Persistent Homology: Computation and applications of a Modern Data analysis tool

Pontificia Universidad Javeriana

Facultad de ciencias

Departamento de Matematicas

David Ricardo Gonzalez Jimenez

Mayo 2020

Contents

1 Introduction **2**

2 Homology **3**
2.1 Simplices and simplicial complex 3
2.2 Calculating homology 6

3 Building Persistent homology **9**
3.1 Building Simplices in Data 9
3.2 Persitent Homology 11
3.3 Example of computing persistent homology 14

4 Applications: **18**

5 Conclusion: **24**

Chapter 1

Introduction

In the last years the amount of data that is available has increased dramatically, such availability of data posses new challenges and is a fertile ground for new methods. Such methods must take into account the growing completeness and complexity of the data; before, we only had information of some demographic characteristics of individuals, now it is possible to infer relationships between individuals and how they interact with each other in complex environments. A miriad of new methods available to use in this new context come from the field of topology, such methods have been proven useful for a multitude of problems, from classification to prediction. The advantages from this methods comes from the fact that it uses topology and geometry as a way to infer robust information about the structure of the data [1].

One of the most used tools in the Topological Data Analysis toolbox (from here on TDA) is Persistent Homology. This method is widely used because allow us to draw a precise qualitative features of the data structure, a tool to compute such features in practice and a guarantee about the robustness of these features [2].

The present work is an effort to review the principles of persistent homology, explain its mathemathical intuitions and why this makes it a powerful tool for data analysis. In order to underline this importance even more we will also review an application to the field of social networks; moreover, it will be shown how the topological insights from persistent homology can help unravel important global relationships of the networks that were not available before.

The current work follow the following structure: The first section is about homology, what it is, the structure needed to build it and how it is calculated. The second section will be about Persistent homology, what it is, how it is build and what it can show us about the data. Finally an application of persistent homology to weighted networks will be tackled and we will see what the authors of this applications gain when using persistent homology.

Chapter 2

Homology

In this section there is going to be a brief introduction to important theoretical elements that are needed to talk about. Take into account that it is not a throughout review of concepts, instead here are listed those that are important for what comes. First as we will be dealing with data points and clouds of data, we will try an approximation of homology from simplicial complex. In this section first simplices and simplicial complexes will be introduced, then the concept of homology will be presented and some important results that allow is computability. Later we will introduce Persistent homology and how it can be computed, finally some examples of how this methods have been used and its usefulness acording to some examples on the literature.

2.1 Simplices and simplicial complex

First we will define what is a simplex and a simplicial complex, here we give both the definition of the geometrical simplicial complex and an abstract simplicial complex [3].

Definition: let $\{a_0, \cdots, a_n\}$ be a lineally independent set in $\mathbb{R}^N$. Define σ as a n-simplex spanned by $a_0, \cdots, a_n$ to be the set of all points x of $\mathbb{R}^N$ such that

$$x = \sum_{i=0}^{n} = t_i a_i \text{ where } \sum_{i=0}^{n} t_i = 1 \text{ and } t_i \geq 0 \text{ for all } i$$

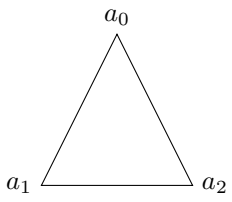


Figure 2.1: example of a 2-simplex

Points $a_0, \cdots, a_n$ are called vertices of σ, n being the dimension of σ. A simplex that is spanned by a subset of vertices of σ is called a face.

Definition: A simplicial complex in K in $\mathbb{R}^N$ is a collection of simplices in $\mathbb{R}^N$ such that:

1. Every face of simplex of K is in K
2. The intersection of any two simplices of K is a face of each of them

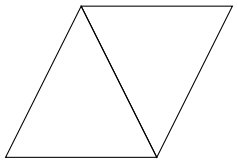


Figure 2.2: Example of a simplicial complex

Definition: An abstract simplicial complex is a collection $\mathscr{S}$ of finite non-empty sets, such that if A is an element of $\mathscr{S}$, so is every nonempty subset of A

Each simplex can have an orientation, that is, each simplex σ is given an ordering of its vertex, such orders are equivalent if they differ by an even permutation. If $dim\, \sigma > 0$ orderings of vectors fall in two equivalence classes. Each class is an orientation, an oriented simplex is just a simplex and its orientation.

Definition: let K be a simplicial complex. A $p - chain$ on K is a function c from the set of oriented p-simplices of K to the integers, such that:

1. $c(\sigma) = -c(\sigma')$ if σ and σ' are opposite orientations of the same simplex
2. $c(\sigma) = 0$ for all but finitely many oriented p-simplices σ

The free abelian group generated by the set of p-chains on K $\{c(\sigma), +\}$ is denoted as $C_p(K)$ and is called the group of oriented $p - chains$.

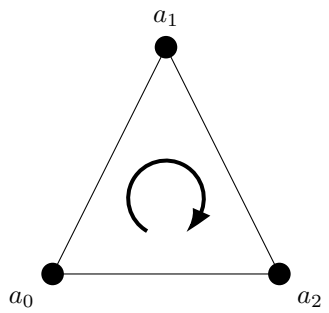


Figure 2.3: example of a 2-chain

Definition: Define the following homomorphism

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

called the boundary operator. If $\tau = [v_1, .., v_p]$ is an oriented simplex of dimension p with $p > 0$ we define:

$$\partial\tau = \partial\,[v_1, .., v_p] = \sum_{i=0}^{p} (-1)^i [v_1, \ldots, \hat{v_i}, \ldots, v_p]$$

where $\widehat{v_i}$ means that the vertex v_i has to be deleted from the array.

