysis aspect of it [16]. Three years later, in 2005 paper, Zomorodian et al. and Carlson et al. published a paper that was devoted to the ore mathematical framework of persistent homology. Particularly the paper aimed to show that the algorithm works not only for $\mathbf{Z} \oplus \mathbf{Z}$ field coefficent but for any field of coefficients. PH discussed in Chapter (4), is the study of inclusive subsets of simplicial complexes that are increasing through filtration. This will give insight on the structure of n-dimensional holes along different sublevel sets of filtration. Interpretation tools, such as barcodes and persistent diagrams, reveal the structural complexity of the holes with its corresponding persistence intervals. Exploring such holes would allow us to see which features are long lasting (i.e., persistent) and which features are temporary shocks in the data.

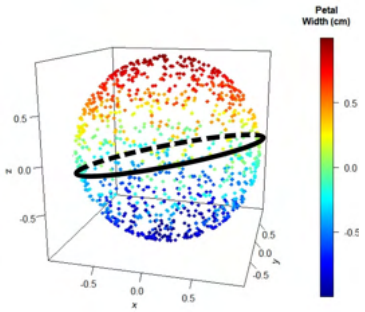


Figure 5: Uniformally distributed Point Cloud Data of 1000 points

Topological data analysis and particularly the novel tool of persistent homology has been a topic of interest to a miscellaneous set of scientific fields such as Data Science, Statistics, Machine Learning and more. Generally, it can be divided into two types of developments: analysing forms of complicated point cloud data and further developments in the concept itself. Point cloud data, or its equivalent input types, can range from weighted network data, wherein the location of vertices in space do not have meaning, to types of data where vertices express their spatial location of units of interest in a given system.

The paper is organised as follows. In chapter (2), and (3), I introduce the required fundamental background in topology. Firstly, I present the concept of simplical representation of data. Secondly, I introduce homology in the simplicial representation of data in chapter (3). Chapter (4) will serve as an introduction to the tool of persistent homology which is based on applying homology to the simplices that are produced in increasing sequence of spaces. In chapter (5), I will implement techniques to provide a PH case study on the California school market.

2 Simplicial Complexes

2.1 Simplex

2.1.1 Vectors

On Euclidean space, two distinct points determine a line segment. The subset of points between the vertices v^0 and v^1 can be expressed by $\lambda^0 v^0 + \lambda^1 v^1$ such that $\{\lambda^0 + \lambda^1 = 1,\ \lambda^0 \geq 0,\ \lambda^1 \geq 1\}$. Similarly a point P which is either on the edges of a triangle or inside the triangle, can be denoted as a convex combination of the three non-colinear vertices v^0, v^1, and v^2. Here, point P can be written as $P = \lambda^0 v^0 + \lambda^1 v^1 + \lambda^2 v^2$ such that the barycentric coordinates $\lambda^1, \lambda^2, \lambda^3$ impose two conditions: $\sum_{i=0}^{2} \lambda^i = 1$ and $\lambda^i \geq 0$ where $i = 0, 1, 2$. These two examples motivate definitions of a-independent and a-dependent vectors which is required for defining simplices.

Definition 2.1. Affinely Independent (a-independent) Vectors
Given $(n+1)$ set of vectors $v^0, \ldots, v^n \in \mathbb{R}^n$ where $n > 1$ are called a-independent of one another if $\{v^1 - v^0, \ldots, v^n - v^{n-1}\}$ are linearly independent.

Definition 2.2. Affinely dependent (a-dependent) Vectors
Conversely, given $v^0 \ldots v^n \in \mathbb{R}^n$, vector v is called affinely dependent if $\exists \lambda^i$ such that

$$\lambda^0 + \cdots + \lambda^n = 1 \quad \textit{and} \quad v = \lambda^0 v^0 + \cdots + \lambda^1 v^1$$

Two arbitrary points are a-independent iff they are not the same. Similarly, three points are a-independent iff they are not colinear. The definition about the independence of the set of vectors are required for defining simplices. In order to mark arbitrary points of an object one can utilize its barycentric coordinates and the boundary points.

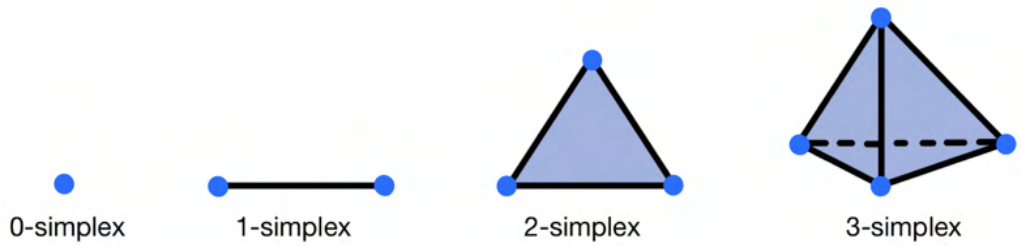


Figure 6: Examples of N-dimensional Simplices

2.1.2 Closed Simplex

Definition 2.3. Closed Simplex
A closed simplex is defined by

$$e_n = e(v^0, \ldots, v^n) = \{\sum_{i=1}^{n} \lambda_i v_i : \ \lambda_i \geq 0, \ \sum_{i=1}^{n} \lambda_i = 1\}$$

where $v^0, \ldots, v^n$ are a set of a-independent vectors and $\lambda_i \geq 0$ are the barycentric coordinates of the points.

The points $v^0, v^1, \ldots, v^n$ are called the vertices of simplex e_n and the number n signifies the dimension of the simplex. In the case where $\lambda^i > 0$ for all i the points are in the interior of the simplex. Expanding on the definition above, an *open simplex* is composed exclusively of the interior points of the simplex. If at least one of the barycentric coordinates is equal to 0 the points of interest lie on the boundary of the simplex. For example, if one of the barycentric coordinates of a 2-simplex is equal to 0 then the points of analysis formulate one of the three edges of the triangle.

Below, are some examples of simplices: 0-simplex constitutes a vertex which is a point itself; The line segment (1-simplex) results from connecting two 0-simplices; The 2-simplex is a triangle produced by three non-colinear points or equivalently three linearly independent 1-simplices; Finally, the 3-simplex is a tetrahedron built from a union of four 0-simplices and four 2-simplices.

Any simplex that is spanned by the subset of $v^0, \ldots v^n$ is called a face of the simplex.

Definition 2.4. Face
The face of a simplex $e_n = (v^0, \ldots, v^n)$ is a simplex created from a subset of vertices $\{v^0, \ldots, v^n\}$. If e_f is a face then we write $e_n > e_f$ or $e_f < e_n$. If the subset does not include all of the vertices of the simplex, the face is called a proper face.

Remark. *(On Boundaries)*
It is important to note that faces are simplices as well. Faces and other lower dimensional elements such as vertices, edges, triangles are connected together to create higher dimensional structures. To better understand simplices realized in $\mathbb{R}^n$, one needs to understand their composite structure. For example, the boundaries of the 3-simplex is composed of four 2-simplices; while the boundary of the 2-simplex itself is created by three 1-simplices. Thus, boundaries of the simplex can be described as the union of all the proper faces e_f of the simplex e_n. Simplices are inductively developed by using building blocks such as faces and lower dimensional topological structures. In an n-simplex there are $\binom{n+1}{f+1}$, f-dimensional faces. Moreover, there is a total of $\sum_{f=-1}^{n}\binom{n+1}{f+1} = 2^{n+1}$ faces. Working

with simplices may seem to be computationally heavy. However, simplicial world provides a medium where complex objects can be easily decomposed into their simpler, lower dimensional topological counterparts.

2.2 Simplicial Complexes

2.2.1 Introducing Simplicial Complexes

A simplicial complex is a set $|K|$ together with its subset of simplices. Simplicial complexes approximate the structure of the data and allow calculation of its properties.

Definition 2.5. Simplicial Complex
Simplicial complex K is a finite set of simplices in $\mathbb{R}^n$ that satisfy the following two properties

(i) If $e \in K$ and f is a face of e $(f < e)$ then $f \in K$
(ii) If $e \in K$ and $f \in K$ then $e \cap f = \emptyset$ or is a face of e and f.

Condition (i) suggests that every face of a simplex of K is also in K. Whereas, condition (ii) implies that the inteersection of any two simplices of K is a face of both simplices. For example, a 2-simplex with all of its faces is a simplicial complex. The dimension of a simplicial complex determines its highest dimensional simplex. The underlying space called the carrier of a simplicial complex can geometrically be realized by a-independent points in $\mathbb{R}^n$. At the same time, these points must be elements of at least one simplex in K. Simplicial complex, referred as a complex, is the union of low-dimensional n-simplices such as vertices, edges, triangles, etc. Later in the paper I introduce the concept of simplicial homology that focuses on the homological features of a given data.

2.2.2 Examples of Simplicial Complexes

The union of simplices however, are imposed by two restricting conditions 2.2.1(i) and 2.2.1(ii). The latter one is called an intersection condition between two simplices. In the diagram below one can see simplices of dimensions ranging from zero to two coming together to form a simplicial complex K. One can check that both of the conditions 2.2.1(i) and 2.2.1(ii) hold true in figure 7.

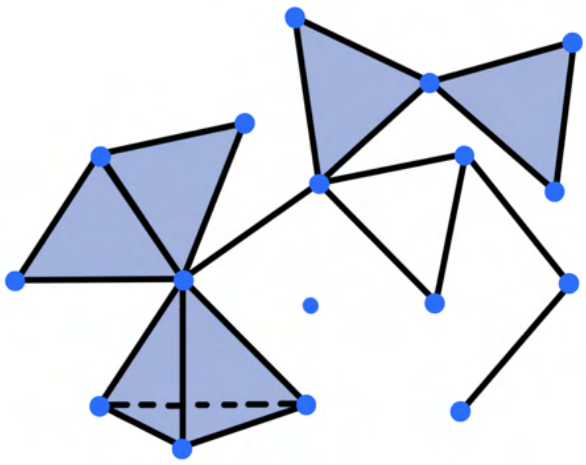


Figure 7: Example of a simplicial Complex

Simplical complexes can be formed in many differnnt ways but generally, violations in forming simplicial complexes are caused by the failure to satisfy the intersection condition. In figures 8 (a), (b), (c) some classical violations are drawn out. In all three examples an intersection of an arbitrary face with a simplex does not result in an empty set or a common face of the simplices in the complex. Figure 8(a) represents a shared partial edge which is not a face of either of the two simplices. In figure 8(a and b), one can easily notice that there are no common faces emerging from the intersection of the two simplices. The 2-simplex in figure 8(c) is slicing through a 3-simplex which is a violation as of itself.

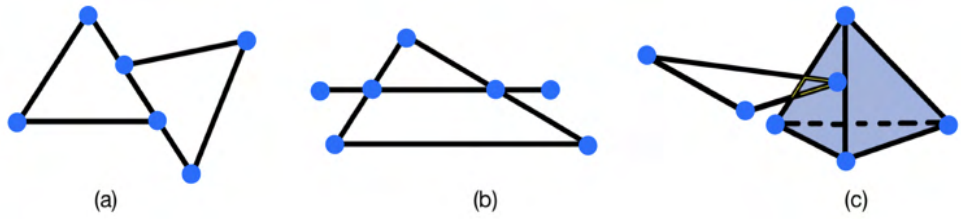


Figure 8: Frequent Violations of Simplicial Complex Construction

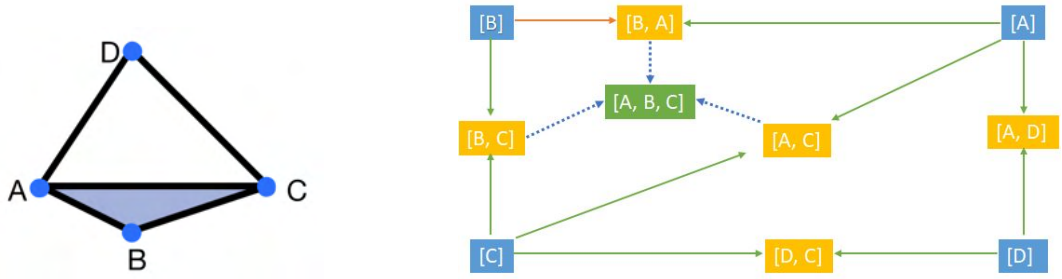
2.2.3 Abstract Simplicial Complexes

This section looks into simplicial complexes from a combinatorial perspective rather than discuss its geometric interpretations. Simplicial complexes gain their underlying geometric information when they are realized in $\mathbb{R}^n$. However geometric realizations of simplicial complexes are not necessary to study topology of complexes. In the 2.1.2 simplicial complex is described as a set with the collection of all its subsets (*power set*). Abstract simplicial complex I defines the union of the underlying subsets of K.

Definition 2.6. Abstract simplicial Complex
Given a finite set σ, an abstract simplicial complex on σ is a collection T of subsets of σ such that
(i) If $v \in \sigma$, then $\{v\} \in T$
(ii) if $\tau \in T$ and $e \subset \tau$, then $e \in T$.
The elements of T are called simplices and if $|\tau| = k + 1$, then τ is called a k-simplex.

The structural units of an abstract simplicial complex (ASC) are families of sets that satisfy the conditions 2.2.2(i) and 2.2.2(ii). Mappings between simplices can be used to present the equivalence set of K without realizing it in $\mathbb{R}^n$. *Vertex scheme* diagram of K in figure 9b is constructed based on figure 9a by removing simplices and retaining their sets of vertices. The dimension of an ASC is defined by its highest dimensional $n-simplex \in T$ that also belongs to the complex. Thus, one can extrapolate information about high dimensional simplicial complexes without its geometric construction in high dimensional space.

Two abstract simplicial complexes I_1, I_2 are called isomorphic if there is a bijection $g : I_1 \longrightarrow I_2$ such that $(\sigma_0, ..\sigma_n) \in T_1$ if and only if $(g(\sigma_0), ...g(\sigma_n)) \in T_2$. Two simplicial complexes are isomorphic if the abstract realizations of those corresponding complexes are isomorphic as well. Abstract graphs composed of points as vertices and edges as line segments between unordered pairs of vertices can be realized in $\mathbb{R}^3$. One can show that any four points of *twisted cubic* are not in the same plane and consequently no three



(a) Abstract simplicial Complex I

(b) Diagram Representing simplicial Complex of Part a) Where Vertices describe the Subsets of n-simplices

Figure 9: Abstract simplicial Complex is Described Using a Family of Sets

points can be on the same line. With a similar concept of proof, i will show that abstract simplicial complexes can be realized in $\mathbb{R}^{2n+1}$.

Theorem 1. Geometric Realization
Every n-dimensional simplicial complex can be realized in $\mathbb{R}^{2n+1}$

Proof. First I need to show that any n-dimensional simplicial complex embeds linearly in $\mathbb{R}^{2n+1}$. Let us define the mapping $f(K(v_i)) \Rightarrow \mathbb{R}^{2n+1}$ of n-dimensional complex K be the injection of vertices to the points in $\mathbb{R}^{2n+1}$, specifically in a way that no hyperplane contains more than $2n+1$ points. Let $v_1, \ldots, v_i$ be vertices. Pick distinct $(t_1, \ldots, t_i) \in \mathbb{R}$ and define $f(v_i) = (t_i,\ t_i^2, \ldots, t_i^{2n+1})$. Our goal here is to show that f is an embedding. Any arbitrary simplex of K has no more than $n+1$ vertices sent to a-independent points. Let $e_1 \in K$ be a simplex with $dim = n_1$ and $e_2 \in K$ be a simplex with $dim = n_2$ where $e_1 \neq e_2$. We have $n_1 + n_2 \leq 2n + 2$. Thus, the points in $e_1 \cup e_2$ are a-independent and unique. This follows from the fact that we require any $2n + 2$ of the points to be a-independent. To check this condition, we need to show that the Vandermonde's determinant does not vanish.

$$\delta = \begin{vmatrix} x_1 - x_2 & x_1^2 - x_2^2 & \cdots & x_1^{2n+1} - x_2^{2n+1} \\ x_1 - x_3 & x_1^2 - x_3^2 & \cdots & x_1^{2n+1} - x_3^{2n+1} \\ \vdots & \vdots & \ddots & \vdots \\ x_1 - x_{2n+2} & x_1^2 - x_{2n+2}^2 & \cdots & x_1^{2n+1} - x_{2n+2}^{2n+1} \end{vmatrix}$$

The determinant δ vanishes if and only if the x's are not distinct. As a result, $f(l)$ and $f(m)$ are disjoint which means that f is an embedding for K. □

2.2.4 Relations Between simplicial Complexes

Up until now, I used simplices as fundamental units for constructing simplicial complexes. However there are many more interesting subdivisions of simplicial complexes, namely: subcomplexes. Sub-collection of $|K|$ the contains all faces of the elements defines the subcomplex.

Definition 2.7. simplicial Subcomplex
A subset L of simplicial complex K, that contains all of its faces, is called a subcomplex if

$$\forall e \in L \ and \ \forall f \in K \ (such \ that \ f < e) \Rightarrow f \in L$$

A subcomplex is called proper when $L \neq K$. To gain a better understanding of subcomplexes, let's take two arbitrary subcomplexes of K: K_1 and K_2. Now, let us explore the union and intersection of these two sets. Firstly, union of empty simplices is a simplicial face in an empty space that is often denoted to have $dim = -1$. Secondly, the union of all the faces of an n-dimensional simplex e is itself a simplicial complex which is denoted by $\bar{e}$.

In figure 6, it can be seen that the 3-simplex is composed of 4 triangles. These triangles are the faces of e_3, that for the simplicial complex. They are not random faces but are proper faces of e_3 with dimensions equal to $n-1$. The union of all proper faces of e_n is the boundary of the n-simplex which is a simplicial complex denoted by $\dot{S}$. Alternatively, each face of a simplex that is also an $n-1$-simplex is called a boundary face. The union of the boundary faces will give us the boundary of the whole simplex. Decomposing simplices into their lower dimensional components would be of paramount importance when defining boundary homomorphism in Chapter (3) and introducing the concept of homology.

Proposition 2.1. If subcomplexes K_1 and K_2 of simplicial complex K are in $\mathbb{R}^n$ then
(i) $K_1 \cup K_2$ is a subcomplex of K
(ii) $K_1 \cap K_2$ is a subcomplex of K.

Proof. a) An abstract simplicial complex (discussed in the next section) is a subcomplex of K where every $f \in I$ belongs to the complex. If K_1 and K_2 are subcomplexes then K_1 and K_2 are abstract simplicial complexes. To prove that $K_1 \cup K_2$ is a subcomplex let us show that all faces of $K_1 \cup K_2$ belong to the simplicial complex. Case(1): f is a face of $K_1 \subseteq K_1 \cup K_2 \Rightarrow f \in K$. Case(2): f is a face of $K_2 \subseteq K_1 \cup K_2 \Rightarrow f \in K$. Case(3): f is a face of $K_1 \cap K_2 \subseteq K_1 \cup K_2 \Rightarrow f < K_1$ or $f < K_2$ which brings us back to case(1) and case(2). It is evident that every face of $K_1 \cup K_2$ belongs to simplicial complex K. Thus, $K_1 \cup K_2$ is a subcomplex of simplicial complex K.

b) Let f be a face of $K_1 \cap K_2 \Rightarrow f$ is a face of K_1 and f is a face of K_2. Thus, f is in simplicial complex K. Since any face f of $K_1 \cap K_2$ belongs to a simplicial complex then $K_1 \cap K_2$ is a subcomplex of K. □

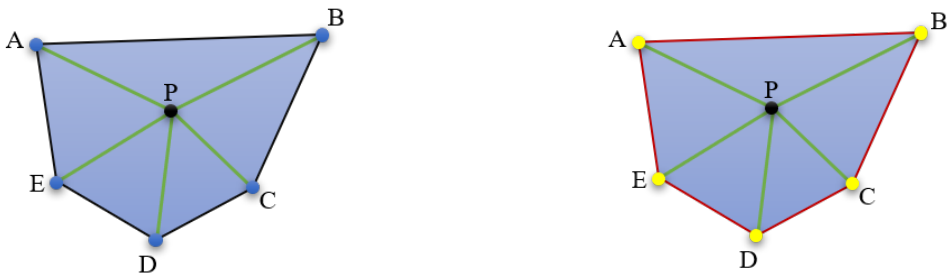
2.2.5 Useful Terms for Simplicial Complex Construction

In this section we will discuss some basic definitions of typical subsets of simplicial complexes. Various scientific fields such as network architecture widely adopted the medium of working with simplicial complexes. The medium makes it easier to analyse local neighbourhoods of elements in the network.