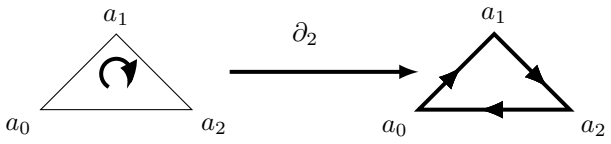


Figure 2.4: Example of a boundary operator over a 2-simplex

Now we can define Homology, in lay language what homology allow us to do is to classify topological spaces by the holes on it.

Definition Homology group: The kernel of $\partial_p : C_p(K) \longrightarrow C_{p-1}(K)$ is called the group of p-cycles and denoted $Z_p(K)$. The image of $\partial_{p+1} : C_{p+1} \longrightarrow C_p(K)$ is called the group of $p - boundaries$ and is denoted $B_p(K)$. Each boundary of a p+1-chain is a p-cycle (as $\partial_p \circ \partial_{p+1}$=0). That is, $B_p(K) \subset Z_p(K)$. Define:

$$H_p(K) = Z_p(K)/B_p(K)$$

and call it the p-th homology group.

This definition has a lot to unpack, first the group $Z_p(K)$ is the group of p-chains that have an empty boundary, hence $Z_p(K) \subset C_p(K)$. The fact that $\partial_{p-1} \circ \partial_p$=0 tells us that the boundary of a p-chain is a $(p-1)$-cycle, hence that $B_p(K) \subset Z_p(K)$. Hence as the homology is the quotient $Z_p(K)/B_p(K)$ what it does is to classify the cycles in a cycle group by putting together those cycles in the same class that differ by a boundary.

It is possible to reduce the calculation of the groups of homology to simple linear algebra. This can be done from the matrices representing the boundary homomorphism. This matrix representation provides the ranks of the cycle and boundary groups, and their difference provide the Betti numbers. The following definition allow us to see it more formally

Definition: A chain complex $\mathscr{C}$ is a sequence:

$$\cdots \to C_{p+1} \xrightarrow{\partial_{p+1}} C_p \xrightarrow{\partial_p} C_{p-1} \to \cdots$$

of abelian groups C_i and homomorphisms ∂_i such that $\partial_{i-1} \circ \partial_i = 0$. The $p-th$ homology group of $\mathscr{C}$ is defined by the equation

$$H_p(\mathscr{C}) = ker(\partial_p)/im(\partial_{p+1})$$

2.2 Calculating homology

Now in order to give an example we will calculate a the homology of real projective plane of dimension 2, $\mathbb{R}P^2$. The simplicial decomposition appears on 2.5.

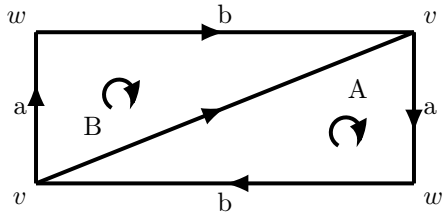


Figure 2.5: Simplicial decomposition of the $\mathbb{R}P^2$

The chain complex for this simplicial complex is the following:

$$0 \longrightarrow C_2 \xrightarrow{\partial_2} C_1 \xrightarrow{\partial_1} C_0 \xrightarrow{\partial_0} 0$$

$$0 \longrightarrow \underset{\langle A,B\rangle}{\mathbb{Z}\oplus\mathbb{Z}} \xrightarrow{\partial_2} \underset{\langle a,b,c\rangle}{\mathbb{Z}\oplus\mathbb{Z}\oplus\mathbb{Z}} \xrightarrow{\partial_1} \underset{\langle w,v\rangle}{\mathbb{Z}\oplus\mathbb{Z}} \xrightarrow{\partial_0} 0$$

now we compute the boundary of the simplices and compute the homology.

$$\partial_1(a) = w - v$$
$$\partial_1(b) = v - w$$
$$\partial_1(c) = 0$$
$$\partial_1 = \begin{pmatrix} -1 & 1 & 0 \\ 1 & -1 & 0 \end{pmatrix} \sim \begin{pmatrix} -1 & 1 & 0 \\ 0 & 0 & 0 \end{pmatrix}$$

where $im(\partial_1) = \mathbb{Z}$ since there is only one independent row vector in the matrix representation of ∂_1. As can be seen in the chain representation $ker(\partial_0) = \mathbb{Z}\oplus\mathbb{Z}$, then

$$H_0(\mathbb{R}P^2) = \mathbb{Z} \oplus \mathbb{Z}/\mathbb{Z} = \mathbb{Z}$$

Now

$$ker(\partial_1) = \left\langle \begin{pmatrix} 1 \\ 1 \\ 0 \end{pmatrix}, \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix} \right\rangle$$

$$ker(\partial_1) = \langle v_1, v_2 \rangle$$

$$\partial_2(A) = a + b + c$$

$$\partial_2(B) = a + b - c$$

$$\partial_2 = \begin{pmatrix} 1 & 1 \\ 1 & 1 \\ 1 & -1 \end{pmatrix}$$

where

$$im(\partial_2) = \left\langle \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix}, \begin{pmatrix} 1 \\ 1 \\ -1 \end{pmatrix} \right\rangle$$

putting it all in terms in v_1 and v_2

$$im(\partial_2) = \langle v_1 + v_2, v_1 - v_2 \rangle$$

$$\begin{aligned} H_1(\mathbb{R}P^2) &= \langle v_1, v_2 \rangle / \langle v_1 + v_2, v_1 - v_2 \rangle \\ &= \langle v_1 + v_2, v_2 \rangle / \langle v_1 + v_2, v_1 - v_2 \rangle \\ &= \langle v_1 + v_2, v_2 \rangle / \langle v_1 + v_2, 2v_2 \rangle \\ &= \mathbb{Z}/2\mathbb{Z} \end{aligned}$$

$$H_1(\mathbb{R}P^2) = \mathbb{Z}/2\mathbb{Z}$$

Now lets calculate the the $H_2(\mathbb{R}P^2)$

$$ker(\partial_2) = \mathbf{0}$$

$$H_2(\mathbb{R}P^2) = \mathbf{0}/\mathbf{0}$$

As it is now clear how to compute the homology with the example of the projective plane, it is now done for the Klein Bottle which we are going to represent with **K**.

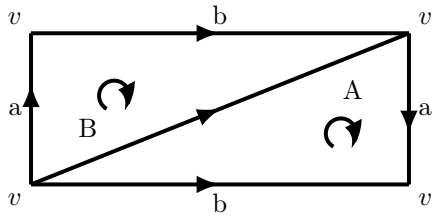


Figure 2.6: Simplicial decomposition of the Klein Bottle

The chain complex is the following:

$$0 \longrightarrow C_2 \xrightarrow{\partial_2} C_1 \xrightarrow{\partial_1} C_0 \xrightarrow{\partial_0} 0$$

$$0 \longrightarrow \underset{\langle A,B\rangle}{\mathbb{Z}\oplus\mathbb{Z}} \xrightarrow{\partial_2} \underset{\langle a,b,c\rangle}{\mathbb{Z}\oplus\mathbb{Z}\oplus\mathbb{Z}} \xrightarrow{\partial_1} \underset{\langle v\rangle}{\mathbb{Z}} \xrightarrow{\partial_0} 0$$

Now in order to calculate the homology

$$\partial_0(a) = \partial_0(b) = \partial_0(c) = v - v = 0$$

$$\partial_1 = \begin{pmatrix} 0 & 0 & 0 \end{pmatrix}$$

Hence

$$H_0(\mathbf{K}) = \mathbb{Z}/\mathbf{0} = \mathbb{Z}$$

$$ker(\partial_1) = \langle v_1, v_2, v_3 \rangle$$
$$\partial_2(A) = a - b + c$$
$$\partial_2(B) = a + b - c$$
$$\partial_2 = \begin{pmatrix} 1 & 1 \\ 1 & 1 \\ 1 & -1 \end{pmatrix}$$

where

$$im(\partial_2) = \left\langle \begin{pmatrix} 1 \\ -1 \\ 1 \end{pmatrix}, \begin{pmatrix} 1 \\ 1 \\ -1 \end{pmatrix} \right\rangle$$

calculating the homology

$$H_1(\mathbf{K}) = \langle v_1, v_2, v_3 \rangle \,/\, \langle v_1 - v_2 + v_3, v_1 + v_2 - v_3 \rangle$$
$$= \langle v_1 - v_2 + v_3, v_2, v_3 \rangle \,/\, \langle v_1 - v_2 + v_3, v_1 + v_2 - v_3 \rangle$$
$$= \langle v_1 - v_2 + v_3, v_2, v_3 \rangle \,/\, \langle v_1 - v_2 + v_3, 2v_3 \rangle$$
$$= \mathbb{Z} \oplus \mathbb{Z} \oplus \mathbb{Z}/\mathbb{Z} \oplus 2\mathbb{Z}$$
$$H_1(\mathbf{K}) = \mathbb{Z} \oplus \mathbb{Z}/2\mathbb{Z}$$

Now lets calculate the the $H_2(\mathbf{K})$

$$ker(\partial_2) = \mathbf{0}$$

$$H_2(\mathbf{K}) = \mathbf{0}/\mathbf{0}$$

Chapter 3

Building Persistent homology

As we have already introduced the concept of homology and how it relates to simplices, we will now draw the connection between this and data. In hindsight what is needed to be done is to find a way to build simplicial complexes from data and then to calculate their homology groups and topological invariants.

3.1 Building Simplices in Data

In most of the applications of data, what we are given is data that is some set of point P. The problem is that Homology is built is over simplicial complexes, hence there is the need to turn a data point a simplicial complex. There are various methods of building simplicial complexes out of data points, we will profit from some results that will allow us to build an abstract simplicial structure on the data and relate it with the concept of Homology.

In order to do this we need a set of points P and a notion of distance between points, hence we will need that P is a suset of the metric space (M,d).

Definition: Let M be a topological space. Let $\mathcal{M}$ be a open cover of M. The Nerve of $\mathcal{M}$ is the abstract simplicial complex $\mathcal{K}$ defined on the set $\mathcal{M}$ where a simplex $\{a_1, a_2, ..., a_k\} \subset \mathcal{M}$ is in $\mathcal{K}$ if

$$\bigcap_{i=1}^{k} a_i \neq \emptyset$$

Čech complexes: Let (M,d) be a topological space, with a topology induced by its metric. Let P be a subset of M. Given a real $r > 0$, the Čech complex $\mathcal{C}^r(P)$ is defined to be the nerve of the set $\{B(p, r/2)|p \in P\}$

$$B(p, r/2) = \{x \in M | d(p, x) < r/2\}$$

is the metric open ball of radius r/2 centering p.

The construction of a Čech complex can be computationally demanding, to tell whether there are any 10-simplices (without additional knowledge) you have to inspect all subsets of size 10. In general for computating a Čech complex the algorithm must store in memory either the entire complex and its boundary operator, or the precise distances between vertices.

The Vietoris-Rips complex is similar to the Čech complex, except instead of adding a n-simplex when there is a common point of intersection of all the (r/2)-balls, we just do so when all the balls have pairwise intersections.

Vietoris-Rips Complex Let (P, d) be a metric space where P is a set of points. Given a real $r > 0$ the Vietoris-Rips complex is the abstract simplicial complex σ where $\sigma \in \mathcal{R}^r(P)$ if and only if $d(p, q) < r$ for every pair of vertices of σ.