Definition 2.8. Closure: Closure of a subset, $\Phi \subseteq K$, denoted $Cl(\Phi)$) is the smallest subcomplex which contains Φ:

$$Cl(\Phi) = \{\phi \in K : \quad e \leq f \quad for \quad f \in \Phi\}$$

The closure of a simplicial complex can be deduced by walking through the faces of simplices in Φ and filtering out the smallest subcomplex that is common. Definition of the closure is closely related to a concept known as r-skeletion of simplicial complex. For $r \leq 0$, the r-skeleton, denoted by K^r, is a subcomplex of K that is defined as the set of simplices in K with dimension $n \leq r$. For example, 0-skeleton is simply the subset of simplices of K which has dimenison $n \leq 0$. The only such subset is the set of 0-simplices which are the vertices of the simplicial complex. The other two important substructures of complexes are stars and links.



(a) Star

(b) Link

Figure 10: Examples of a Star and a Link of simplicial Complex K

Definition 2.9. Star: Given a simplicial complex K with simplices $\Phi \subseteq K$, the star $st(\Phi)$ is defined as

$$st(\Phi) = \{e \in K : \quad f \leq e \quad for\ some \quad f \in \Phi\}$$

Definition 2.10. Links: Given a simplicial complex K with simplices $\Phi \subseteq K$, the link $lk(\Phi)$ is defined as

$$lk(\Phi) = \{e \in Cl(st(\Phi))) : \quad e \cap \Phi = \emptyset\}$$

The star of a vertex P in figure 10a is the face $ABCDE$ with its interior of the simplicial complex. It is called a closed star. The link of the same simplicial complex is the closure of $st(P)$ or equivalently the intersection between closed and open stars of p. In figure 10b, the $lk(P)$ is the union of line segments AB, BC, CD, DE, EA that are painted in red.

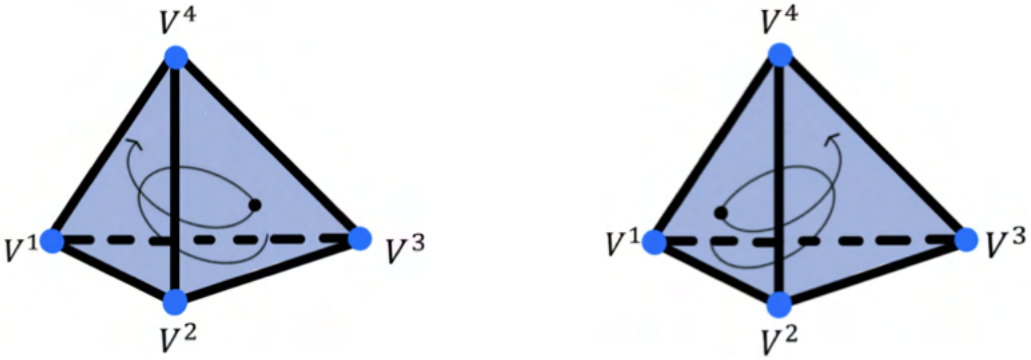
It is important to note that open stars, such as the interior of $ABCDE$, do not always form a simplicial complex. However, closed stars and links result in a proper subcomplex.

3 Homology

3.1 Oriented simplices

By the definition of simplices one can choose different permutations of vertices to construct the simplex. 1-simplex can be defined either $e_1 = (v^0v^1)$ or equivalently $e_1 = (v^1v^0)$. In this section I define the orientation of simplices which is important to further discuss boundary homomorphism in homology groups. Oriented simplices are simplices that have an orientation assigned to them. Thus, the vertices of an oriented simplex have a fixed order.

Definition 3.1. Oriented Simplex: The orientation of a given simplex $e_n = (v^0...v^n)$ is a particular ordering of its vertices. An oriented simplex $\overline{e_n}$ is the n-simplex e_n with a fixed orientation.



(a) Clockwise Orientation

(b) Counterclockwise Orientation

Figure 11: Examples of Two Oppositely Oriented 3-simplices

Clearly any 0-simplex has only one orientation but any other n-simplex has $(n+1)!$ ordered simplices which corresponds to the number of permutations of vertices $v^0, \ldots, v^n$. Two distinct orderings of simplices define the same orientation if and only if they differ by even permutations. In other words, two simplices $\nu_n = (v^0 ... v^n)$ and $\mu_n = (v^{\pi(0)} ... v^{\pi(n)}$ have the same orientation $\nu_n = \mu_n$ if and only if π is even. For example, $\overline{e_n}$ when $n > 0$ has only two distinct orientations which are denoted by $\overline{e_n}$ and $-\overline{e_n}$. Although, the 2-simplex (v^0, v^1, v^2) can have $(2+1)! = 6$ distinct orderings, orientation of the simplex can be described either by $(v^0 v^1 v^2)$ or $(v^1 v^0 v^2)$. Six orientations fall into two categories: clockwise and counterclockwise orientations.

The figures 11a and 11b, drawn above represent two different orientations that a 3-simplex can have. In 11a the clockwise oriented 3-simplex can be written out as $\overline{e_3} = (v^3, v^2, v^1, v^4)$. Contrarily, 11b describes counterclockwise orientation of the 3-simplex $-\overline{e_3} = (v^3, v^1, v^2, v^4)$. The same way, oriented simplicial complexes, are simplicial complexes where every simplex has its own unique orientation. When discussing chain complexes and boundary homomorphsim, we will give more rigorous definition of oriented simplicical complexes. More information on the orientation can be found [1]. Later, I will introduce an appropriate convention that will allow us to work with boundaries of simplices regardless of having elements with conflicting orientation within the same complex.

3.2 Chain Groups Boundary Homomorphism

3.2.1 Groups

Since 19-th centuary mathematicians realized that a lot of problems require fundamentally similar/symmetric solutions which can be generalized into groups of problems. Within those groups, one can derive various subsets such as graded/ungraded rings, fields that are defined under certain operations. Creating this environment may simplify the initial question by breaking it down into smaller problems.

Definition 3.2. Group
A set G is a group if its elements satisfy the conditions listed below
(1) G is closed under binary operation
(2) All elements of G have their inverse in G
(3) Operations defined under G are associative

(4) There is an identity element in G

A particular type of groups, called abelian groups extend the definition of groups by imposing one more addition condition on the set. Abelian groups, modeled on integers $\mathbb{Z}$ are groups that are powered by commutative property. Based on the 4 conditions mentioned in 3.2.1, we exclude many examples such as the group of symmetries (i.e. if $a,\ b \in G$ we cannot assume that $a + b = b + a$). Thus, groups are called abelian if $a * b = b * a$ for all $a, b \in G$.

Definition 3.3. Abelian Groups
Groups are called abelian if the underlying operation of the group is commutative.

Basic understanding of groups would allow us to talk about chain complexes which are created upon free abelian groups. However, more information can be found in [2].

Chain groups are commutative free abelian groups that append an integer coefficient to the elements that form an n-simplex. Chain groups appear in every dimension of simplicial complex K.

Definition 3.4. m-chain
Let K be the oriented simplicial complex composed of oriented m-simplices denoted by $\theta^1, ..., \theta^n$ where n is the number of m-simplices. The m-chain group $C(k)$ is the free abelian group on the set $\theta^1, ..., \theta^n$. Elements of the set are called m-chains and can be written as $\sum_m \lambda_m * \theta_m$ where $\lambda_m \in \mathbb{Z}$ and $\theta_m \in K$

Chain groups carry natural abelian group structure. If $c_m, c_n \in C(K)$ then we define $(c_m + c_n)(\theta) = (c_m(\theta) + c_n(\theta))$ The concept of homology aims to identify the relationship between n and $n-1$ dimensional components of simplices. Identifying homology groups allows for the study of the relationship of two consecutive chain groups. The idea of boundary homomorphism allows decomposition of these higher dimensional chain groups into its subsequent lower dimensional versions.

In examples above, the n dimensional simplex $(v_0...v_n)$ is composed of various $n-1$-dimensional simplices. Similarly, for the boundary of an n-dimenional simplex is a $(n-1)$ dimensional chain. To compute the homology groups of a simplex, one needs to identify the chain groups $C(k)$. However, one is interested in simplicial complex K where particular orientations of v_1 and v_2 may not be coherently consistent with the orientation of the rest of the complex. Thus, faces of the simplex cannot be simply added to get the boundary. The reason is that one should account for the orientation of the different components of the complex. If the line segment $v_1v_2 \notin K$ then it will not result in a 1-chain v_1v_2 when computing the group representing boundaries. The algebraically equivalent orientation of v_1v_2 is denoted by $-v_2v_1$. Thus the boundary of the 2-simplex can be written as $d(2 - simplex) = (v_0v_1 - v_2v_1 + v_2v_0)$.

Last paragraph motivates the convention, according to which the boundary of oriented simplex is the $(n-1)$ chain denoted by $\sum_i (-1)^i[v_0, ...\hat{v_i}...v_n]$ where the hat on v_i means that the particular vertex is missing. The -1 in the beginning controls for the overall orientation of the simplex depending on odd or even permutations of its elements. Below, I give the rigorous definition of boundary homomorphism based on the convention and the discussion of the boundary operator above.

Definition 3.5. Boundary Homomorphism
The boundary operator $d_p : C_k \longrightarrow C_{k-1}$ is a homomorphism that is defined on chain c

as

$$d_p(\overline{e_p}) = \sum_i (-1)^i \, \overline{e_p} \mid [v_0, ...\hat{v_i}...v_p]$$

where $\hat{v_i}$ is the omitted simplex and $\overline{e_p} \mid [v_0, ...\hat{v_i}...v_p]$ denotes the simplex $\overline{e_p}$ with vertices $(v_0,\ v_1, \cdots, v_p)$.

It is important to see that the right side of the equation describes $k-1$-chain group. The reason is that every element of the summation is mapped to $n-1$-dimensional simplex. Alternatively, the boundary homomorphism $d_p : C_k \to C_{k-1}$ can be described as $d_p(\sum \lambda_m \overline{e_p}) = \sum \lambda_m d_p(\overline{e_p})$ where $m \in [0, n]$. Such a representation is very similar to operations performed in linear algebra. Elements of the groups can be imagined as vectors. However, contrary to linear algebraic mathematics coefficient, λ must be an integer.

(a) Each edge in the figure on the right side has its own boundaries described by a pair of vertices. I.e., $d(a) = V^2 - V^1$, $d(b) = V^3 - V^2$, $d(c) = V^1 - V^3$ and $d(d) = V^1 - V^3$). looking at definition 3.2.4 The boundary homomorphsim $d : C_1 \to C_2$ can be written as $\mathbf{d}(\lambda_1 a + \lambda_2 b + \lambda_3 c + \lambda_4 d) = \lambda_1 \mathbf{d}(a) + \lambda_2 \mathbf{d}(b) + \lambda_3 \mathbf{d}(c) + \lambda_4 \mathbf{d}(d) = \lambda_1(V^2 - V^1) + \lambda_2(V^3 - V^2) + \lambda_3(V^1 - V^3) + \lambda_4(V^1 - V^3) = V^1(\lambda_3 + \lambda_4 - \lambda_1) + V^2(\lambda_1 - \lambda_2) + V^3(\lambda_2 - \lambda_3 - \lambda_4)$ It is also interesting to note that $d(a + b + c) = d(a) + d(b) + d(c) = (V^2 - V^1) + (V^3 - V^2) + (V^1 - V^3) = 0$ and $d(a + b + d) = d(a) + d(b) + d(d) = (V^2 - V^1) + (V^3 - V^2) + (V^1 - V^3) = 0$. This finding is true for all cycles. When discussing homology groups we will generalize cycles as closures whose boundaries are equal to zero.

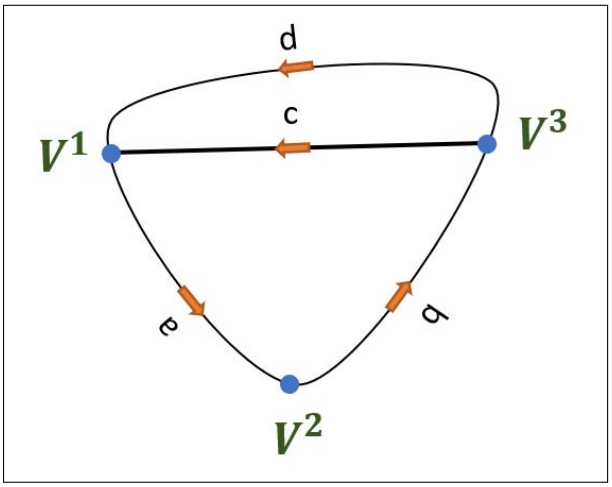


Figure 12: Simple Example of Boundary Calculations

Definition 3.6. Augmentation
The augmentation $\eta :\ C_0(K) \to \mathbf{Z}$ is the homomorphism defined such that $\eta(\sum \lambda_m \overline{e_p}) = \sum \lambda_m$

The sequence of such group homomorphisms can be described as

$$... \to C_2(K) \xrightarrow{d_2} C_1(K) \xrightarrow{d_1} C_0(K) \xrightarrow{\eta} \mathbf{Z} \to 0$$

which is called augmented chain complex of K. There is a clear pattern which suggests that the result of boundary operator acting on a boundary operator is trivial. It means that $d^2 = d \circ d = d_p \circ d_{p-1} \equiv 0$

Theorem 2. The sequence $C_p(K) \xrightarrow{d_n} C_{p-1}(K) \xrightarrow{d_{n-1}} C_{p-2}(K)$ is trivial, meaning the composition is equal to zero.

Proof. We have that

$$d_p(\overline{e_p}) = \sum_i (-1)^i \, \overline{e_p} \mid [v_0, ...\hat{v_i}...v_p]$$

then $d_{p-1}d_p(\overline{e_p}) = \sum_i (-1)^i\, d_{p-1}(v_0, ...\hat{v_i}...v_p) = \sum_{i<j}(-1)^i(-1)^{j-1}\overline{e_p} \mid [v_0, ...\hat{v_i}...\hat{v_j}...v_p] +$
$+ \sum_{i>j}(-1)^i(-1)^j\overline{e_p} \mid [v_0, ...\hat{v_j}...\hat{v_i}...v_p] = -\sum_{i<j}(-1)^i(-1)^j\overline{e_p} \mid [v_0, ...\hat{v_i}...\hat{v_j}...v_p] +$
$+ \sum_{i>j}(-1)^i(-1)^j\overline{e_p} \mid [v_0, ...\hat{v_j}...\hat{v_i}...v_p] \equiv 0$ □

Finally, singular chain complex of an arbitrary topological space is a chain complex where d_p is the boundary map.

3.3 Homology Groups

Definition 3.7. *p*-**cycles**
$Ker\ d_p$ are the p-cycles which are denoted by Z_p: $Z_p(K) = ker(d_p : C_p \to C_{p-1})$.

Definition 3.8. *p*-**boundaries**
$Im\ d_{p+1}$ are the p-boundaries which denoted by B_p :, $B_p(K) = Im(d_{p+1} : C_{P+1} \to C_p)$

Theorem 3.2.1 suggests that $Im(d_{p+1}) \subset Ker(d_p) \subset C_p(K)$. This, allows to define p-th homology group of the augmented chain complex as a quotientent group $H_p(K) = Z_p(K)/B_p(K)$. Homology group H_p is a finitely generated abelian group. The rank of these groups measures the number of p-dimensional cycles in a given simplicial complex K. More generally p-th homology group can be thought as a group of p-cycles of K that are not boundaries.

0-th homology group $H_0(K)$, where $p = 0$ describes the connected components of a simplicial complex. In this case $Z_0(K)$ are cycles that are a linear combination of 0-simplices in K. Moreover, boundaries $B_p(K)$ are a linear combinations of 0-simplices that are in the same connected component. As a consequence, cycles modulo boundaries are the free abelian group of connected components.

First homology group, where $p = 1$ examines cycles which are the linear combination of closed surfaces composed of 1-simplices. Boundaries on the other hand are a linear combination of cycles in K that leap 2-simplices. Correspondingly, when $C_2(K) \xrightarrow{d_2} C_1(K) \xrightarrow{d_1} C_0(K)$ then $H_1(K) = Ker\ d_1/Im\ d_2 = Z_1/B_1$ describes the free abelian group of independent loops. In the example of torus discussed in the introduction we encounter 2 different types of holes that result in $H_1(T) = \mathbb{Z} \oplus \mathbb{Z}$. Describing cycles and boundaries is much more complicated in cases of $p \geq 2$. The theorem below, on the sequence of homology groups, makes it easier to compute higher dimensional homology groups.

Theorem 3. If $K_1, ..., K_n \in K$ are connected components of simplicial complex K then $H_m(K) \cong H_m(K_1) \oplus ... \oplus H_m(K_p)$

Proof. I want to show that the decomposition of the homology groups holds for both the p-cycles and p-boundaries. Let $X_m(K)$ represent the group of m-chains of K. $X_n(K_i)$ is a subgroup of $X_n(K)$. Furthermore it can be written as $X_m(K) = X_m(K_1) \oplus ... \oplus X_m(K_p)$. Let the image of f on the m-chains be $B_m(K_i) = f(X_{m+1}(K_i)$ then $B_m(K) = B_m(K_1) \oplus ... \oplus B_m(K_p)$. Let $Z_m(K_i) = ker\ f \cap\ X_m(K_i)$ then $Z_m(K_i) = Z_m(K_1) \oplus ... \oplus Z_m(K_P)$ Thus,

$$\frac{Z_m(K)}{B_m(K)} = \frac{Z_m(K_1)}{B_m(K_1)} \oplus ... \oplus \frac{Z_m(K_p)}{B_m(K_p)} \iff H_m(K) \cong H_m(K_1) \oplus ... \oplus H_m(K_p)$$

□

In the example below I present two distinct ways of computing homology groups of a $S^1 \cup S^1$ (figure eight) and a S^2 (a sphere.) Unlike the first one, the second example uses reduction algorithm to identify generators (the basis of cycles) using boundary operations. Note that curves are topologically equivalent to straight lines. In case of a circle, one can include three 0-dimensional simplices on the contour and also add 1-dimensional line segments. As a result, under affine transformations in $\mathbb{R}^n$ the circle can be replaced by its topologically homeomorphic triangle.

3.3.1 Examples of Computing Homology Groups

In figure 13a, one can see that in the sequence of $0 \to C_1 \xrightarrow{d_2} C_0 \xrightarrow{d_1} 0$: C_0 is generated by a single cell of vertex v. C_1 is generated by edges of e_1 and e_2. The sequence can be rewritten such as $0 \to \mathbf{Z} \oplus \mathbf{Z} \xrightarrow{d_2} \mathbf{Z} \xrightarrow{d_1} 0$. Additionally, $Im(d_1) = Im(d_2) = d(e_1) = d(e_2) = v - v = 0$, while $Ker(d_0) = C_0 = \mathbf{Z}$ and $Ker(d_1) = C_1 = \mathbf{Z} \oplus \mathbf{Z}$. Thus,

$$H_0(S_1 \; v \; S_2) = \frac{Ker(d_0)}{Im(d_1)} = \mathbf{Z} \quad H_1(S_1 \; v \; S_2) = \frac{Ker(d_1)}{Im(d_2)} = \mathbf{Z} \oplus \mathbf{Z}$$

and $H_p(S_1 \; v \; S_2) = 0 \; \forall p \geq 2$ as there are no other higher dimensional generators.

For the two dimensional sphere in 13b first I triangulate the object to its equivalent tetrahedron (Stock's Theorem). In particular, the 3-simplex is composed of

vertices: $\{V^1\}, \{V^2\}, \{V^3\}, \{V^4\}$

line segments $\{V^1, V^2\}, \{V^1, V^3\}, \{V^1, V^4\}, \{V^2, V^3\}, \{V^3, V^4\}, \{V^4, V^2\}$

triangles: $\{V^1, V^2, V^4\}, \{V^1, V^2, V^3\}, \{V^1, V^3, V^4\}, \{V^2, V^3, V^4\}$.