The edges in the Vietoris-Rips complex are the same as in the Čech complex. This can be seen in the next result:

Lemma: Let P be a subset of the metric space (M,d) then

$$\mathcal{C}^r(P) \subset \mathcal{R}^r(P) \subset \mathcal{C}^{2r}(P)$$

Proof. The first inclusions follows from the fact that if there is a point x in the intersection of $\bigcap_{i=1}^{k} B(p_i, r/2)$ the distances $d(p_i, p_j)$ for every pair (i, j), $1 \leq i, j \leq k$ are at most r, it follows then that if the simplex $\{p_1, p_2, .., p_k\} \in \mathcal{C}^r(P)$ is also in $\mathcal{R}^r(P)$. For the second inclusion consider the simplex $\{p_1, p_2, .., p_k\} \in \mathcal{R}^r(P)$ since by definition of the Rips complex $d(p_j, p_1) < r$ for every p_j we have $p_1 \neq \emptyset$ and $p_1 \subset \bigcap_{i=1}^{k} B(p_i, r)$, then by definition $\{p_1, p_2, ...p_k\} \in \mathcal{C}^{2r}(P)$. □

Even though the Vietoris-Rips complex might be more computationally efficient the Čech complex preserves better the topological qualities of the data because of the Nerve theorem, that says that if F is a finite collection of closed, convex sets in Euclidean space then the nerve of F and the union of the sets in F have the same homotopy type.

In the following Figure there is a very simple example of the difference between constructing a Vietoris-Rips complex and a Čech complex. While for the Vietoris-Rips a pairwise intersection between balls is enough to draw a 2-simplex, for the Čech complex it is not enough as there is no triple intersection between balls.

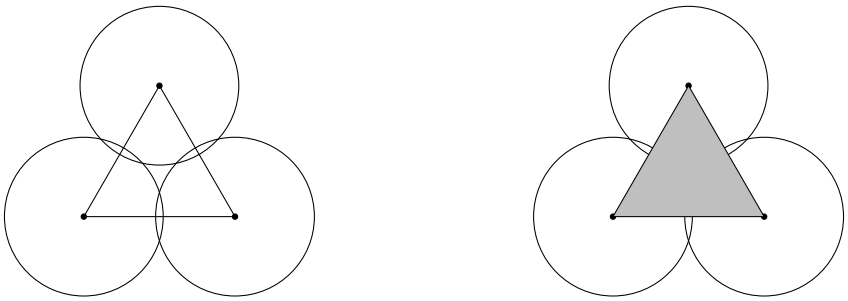


Figure 3.1: Example of a Čech complex in the left and a Vietoris-Rips complex in the right hand

3.2 Persitent Homology

The goal with Persistent Homology as it is with other tools of TDA is to extract topological qualities from an underlying space just by studying a sample. Hence it is important that those qualities that are extracted are robust to small perturbations in the data. Persistent Homology is useful as it allows to measure the scale or resolutions of topological features, it does so by capturing how the homology of the complexes change as parameters change and which features persist.

In order to grasp how this features and classes change it is important to determine when and how a feature "survives". What this means is that as parameters change it will be important to measure when for each value of the changing parameter the feature starts and when does it end. A graph representation can be done out of the paths of the features and the only way to generate a unique graph representation is through the elder rule [4]:

The elder rule: at a juncture, the older path continue and the younger path ends.

Persistent homology is obtained when the elder rule is formulated in homology groups of all dimensions. In order to observe and evolution, the space must be constructed in such a way that simplices are added as the space evolves, for this we use the concept of filtration

Definition if $\mathscr{X}$ is a topological space, a filtration of $\mathscr{X}$ is a sequence $X_0 \subset X_1 \subset \dots$ of subspaces of $\mathscr{X}$ whose union is $\mathscr{X}$. A space $\mathscr{X}$ together with a filtration of $\mathscr{X}$ is called a filtered space.

It is important to define the concept a filtration, first consider a topological space X and a real valued function $f : X \to \mathbb{R}$. Each value $i \in \mathbb{R}$ gives rise to two excursion sets, the sublevel set $f^{-1}((-\infty, i])$ and the superlevel set $f^{-1}([i, \infty))$,

the sublevel set will be called $X(a)$. For this sublevel sets the following is true $X(a) \subset X(b)$ for $a \leq b$, so if the inclusion map denoted by $i : X(a) \to X(b)$ then we have an induced map

$$f = i_* : H_p(X(a)) \to H_p(X(b))$$

hence for $i \in \{1, \cdots n\}$, as $1 < \cdots < n$ then $X(1) \subset \cdots \subset X(n)$ calling $X(i) = X_i$ it follows that $H_p(X_a) \to \cdots \to H_p(X_n)$. So we call $f_p^{i,j} : H_p(X_i) \to H_p(X_j)$ [5].

Now in order to build a filtration of a simplicial complex and its homology groups as in [4] we define K a simplicial complex consider a function $f : K \to \mathbb{R}$ where f is monotonic, that means that is non-decreasing in increasing chains of faces, so if $\sigma \subset \tau$ then $f(\sigma) \leq f(\tau)$. As f is monotonous we have that $K(a) = f^{-1}(-\infty, a]$ is a subcomplex of K for every $a \in \mathbb{R}$. The function f gives an ordering of the subcomplexes, allowing to arrange the diferent subcomplexes in increasing sequence

$$\emptyset = K_0 \subset K_1 \subset \ldots \subset K_n = K$$

where m is the number of simplices and $n + 1 \leq m + 1$. We have that $a_1 < a_2 < \ldots < a_n$ are the function values of the simplices in K and $a_0 = -\infty$, then $K_i = K(a_i)$ for each i, hence K becomes a filtered space.

Now that there are sequences of simplicial complexes it is possible to study the sequences of the corresponding homology groups. As there is an inclusion for every $i \leq j$ there is an inclusion map from K_i to K_j and therefore an induced homomorphism $f_p^{i,j} : H_p(K_i) \to H_p(K_j)$ for each dimension p. Hence taking into account the increasing sequence of K

$$0 = H_p(K_0) \to H_p(K_1) \to \ldots \to H_p(K_n) = H_p(K_n)$$

As we move from the sequence homology classes change and those changes is what is of interest in persistent homology.

Definition: The $p-$th persistent homology groups are the images of the homomorphisms induced by inclusion $H_p^{i,j} = Imf_p^{i,j}$, for $0 \leq i \leq j \leq n$. The Betti numbers are the ranks of these groups, $\beta_p^{i,j} = rankH_p^{i,j}$.

Persistent homology groups for every dimension p and index $i \leq j$ are the homology classes of K_i that ares still alive in K_j hence the quotient in the persistent homology changes and it becomes $(B_p(K_j) \cap Z_p(K_i))$, so $H_p^{i,j} = Z_p(K_i)/(B_p(K_j) \cap Z_p(K_i))$. Now is important to define when a class is "born" and when it "dies", a class is born when $\delta \in H_p(K_i)$ and $\delta \notin H_p^{i-1,i}$, by the elder rule δ "dies" in K_j when it merges with another class, hence $f_p^{i,j-1}(\delta) \notin H_p^{i-1,j-1}$ but $f_p^{i,j}(\delta) \notin H_p^{i-1,j}$. Where the difference in the values of the functions is denoted the *persistence* $pers(\delta) = a_i - a_j$, where a_i, a_j are the image of

the filtration function for $K(a_i), K(a_j)$ and $K(a_i) \subset K(a_j)$.

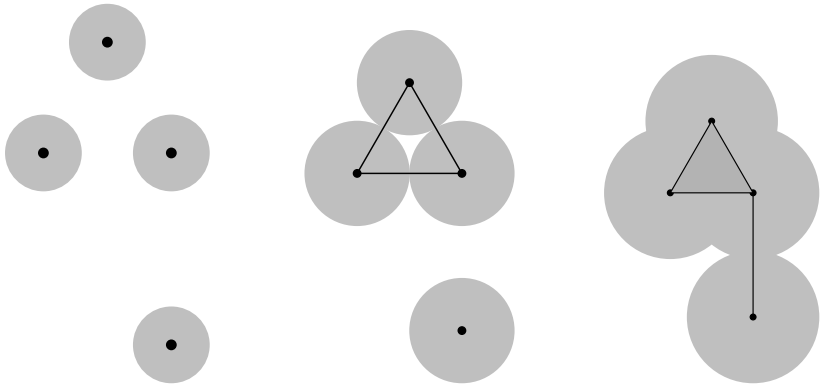
This is what is advantageous of the pth persistent homology in contrast with the pth homology it gives information not only of one of the subcomplexes of the filtered complex, but information of all the previous and following complexes. It gives us information of in which subcomplex connected components are "born" or when they "die". This is important as it allow us to give one of the most important results in this analysis that are the barcode diagrams and persistence diagrams, where it is illustrated the born and death of connected component.

It is now important to show how persistent diagrams are constructed and how they should be interpreted. The persistent Betti numbers drawn in two coordinates in the space $\overline{\mathbb{R}}^2$ as classes that never die are said to go to infinity. Let $\mu_p^{i,j}$ be the number of p-dimensional classes that are born at K_i and die entering K_j then

$$\mu_p^{i,j} = (\beta_p^{i,j-1} - \beta_p^{i,j}) - (\beta_p^{i-1,j-1} - \beta_p^{i-1,j})$$

for all $i < j$ and all p. The first difference counts classes that are born at or before the subcomplex K_i and die entering K_j while the second counts the ones that are born at K_{i-1} and die entering K_j [4]. Drawing each point (a_i, a_j) with multiplicity $\mu_p^{i,j}$ we get the persistence diagram.

The following graph shows how persistent homology is build onto a very simple cloud point, subcomplexes are build according to a filtration. In the first complex 3.2a there are no connected components only points, all the elements of the 0-d homology in 3.3 are born at 0 as shown . The second subcomplex 3.2b shows the birth of a connected component (represented by the dots in red)and the dead of two zero homologies as they become the radius interlap. In the last subcomplex 3.2c the last radius that was separated interlaps and hence this marks the dead of another 0-homology class. For simplicity we assume all components die in after this last subcomplex.



(a) 0-subcomplex (b) 1-subcomplex (c) 2-subcomplex

Figure 3.2

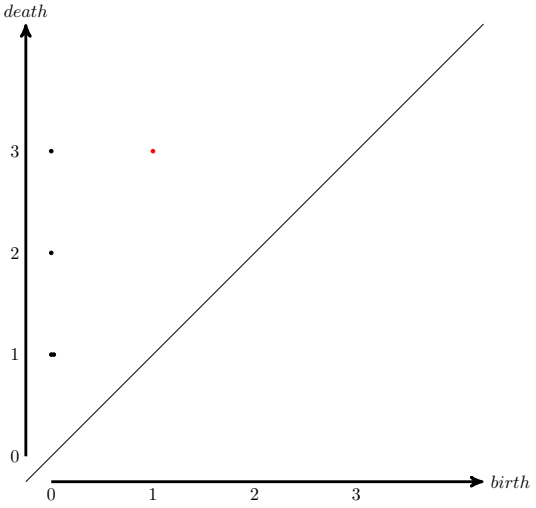


Figure 3.3: Persistence Diagram of the complex on 3.2

3.3 Example of computing persistent homology

As it has been introduced theoretically what is homology and what is a persistent homology. Now it is going to be shown how it is calculated in a point cloud with a known intrinsic geometry, for this exercise we are going to use the parametrization of the Klein Bottle and the TDA package [6].

The TDA package produces an interface in R for several packages of TDA such as GUDHI. As the focus of this work is in persistent homology it is impor-

tant to underline the tools in this package that are going to be used to tackle this topic. First it is important to remark what type of simplicial structure is going to be used and over what it is going to be build. The function *ripsDiag* is going to be used to build the simplicial complex and to calculate its persistence, such function builds a Vietoris-Rips complex over the point-cloud. It is possible to feed this function with the maximun dimension of the simplices to be build and the maximun dimension of the homology.