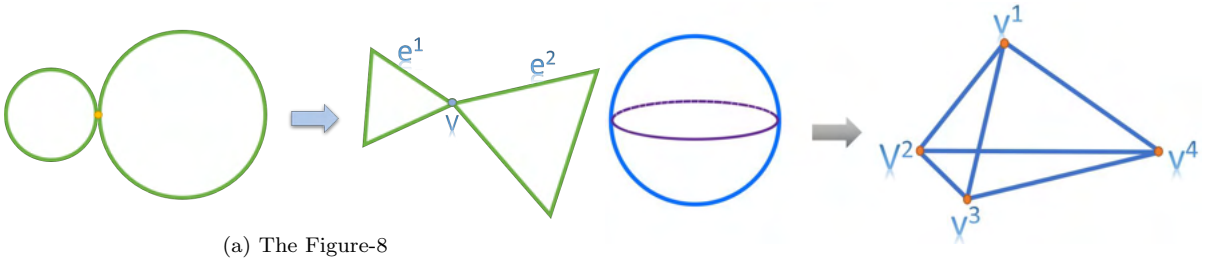
Boundary operators of d_1 can be represented by forming a matrix that describes pairwise vertices of the 1-simplices which are in C_1.

$$d_1 = \begin{pmatrix} 1 & 1 & 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 & 1 \\ 0 & 1 & 0 & 1 & 1 & 0 \\ 1 & 0 & 1 & 0 & 1 & 1 \end{pmatrix}$$

The matrix has 4 basis which are the vertices and the $rank(d_1) = 3$. It follows that 0-th homology group $H_0(S^2) = \mathbf{Z}$. To identify first homology group of a two dimensional sphere, I decompose the triangles into 1-simplices to identify independent generators.

$$d_2 = \begin{pmatrix} 1 & 1 & 0 & 0 \\ 0 & 1 & 1 & 0 \\ 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 1 \\ 1 & 0 & 0 & 1 \end{pmatrix}$$

The rank of the above mentioned matrix is three and the rank kernel of d_1 is equal to three as well. Consequently, $H_1(S^2) = 0$. In addition, S^2 composed of four triangles lacks d_3 generators despite the 2-cycles by itself. As a result $H_2(S^2) = \mathbf{Z}$. All higher dimensional ($p \geq 3$) homology groups of a 2-dimensional sphere are empty.



(a) The Figure-8

(b) The Sphere

Figure 13: Distinct Ways of Computing Homology Groups

4 Persistent Homology

This chapter is devoted to developing the notion of persistent homology which is based on computing homology groups over inductive set of simplicial complexes. The construction called a filtered complex is an increasing sequence of simplicial complexes such that: $E^0 \subset E^1 \subset E^2 \subset E^3 ....$ In the example given below, the filtration process takes six different periods starting with three 0-simplices (at $t = 1$) that results into K built from three 2-simplices and two 1-simplices (at $t = 4$.) More rigorously such an inductive system is $E_0^0 \subset E_0^1 \subset E_0^2 \subset ... \subset E_0^5$ where the superscript indicates the time during the filtration process and the subscript suggests the dimension of simplices we are dealing with.

Definition 4.1. The degree of a simplex l_i is denoted by $deg(l_i)$ is equal to the time when l_i enters the filtration process.

The end goal of persistent homology as a tool is to identify higher dimensional homological features that persist over the duration of the filtration process. Here, I create simplicial complex by either adding n-dimensional n-simplices to pre-existent vertices or by introducing previously non-existent components. In the nested sequence of complexes that are triangulated by definition, I am interested in boundaries and cycles that "emerge" or get "destroyed."

4.1 Theoretical Background

Besides identifying homology groups in each period, by computing persistent homology, one can learn how long particular homology groups persisted over the filtration period. Instead of computing cycles mod boundaries, PH requires computing cycles mod boundaries emerging in the future dimensions of filtration such that $H_n^{i,p} = Z_n^i / B_n^{i+p} \cap Z_n^i$. It is important to note the boundaries should be restricted to the components that only existed in the point of outlook regardless of the new structures that came up as a result of the filtration. Thus, cycles mod boundaries at time $i + n$ are intersected with cycles at time $t = i$.

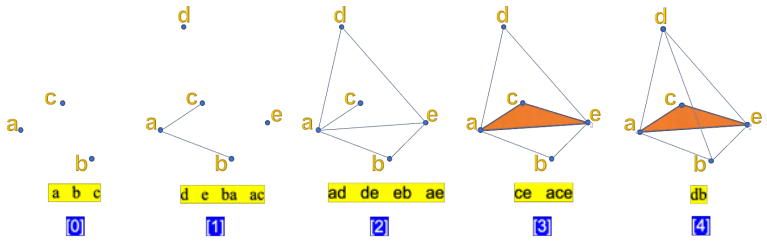


Figure 14: Filtered Complex

Persistent homology's objective is to compute the i-th homology with the coefficient in field F. Inclusion maps between spaces $E^0 \subset E^1 \subset E^2 \subset E^3 \subset E^4$ indicate the maps between homology groups: $H_i(E^0) \to H_i(E^1) \to H_i(E^2) \to H_i(E^3) \to H_i(E^4)$. Thus, homology groups are just vector spaces equipped with linear maps between them. In the diagram below I switch boundary maps to illustrate the vertical direction. The filtration process which is the second dimension of interest is depicted on the horizontal direction going from the left to the right of the page.

$$\begin{array}{ccccccc}
 & \downarrow{\scriptstyle d_3} & & \downarrow{\scriptstyle d_3} & & \downarrow{\scriptstyle d_3} & \\
C_2^0 & \xrightarrow{f^0} & C_2^1 & \xrightarrow{f^1} & C_2^2 & \xrightarrow{f^2} & \cdots \\
 & \downarrow{\scriptstyle d_2} & & \downarrow{\scriptstyle d_2} & & \downarrow{\scriptstyle d_2} & \\
C_1^0 & \xrightarrow{f^0} & C_1^1 & \xrightarrow{f^1} & C_1^2 & \xrightarrow{f^2} & \cdots \\
 & \downarrow{\scriptstyle d_1} & & \downarrow{\scriptstyle d_1} & & \downarrow{\scriptstyle d_1} & \\
C_0^0 & \xrightarrow{f^0} & C_0^1 & \xrightarrow{f^1} & C_0^2 & \xrightarrow{f^2} & \cdots \\
 & \downarrow{\scriptstyle d_0} & & \downarrow{\scriptstyle d_0} & & \downarrow{\scriptstyle d_0} & \\
 & 0 & & 0 & & 0 &
\end{array}$$

Definition 4.2. The persistence module M is a group of graded modules over a graded ring with homomorphisms $\theta^i : M^i \to M^{i+1}$.

Zomorodian et al. and Carlson et al. show that there is an equivalence relation between persistent and $F[t]$ modules. Let the elements of the the set of $F[t]$ module be $h_i(E^0), h_i(E^1), h_i(E^2), h_i(E^3), h_i(E^4)$ then the list of these elements is in the direct sum of vector spaces $H_i(E^0) \oplus H_i(E^1) \oplus H_i(E^2) \oplus H_i(E^3) \oplus H_i(E^4)$. Moreover, the action required by $F[t]$ module can be given by $t * (h^0, h^1, h^2, \ldots) = (0, g^0(h^0), g^1(h^1), \ldots)$. As a result variable t acting on h^n allows to encode the maps between the vector spaces. Furthermore, finitely generated graded $F[t]$ module can be decomposed into

$$H_k = \left(\bigoplus_{i=1}^{n} \sum^{\alpha_i} F[t] \right) \oplus \left(\bigoplus_{i=1}^{n} \sum^{\gamma_j} F[t]/(t^{k_j}) \right)$$

The grading part in the equation is signified by the $\sum^{\alpha_i}$ and the $\sum^{\gamma_j}$ which allow for the shifts of subsequent spaces. Torsion free part correspond to the first component in the equation that goes forever in persistent barcode (barcodes are introduced in the next section.) However, the second part of the equation describes the finite bars where cycles that started at some point during the filtration and were destroyed further in the filtration process.

4.2 Computing PH

In this section, I show that persistence algorithm can be thought of as Gaussian elimination. First, I list generators and relators across rows and columns of matrices. Let us write out M_1 where $\{d, e, c, b, a\}$ are the generators of H_1 and $\{ba, ac, ad, de, eb, ce, db\}$ are the relators of H_1

$$M_1 = \left[\begin{array}{c|cccccccc} & ba & ac & ad & de & eb & ae & ce & db \\ \hline d & 0 & 0 & t & t & 0 & 0 & 0 & t^3 \\ e & 0 & 0 & 0 & t & t & t & t^2 & 0 \\ c & 0 & t & 0 & 0 & 0 & 0 & t^3 & 0 \\ b & t & 0 & 0 & 0 & t^2 & 0 & 0 & t^2 \\ a & t & t & t^2 & 0 & 0 & t^2 & 0 & 0 \end{array}\right]$$

M is a graded module over a polynomial ring. We call polynomials homogeneous if all terms have the same degree. For example, $(t)a + (t)b - (1)ab \in C^1$ is a homogeneous polynomial where all terms are from the same filtration period. Grading the module allows us to represent the filtration period through degrees of polynomials $< t >$. Let us define $\{l_i\}$ to be homogeneous basis for C_k and $\{\hat{l}_i\}$ be homogeneous basis for C_{k-1}. Then

$$deg\ (\hat{l}_i)\ +\ deg\ M_l(i,j)\ =\ deg\ (l_i)$$

where $M_l(i,j)$ is the element of (l,j) entry of the matrix. Looking at the matrix, $M_l(a, ad) = t^2$ because $deg\ (a) = 0$ and $deg\ (ad) = 2$. It ends up that $deg\ M_l(a, ad) = 2-0 = 2$. Now we can conduct column operations to find kernel and image of the matrix. Below we use the *Basis Change* lemma that can be found in the original paper [3].

Step1 Reduce M_l^i to obtain Z_k^i

$$\bar{M_1} = \left[\begin{array}{c|cccccccc} & db & ce & ac & ba & z_1 & z_2 & z_3 & z_4 \\ \hline d & t^3 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ e & 0 & t^2 & 0 & 0 & 0 & 0 & 0 & 0 \\ c & 0 & t^3 & t & 0 & 0 & 0 & 0 & 0 \\ b & t^2 & 0 & 0 & t & 0 & 0 & 0 & 0 \\ a & 0 & 0 & t & t & 0 & 0 & 0 & 0 \end{array}\right]$$

where z_1, z_2, and z_3 are homogeneous basis such that $z_1 = ad - de - ae$, $z_2 = de - ad - ae$, $z_3 = eb - (t)ba - ae$, and $z_4 = ae - de - ad$. Now I can compute $B_0 = ColSpace\ M_1 =< t^3d + t^2b,\ t^2e + t^3c,\ tc + ta,\ tb + ta >$ and $Z_1 = nullSpace\ M_1 =< ad - de - ae,\ de - ad - ae,\ eb - (t)ba - ae,\ ae - de - ad. >$ From here on $H_0 = Z_1/B_0$ can be easily computed but for $H_1 = Z_1/B_1$ I need column space of M_2.