Now in order to easily conceptualize how persistent homology works and how it is computed we are going to calculate the persistent homology of a point cloud derived from a uniform sampling of points of a Klein Bottle. This allows us to compare the exercise of computing the persistent homology of a point cloud and compare it with results in the previous section. The parametrization of the Klein Bottle to be followed is one of the simplest and is known as the "Pinched Torus", where the parameneter R is simplified as $R = 1$, $0 \leq r \leq 1$ and $(\theta, \phi) \in (0, 2\pi] \times (0, 2\pi]$ hence the parametrization is as follows:

$$w(\theta, \phi) = (1 + rcos\theta)cos\phi$$
$$x(\theta, \phi) = (1 + rcos\theta)sin\phi$$
$$y(\theta, \phi) = rsin\theta cos\frac{\phi}{2}$$
$$z(\theta, \phi) = rsin\theta sin\frac{\phi}{2}$$

The sampling model of the klein bottle comes from [7] and its graphical representation allows us to represent a complex figure in a series of 2D figures in a matrix like form. To interpret each graphic bear in mind that the horizontal axis for each graphic is the letter in the same column of the matrix and the vertical axis is the letter in the same row.

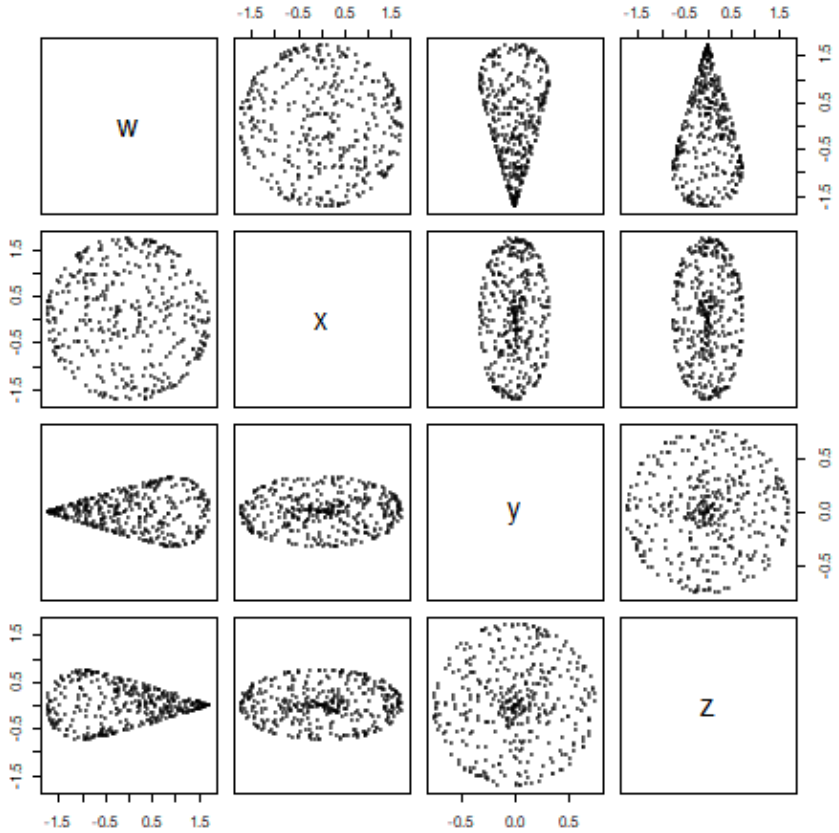


Figure 3.4: A point cloud from a Klein Bottle

Now we can compute the persistent homology of this point cloud using the TDA package and the *ripsDiag* function.

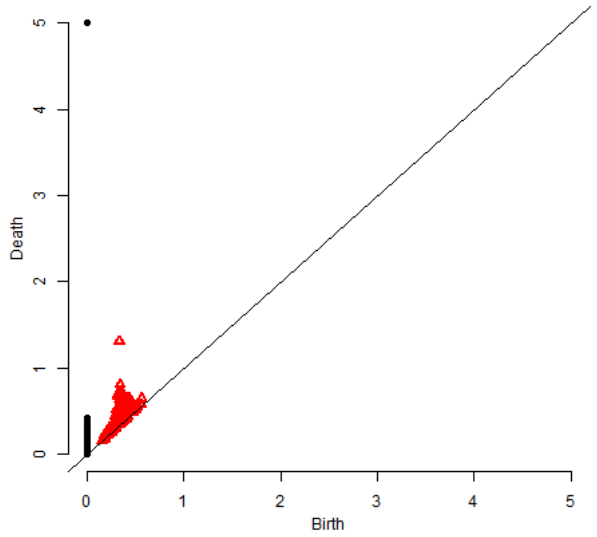


Figure 3.5: Persistent homology of Klein Bottle

As this is a representation of the kleinn bottle it is useful to remeber the results of the homology in Section 2.2. where $H_1(K) = \mathbb{Z} \oplus \mathbb{Z}/2\mathbb{Z}$ and $H_0(K) = \mathbb{Z}$ and see how this is reflected in 3.5.The circles in black are the persistance points of the elements 0d-Persistent homology group and on red there are the points of the 1d-Persistent homology group. The diagrams of the 0d-persistent every one of this points is born at time zero, as is expected with an 0 homology, and as the data cloud is very populated each homology dies very quickly. The 1d-persistent homology tells us about the loops and what is interesting is the point that stand outs of the others because it dies out latter this point illustrates one of the connected components of the homology, while the torsion we are not able to see it because of limitations of the GUDHI package.

Chapter 4

Applications:

In this section an application of persistent homology is exposed in the hopes of underlining its importance in the TDA toolkit. In one of the fields in which Persistent homology shows more promise is in the area of networks, where there has been an interest for methods outside of those of graph theory where it was born, into methods that allow the study of intrinsic geometry and topologies of the networks. This offers advantages that graph based methods do not, as structural holes and voids give us information about the network and it is possible to derive measures of similarity that where not possible before [8].