Step2 Reduce M_{l+1}^{i+p} to obtain B_k^{i+p}

$$M_2 = \left[\begin{array}{c|c} & ace \\ \hline db & 0 \\ ce & 1 \\ ae & t \\ eb & 0 \\ de & 0 \\ ad & 0 \\ ac & t^2 \\ ba & 0 \end{array}\right] \Rightarrow \bar{M_2} \left[\begin{array}{c|c} & ace \\ \hline ba & 0 \\ ad & 0 \\ de & 0 \\ eb & 0 \\ db & 0 \\ ac & 0 \\ ae & 0 \\ z & 1 \end{array}\right]$$

where $z = ce - (t)ae - t^2(ac)$. It is evident that $B_1 =< 1 * z >=< ce - (t)ae - t^2(ac) >$ while $Ker(d_2) = 0$.

Step 3 Compute any $H_n^{i,p}$. In this example $H_1^{0,b}$ and $H_1^{0,b}$ are both 0 because in filtration we only get cycles starting at $t = 2$. To compute $H_1^{2,p}$ first I need to compute $= Z_1^2/B_1^{2+p} \cap Z_1^2$ which is equal to $Z/2Z$ as $deg(z) = 2$. The cycle abe persists for 2 periods when $p = 0,\ 1,\ 2$. Thus, $H_1^{2,p} = (B_1^{2+p} \cap Z_1^2) = Z_2$ for $p = 3$. Using this algorithm I can compute the rest of $H_n^{i,p}$ which would result in various $P-intervals$ that can be demonstrated on a barcode. In the original paper, one may find a pseudocode for computing p-intervals [3]. When the ground ring is a field one can use the correspondence to compute the infinite barcode.

4.3 Sparse Computation of PH

Let us take a look at homology groups that arise during filtration process at $t = 0$ and get destroyed at $t = 2$. The objective is to compute $H_0^{0,2} = Z_0^0/(B_0^2 \cap Z_0^0)$ when $Z_0^0 = a, b, c$, and $B_0^2 = ac, ab, be, da, de$. Intersection of B_0^2 and Z_0^0 emphasizes on the cycles that came from $t = 0$ and were destroyed 2 units later. This allows to restrict the sample into pairs of cycles that were created at a particular period of filtration and destroyed further in the later periods. Cycles such as de, da, be are moded out because their components d, e arise at $t = 1$. $a + c = 0$ and $a + b = 0$ suggests that $H_0^{0,2}$ results in two components. The vertices a, b, c which are cycles themselves are destroyed when line segments ac and ab are introduced. Such a representation of cycles and boundaries allows us to observe cycles that are created at $i = 0$ and vanished p-units later.

It is intuitive that introduction of new simplices in the filtration can either create new cycles or destroy them. Let *positive simplex* signify the components whose entrance creates a cycle. Let a *negative simplex* represent components whose introduction into the filtration destroys a cycle. In computing homology, generators are denoted as cycles and relators as boundary mappings. In persistent homology, generators are the positive simplices whose entrance results in a creation of cycles. On the other hand, relators arise when there is a negative simplex that destroys the generators.

Marking positive and negative simplices through the whole filtration process would allow us to learn how long particular simplices "persisted" through filtration. The figure below shows the sequence of components through the filtration process with the corresponding + / − signs meaning *positive simplex / negative simplex*. The first yellow row represents filtration process corresponding to the entrance of simplices of the second row. The last row representing filtration value counts the number of processes happening during the filtration. For example, a, b, and c are positive 0-simplices because these vertices are cycles by themselves. Later on, in the the first filtration process there are multiple positive and negative simplexs. Arbitrarily, first I list relators and then generators within the same period. Thus, negative simplices ba and ac come before e, d positive simplices on the same filtration level.

0	0	0	1	1	1	1	2	2	2	2	3	3	4
a	b	c	ba	ac	e	d	ad	de	eb	ae	ce	ace	db
+	+	+	-	-	+	+	-	-	-	+	+	-	+
0	1	2	3	4	5	6	7	8	9	10	11	12	13

Figure 15: Sparse Data Structure Storage

Ultimately I need to take these cycles and mod them out with the boundaries in future periods according to persistence homology principles. For example ba destroys b and a cycles but during the pairing process I will arbitrarily choose the latter one: a to pair it with. Looking at the last row of the table, I pair:

$$\{0,3\}\ \{1,9\}\ \{2,4\}\ \{5,8\}.\ \{11,12\}.$$

In the graph 16 one of the axis describes index (i) and the other one persistence (p). Here I connect the pairs identified above in order to see how long it took until the generated cycles were destroyed. The diagram 16 called *barcode* is a sequence of horizontal line segments that represents the orderings of homology generators which persisted through the filtration. The x-axis corresponds to the filtration value while the y-axis does not have a meaning. Software functions when generating barcodes place the longer bars closer to number line and shorter ones stacked above the long ones. The longer the intervals, the more significant are the high dimensional features. Shorter intervals are interpreted as small shocks which do not have a major effect.

Definition 4.3. The P-intervals are ordered pair of i, j where $0 \leq i < j$ and $j \in \mathbb{Z}+\{\infty\}$

The P-intervals corresponding to the pairs above are

$$[a^0, ba^1]\ [b^0, eb^2]\ [c^0, ac^1]\ [e^1, de^2]\ [ce^3, ace^3]$$

The right triangles resemble entrance of a cycle where the dotted line points out the filtration value of when its destroyed. Now, one can compute *betti numbers* which indicates the number of n-dimensional generators at a particular level of filtration denoted by $B_n^{i,j}$. For example $B_0^{2,4} = 1$ because the diagram shows that there is one 0-dimensional cycle that enters at filtration value 2 and is destroyed 2 steps later at filtration value equal to 4. Going over the mechanics of persistent homology is good way for visualizing the process underlying PH.

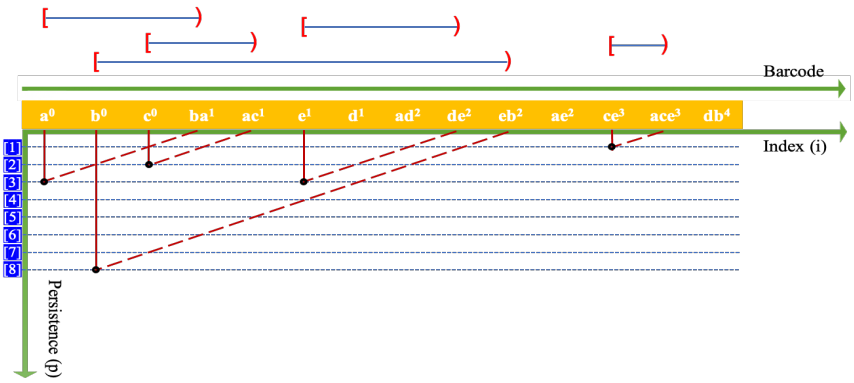


Figure 16: Barcode

5 Application on the US Schools Data

Researchers from various fields such as medicine, biology, image processing, and Researchers from various fields such as medicine, biology, image processing, and statistics adopted PH as a method for data processing. The input type of the data needs to be converted to a point cloud for computing the PH algorithm. The points on the metric space can represent an ally of data structures such as weighted networks, functions, or multidimensional spreadsheet data. [6] Muthu Alagappan used data on NBA player performance to cluster the players into multiple beans. The paper revealed that there are actually 13 new playing positions/styles that were unknown before. Another paper, [7] used high dimensional breast cancer data from participants and identified a new type of cancer. The new type was not lethal and was the result of 7.5% of all kinds of breast cancers. Furthermore, [8] uses brain network data (locations of ROI) and finds that people who have ADHD tend to have more persistent brain network connections. My aim in this paper, is to explore California private and public school data in order to find regions that historically have suffered the most from a lack of educational resources. Also, I aim to run different clustering algorithms to unveil whether there are any hidden topological features for these schools.

5.1 Filtration

There are various methods of filtration that one can use to compute the persistent module. The three main ones are called: Cech-complex filtration (runtime: $2^{O(N)}$) , Alpha-complex filtration (runtime: $N^{O([d/2])}$) (which are based on Nerve theorem), and lazy witness filtration (runtime: $(2^{O(|l|)}$). Lazy witness complex is used when the datasets are big big enough and especially when it encompasses curves and surfaces in euclidean space [4]. There are many other filtration methods. However, they are derivatives of these three methods mentioed above.

Lazy witness complex uses samples from data either randomly or by maxmin point selection algorithm to compute the persistent module. In case of maxmin algorithm, the system chooses initial random point cloud data. The rest of the point cloud enters the filtration by maximizing the distance between the initial and randomly chosen vertices. This can be useful when one works with big data. Re-sampling of the set assures that the entrance of few components will not make sensitive changes to the outcome. Lazy witness complex filtration method does not exist in the statistical packages I use. However, a great deal of information can be found [6].

In the analysis of school data we use Vietoris-rips filtration which is a derivative of cech filtration. Vietoris-rips complex $R(x, \epsilon)$ is composed of vertices $X = (x_1, \ldots, x_n) \in \mathbb{R}^n$ and epsilon balls with diameter d. When d the diameter of the epsilon balls is increasing new components arise in the complex with a condition: include a and b in the $R(X, \epsilon)$ when $d(a, b) = \epsilon$ such that $a,\ b \in X$. The picture in 18 below shows that as ϵ grew larger while two of the balls intersected. As a result an edge needs to be drawn between the two points. In figure 17, I create a randomly generated point cloud data in order to demonstrate the filtration process. The point cloud is drawn within the geo-boundaries of California. The generated data going through different steps of the filtration are depicted in figure 19. On the left side of the figure the epsilon balls are drawn when ϵ is 0.3 while on the right side I draw their corresponding 0 and 1 dimensional simplices.