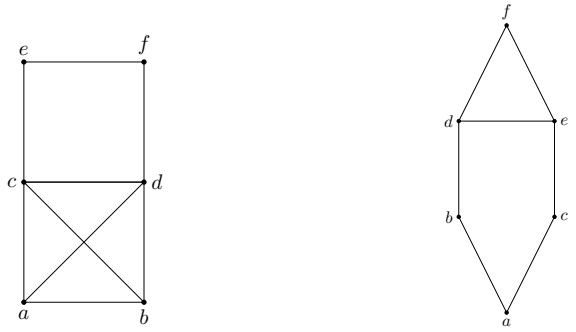
In this work it is going to be reviewed how Persistent Homology is applied in weighted directed networks. There are two important features in this networks, first not all the edges have the same weight and second they are not symetrical networks. There are two main proposed approaches to this type of networks, a Persistent Path Homology [9] and the clique complex and links thresholding approach [10]. In this work the latter approach will be touched with special emphasis on its application in migration and remittances as shown in [11], as this authors say the main idea is that: *given a directed network, construct a mathematical object C (called a Simplicial complex) in such a way that resulting topological features of C corresponds to patterns in the network.*

In their paper Ignacio & Darcy [11] study the migration and Remittances networks of Asia. They used Persistent Homology as a way of capturing higher dimensional pattern that are not captured in the usual methods in graph theory that focus on specific statistics of clusters or vertices, such as centrality, indegree or outdegree. Instead looking on persistent homology will allow a different characterization and characterizations of the flows of this network.

As we have seen in the preliminaries in order to compute the persistent homology of a cloud point a simplicial complex and a filtration should be build. In networks it is important to take into account is that edges carry essential information of the network structure, information such as the direction of the

relationship between vertices or its relative importance (weight). It is then imperative in the analysis of this kind of network that this information is not ignored, hence the normal approach outlined in the preliminaries should be slightly modified. In the following pages this approach will be exposed and explained, in order to understand what the authors gain from it and how it could be used.

First let define the object of study and why there needs the tools from persistent homology must be modified. A symmetric network $N = (G, c)$ is a graph $G = (V, E)$,where V are the vertices and E are the edges, with a nonnegative capacity function $c : E \to \mathbb{R}$, where $c(xy) = c(yx)$[12]. Notice that if a network is non symmetric then we will have that $c(xy) \neq c(yx)$. Second in order to properly understand how the authors use persistent homology it is important to explain what kind of filtration they use, the approach is called a **Ranked weighted clique filtration**. The advantage of this approach is that cliques are a natural structure of networks and it uses the already hierarchical structure of them in weighted networks[10].

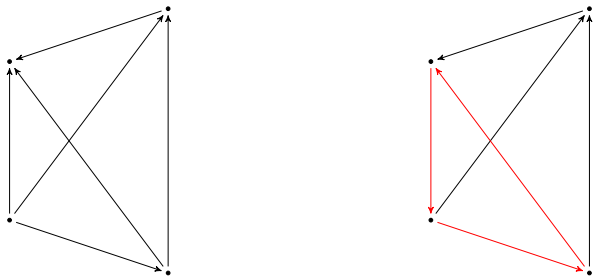


(a) Subfigure A (b) Subfigure B

Figure 4.1: In Subfigure A points there is one clique in $\{a, b, c, d\}$ in Subfigure B there is one clique $\{d, f, e\}$ and two $2-$cliques $\{a, b, c, d, e\}$ and$\{f, d, e, b, c\}$

Before talking about the filtration it is important to define what is a clique, what is an $n - clique$ and what is a directed $n-$clique as they are fundamental concepts to the work of this authors. A clique in a graph is simply a set of pairwise adjacent vertices[13]. Now this definition is stringent, so what is going to be used is an $n - clique$, an $n - clique$ is a subgraph where $n + 1$ vertexes are connected by paths of n length [14] as shown in Figure 4.1. A directed $n-$clique is a subgraph in which every pair of distinct vertices is connected by a unique edge, such that an order can be made any pair of nodes so there is a directed link pointing from the node with the higher order(a source) towards the lower one (a sink), this definition is equivalent on not having a loop on the $n - clique$ as shown in 4.2a in opposition to 4.2b [15]. Hence, This definition induces an

ordering on the set of k vertices and a maximal directed path, because of this when it is possible the simplices are going to be represented with an unidirectional flow [11].



(a) Subfigure A (b) Subfigure B

Figure 4.2: Subfigure A shows a directed $n-$clique while Subfigure B shows a non-directed n-clique

In order to be able to compute an homology what is done to build a directed clique, a *clique complex* $X(G)$ is done according to [16] by *coloring in* the cycles of the graph G, that is a directed $k+1-$clique is defined to be a simplex of dimension k. We have then that the clique complex $X(G)$ is an abstract simplicial complex, as one it is closed under taking subsets so is possible to write $X(G) = \{X_0(G), \cdots, X_M(G)\}$, with each $X_n(G)$ being a subcomplex of G made of simplices of dimension at most n called the $n-$skeleton [17]. Note that as simplicial complex has been defined, it is possible to take $C(X(G))$ and $Z(X(G))$ to compute the homology group.