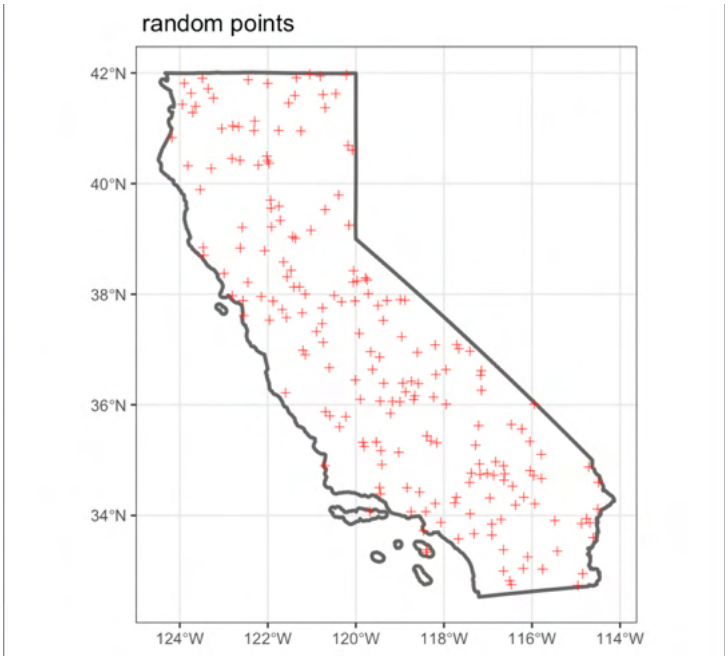


Figure 17: Randomly Generated Point Cloud Within California Boundaries

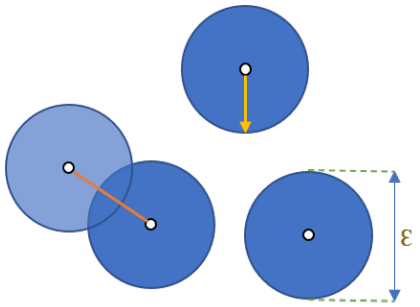


Figure 18: Example of Vietoris-Rips Filtration

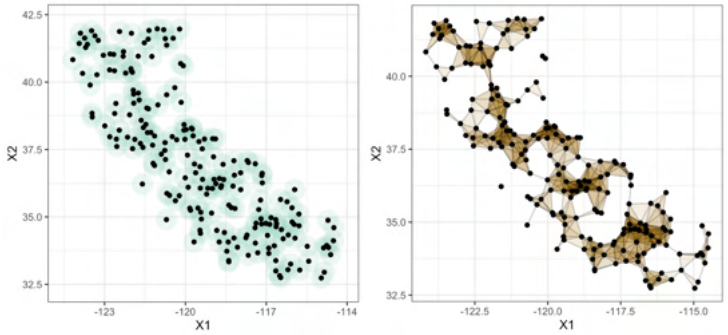
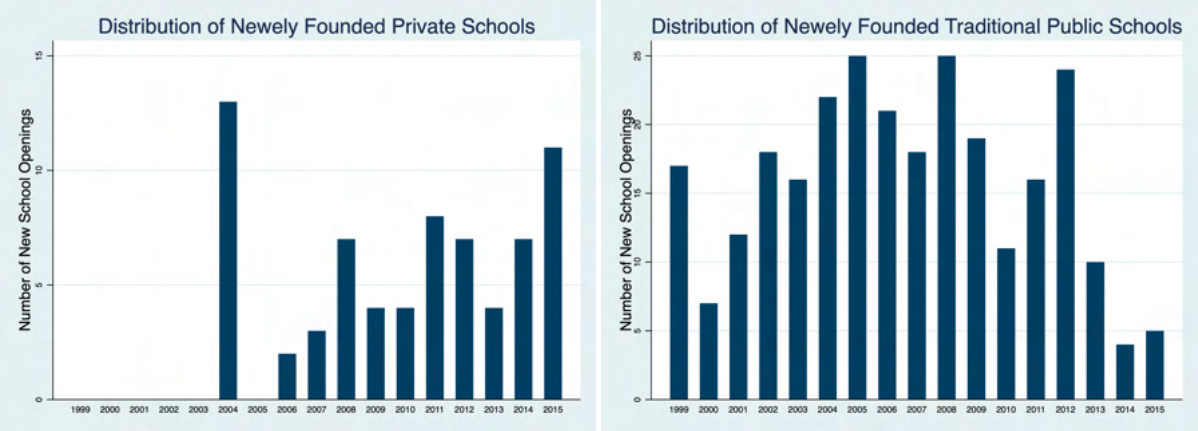


Figure 19: Filtration Process on a Randomly Generated Point Cloud Within California Boundaries



(a) Private Schools

(b) Public Schools

Figure 20: Number of Schools Per Year

5.2 Data & Software Used

The main administrative data sources I use in this study is NCES (National Center for Education Statistics.) NCES uses CCD (Common Core Data) and PSS (Private School Universe Survey) to provide locations, population count and other basic information about the private and public schools. The active school sample I use in my analysis is composed of 219 private high schools and 1370 traditional public schools. The high schools in my sample are defined to be schools serving students from grades 9 to 12.

Most of the data analysis is done in statistical software R. There are various topological data analysis packages, however, we will concentrate on the main ones "TDA" and "TDAstats". These packages run on libraries "GUDHI" and "Dionysus" that are written in C++.

When computing persistent homology the output of a fileration proccess has three main components: betti number, R_0, and R_1. R_0 represents the filtration value at which the homological cycle appears and R_1 is when the homological cycle in topology is destroyed. The betti number refers to the dimension of the particular feature. The relative dominance of the feature is defined by $\frac{R_1-R_0}{L}$ where L is the filtration period when the complex becomes a single connected component. Smaller relative dominance means that particular features are more of shocks rather than persistent features.

5.3 Results

I use point cloud data of public and private schools to run persistent homology algorithms via Vietros Rips filtration. First, I employ the public and private school data to identify the most persistent generators. These generators are going to be the most persistent cycles on the map. There are two graphs we aim to generate, persistence diagrams and the representative loop on the point cloud data.

The 21(a) is called persistence diagram where the x-axis represent the births and the y-axis represent the deaths of the cycles. The black dots represent the 1-dimensional cycles and the triangles mark the 2-dimensional loops. The triangles that are closer to the identity line are considered to be shocks, shortly after the birth they get destroyed. The shaded region is the 90% confidence band for our expected elements on the diagram. More information about the algorithm used can be found in [9]. The most persistent 2-cycle

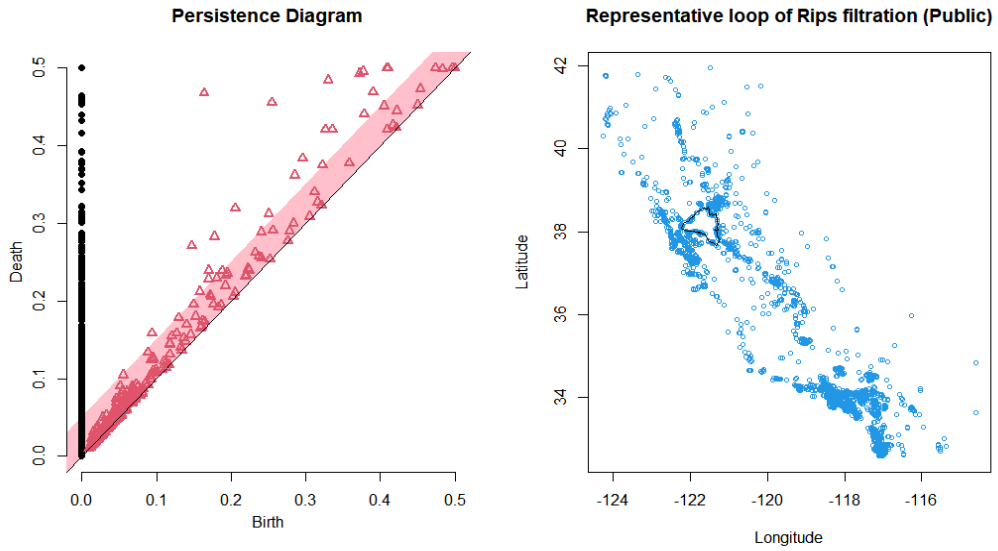


Figure 21: Public Schools 1850-2020

generator was born when $d = 0.2$ and was knocked down at $d = 0.49$. The right side of 21 represents the loop of the most persistent generator. One can infer the circled region to be secluded from wide access of schools. The area represents San Joaquin County which includes cities like Stockton, Manteca, Tracey Morada, Lodi with an estimated population of 700000. According to NCES, there are 668 (472 male and 196 female) students while there are only three existing public schools registered in 2018-2019 school year.

Similarly in 22 I create the persistent diagram conditional on California private schools are active through the period from 1999 to 2015. One can see, the most persistent cycle is represented by the black curves on the right side. There are a few 2-dimensional cycles that are represented above the bootstrapped band. The most persistent cycle is in northern California, where there are only a few private schools founded. This result is consistent with the topmost 1-dimensional generator which is further from the identity line on top. This is due to the fact that observed few private schools in the north are further from the rest of the point cloud, which results in d overextending to connect the components.

There has been a continuous discussion whether education should be privatized. The question needs to account for a wide variety of factors including how a private school in a region affects the composition of student body and their learning outcomes. I use Wasserstein metric to unveil whether private schools locations in 1999 had more ties with the public schools of the time. Or whether the private schools were effective at covering the map of California and the locations are more closely related with the public schools at later periods. Q-th Wasserstein distance between two persistent diagrams is defined as

$$W_q(X,Y) = \left(\min_{f:X\to Y} \left(\sum_{x\in X} ||x - f(x)||_\infty^q \right) \right)^{1/q}$$

Information on its equivalent metric can be found in [10] where the authors discuss the quality of the metric and its advantages. Entrance of new components in the persistent diagram will not effect the metric much. Behind the equation there is a process of

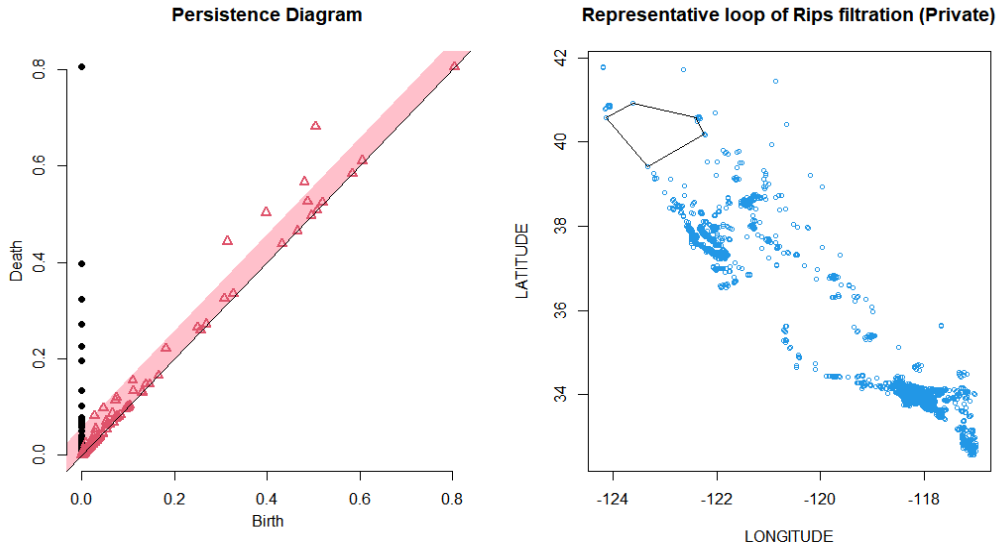


Figure 22: Private Schools 1851-1995

(a) The 1-dimensional barcode for private schools is represented in blue and 0-dimensional generators are colored in red. The barcode confirms that most of the generators were short-lived. However, as we go down closer to red we end up identifying longer p-intervals. The longest one on the bottom runs forever because after d gets big enough the whole point cloud turns into one single generator. However, long but not infinite lines indicate the secluded schools that were in the northern region of California.

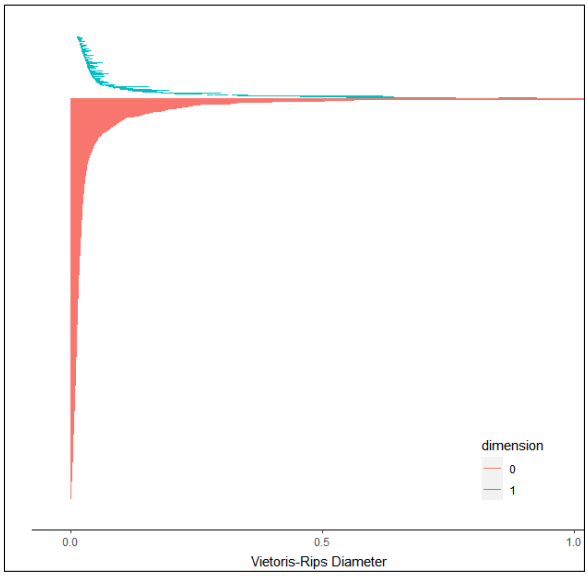


Figure 23: Barcode for Private Schools

mapping the generators of one persistence to another and finding the minimal effort needed to relocate the point cloud to match with another dataset. Wasserstain distance between private and public schools that were open before 1995 is 0.21. However the metric between private schools that opened before 1995 and public schools that opened up until 2020 is 0.29. It means that the grid created from private and public school locations in 1995 evolved greatly. As a result, one can observe a difference in the underlying structure of the grids of private schools in the past and public schools of the day. Thus, there is a possibility that private and public schools are founded in close relation with one another when looking at the 2-dimensional cycles they create.

Private schools are different from one another and there is great heterogeneity when looking at their characteristics. NCES data provides a wide range of characteristics that describe the schools. The following parameters are included: population, enrollment, teacher count, starting school grade, ending school grade, zip code foundation year. See the correlation matrix below:

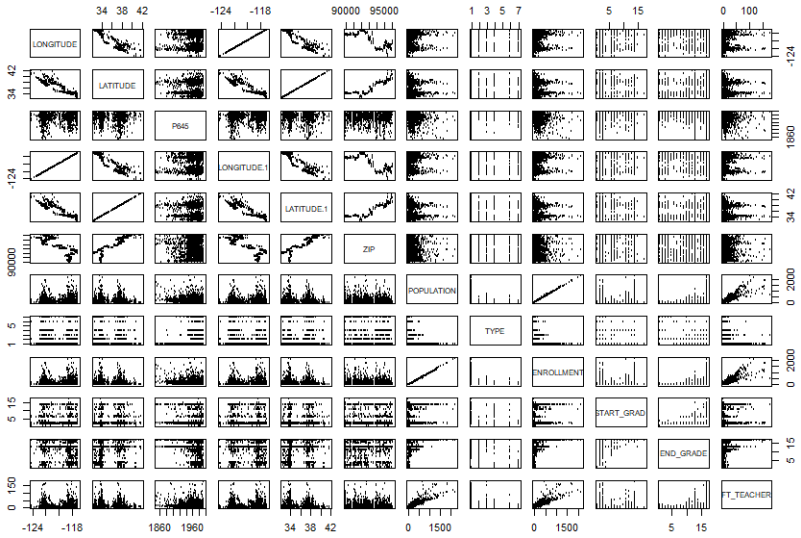


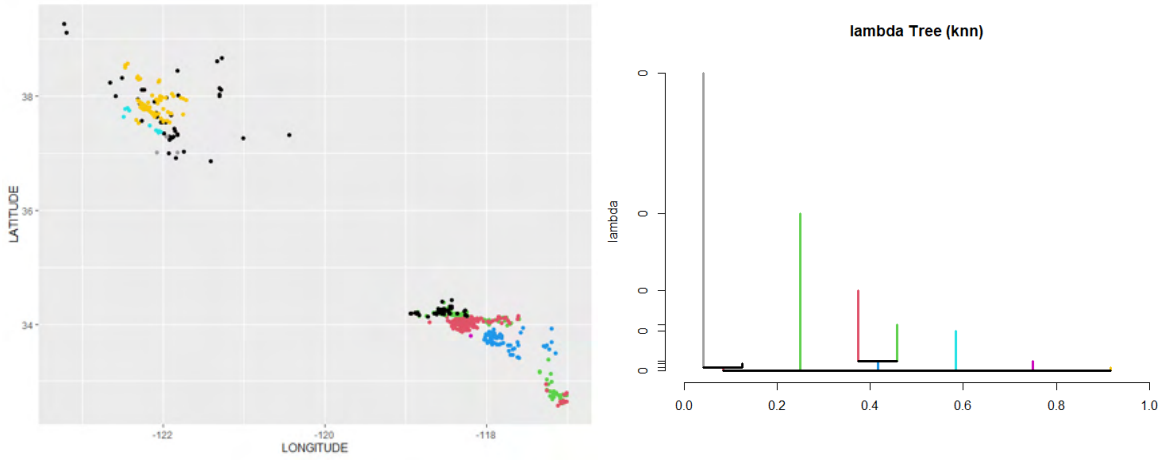
Figure 24: Correlation Matrix Private Schools

To better understand private schools, in the last step of our analysis, I use the characteristics of our school to do hierarchical clustering. The cluster density tree algorithm is based on [11]. If we assume $V = (v_1, \ldots, v^n) \subset \mathbb{R}^n$ is the observed dataset, the level set of f is defined as $L_f(\alpha) = cl(v \in \mathbb{R}^m : \ f(x) > \alpha)$. The density regions together at different levels define the cluster tree. Collection of sets $L_f(\alpha)$ where $T = L_f(\alpha), \alpha \geq 0$ is the α cluster tree.

I ran the clustering algorithm and, the result indicates that there are 7 relatively big clusters that come up. One can suggest that private schools which are located next to each other share a lot of characteristics. The only apparent cluster that is somewhat separated is the one which is painted black. However, it is important to note that these are the schools that were the first to open up in early 1900s.

5.4 Concluding Remarks

Persistent homology secures the purpose of analysing high dimensional complex data. The type of data can vary significantly, it needs to be transformed into the form of point



(a) Representative Clusters On The Map

(b) α Cluster Tree

Figure 25: Private School Clustering

cloud data. Persistent homology allows access to the information about generators and relators of cycles. Simplical homology in abstract algebra provides mathematical basis for identifying these cycles using boundary maps. The output of persistent homology can vary greatly from minute changes in data structure. In the future, it will be interesting to explore persistent co-homology, which has similar roots in abstract algebra and provides more stable output.

References

[1] A. Hatcher, “Algebraic Topology: A First Course,” 2002

[2] Munkres, J. R. Elements of Algebraic Topology. Addison-Wesley, Reading, MA, 1984.

[3] A. Zomorodian and G. Carlsson, “Computing Persistent Homology,” 2004.

[4] N. Otter, M.A. Porter, U. Tillmann, P. Grindrod, and A. Harrington, “A Roadmap for the Computation of Persistent Homology,” 2017.

[5] V. Silva, G. Carlsson, “Topological estimation using witness complexes,” Eurographics Symposium on PointBased Graphics 2004.

[6] Muthu Alagappan. From 5 to 13: Redefining the positions in basketball. In MIT Sloan Sports Analytics Conference, 2012.

[7] M. Nicolau, A. J. Levine, and Gunnar Carlsson ”Topology based data analysis identifies a subgroup of breast cancers with a unique mutational profile and excellent survival”, 2011

[8] H. Lee, M. K. Chung, H., Kang, B., Kim, D., and S. Lee ”Discriminative Persistent Homology of Brain Networks”, 2011

[9] Fasy BT, Lecci F, Rinaldo A, Wasserman L, Balakrishnan S, Singh A “Confidence sets for persistence diagrams”, (2014).

[10] M. Kerber, D. Morozov, A. Nigmetov, ”Geometry Helps to Compare Persistence Diagrams”, 2017

[11] Hartigan JA “Consistency of single linkage for high-density clusters.” Journal of the American Statistical Association, (1981)

[12] TDAstats to calculate persistent homology (Ripser): Bauer U. Ripser: Efficient computation of Vietoris-Rips persistence barcodes. 2019;

[13] Maria C (2014). “GUDHI, Simplicial Complexes and Persistent Homology Packages.”

[14] Morozov D (2007). “Dionysus, a C++ library for computing persistent homology.”

[15] Peter Giblin, ”Graphs, Surfaces and Homology”, 2011

[16] A. Zomorodian, G. Carlsson ”Computing persistent homology” 2002

List of Figures