Now it is important to see how the authors of the study build the network they, which is going to be the one in which the simplicial complex seen in the last paragraph are going to be build upon on. The authors use estimations of migration from the UN Global Migration Database and from the bilateral remittances from the World Bank. In order to deal with double directed edges the net remittance and net migration network are used. They define the net migration networks as $G_M = (V, E_M)$ where V are the countries and the edges are the net migration flow, with direction to the country that receives more. The information of the directed matrix is stored in an incidence-weight matrix $W = [w_{ab}]$ where m_{ij} is the migration flow from country i to country j, then

$$w_{ab} = \begin{cases} m_{ab} - m_{ba} \; if \; m_{ab} > m_{ba} \\ 0 \; otherwise \end{cases}$$

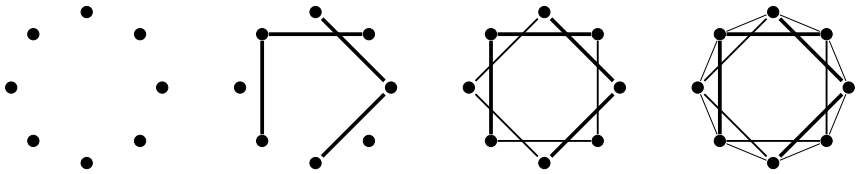
In the case of remittances something similar is done where $G_R = (V, E_R)$ is the weighted remittance network and is defined in a similar maner as the weighted remittance network, V are the vertices, nottice they do not change as

the countries are the same, and the edges are E_R.

Now that it has been stablish how the network is build and the characteristics of the information it has, it is time to discuss the methods of filtration used over the clique complex to build the persistent homology and why they are relevant for this type of network. According to [10] doing a filtration changing the clique structure by changing the path lenght of the connected subgraph does not give any relevant information, hence another type of filtration is needed. This authors suggest another approach using the weight of the edges to create a filtration in which given the adjacency matric $[w_{ab}]$, ε is allowed to vary between (min w_{ij},max w_{ij}) and consider a sequence of graphs, such that the network at ε contains all links with $w_{ij} > \varepsilon$. This filter creates an ordered clique structure, such structure is a series of graphs in which $G_{w_{max}} = G_1 \subset G_2 \subset \cdots \subset G_{w_{min}}$, in this sequence it is now possible to compute the homology and observe the sequence of homological groups.

As both remittances and migration deal with flow patterns, the simplicial complex they build on top of the data is a directed clique complex. This type of complex is done by definining simplices of dimension k to be directed $k+1$ simplices, where a k simplex represents a group of $k+1$ countries with pairwise interaction between any each member. The flow structure should be able to be characterized a streaming from an unique source to a unique sink, in order to be able to draw the directed $n-$clique this is one of the main difference with this approach and path homology approach.

In this step when the authors have defined topological object from which they can extract the homology groups and hence introduce a filtration and compute the persistent homology. As said before building the filtration and computation of persistent homology are influenced by the fact that the authors want to study directed weighted networks. For this type of networks the approach taking complex of the $n-$clique computing its homological grous and then for the complex of the $n+1-$clique, which is a method that somewhats resembles the filtration introduced in the past sections, does not seem to add any relevant information[10]. The authors chose to follow the methods discused in [10] with some modifications to cover the fact that both the migration network and the migration network are directed networks. The main difference of this approach is that the filtration is going to be a function over net weights of the edges E_R and E_M, in 4.3 illustrates the idea of how the filtration works in the edge-weighted network. Such a function should allow us to produce the sequence of subgraphs $\{G_0, \cdots, G_N\}$ and the maps $i : G_i \to G_{i+1}$ needed for the study the of the persistence of homological features in the sequence.



(a) G_0 (b) $G_0 \hookrightarrow G_1$ (c) $G_1 \hookrightarrow G_2$ (d) $G_2 \hookrightarrow G_3$

Figure 4.3

The filtration used to build the complexes in this work is the max-to-min weight filtration, for this it is important to do a transformation on every nonzero element of $[w_{ab}]$:

$$f(w_{ij}) = (max([w_{ab}] + 1) - w_{ij})$$

and induce the filtration of clique complexes of transformed values. This transformations permits to filtrate from the largest weight in he network to the smallest, hence persistence identifies the most relevant flow patters in magnitude. In this transformation cycles with larger weights are born earlier, thus even if they get killed by lower weighted cycles they have larger persistence, following the weight rank clique filtration introduced in [10].

In addition to the characterization done by the persistent homology the authors add another classification in the barcodes of the second and first dimensional homology. This characterization is done by taking advantage of the fact that homologycal cycles correspond to actual subnetworks, what they do is to compute the standard deviation of all the edge weights present in the oldest linear combination of simplices that surround a topological features. What this allows is to further classify cycles of the same dimension, according to the authors it is useful for distinguishing cycles that are born at a later stage in the filtration. The complete network characterization with the classification based on the bar codes and an example of the migration and remittances cycles in both one dimension and 2 dimensions can be seen in 4.4

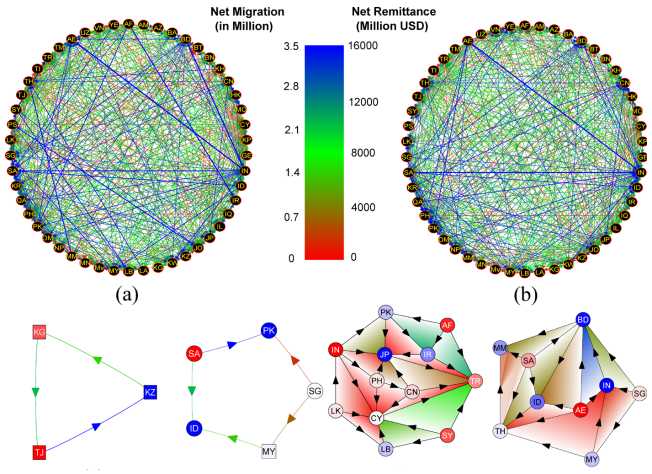


Figure 4.4: Complete network characterization and examples from [11]

For the migration network this work allowed to recognize that individual pairwise migration between countries was relatively small compared with the total flux, because the 0-dimensional homology had relatively large bars in the barcodes. While the 1-cycle reveal a broader picture of countries with lot of inmigration like Saudi Arabia or countries with emigration like India, in some cycles they are revealed to be transitory countries. The 3 cycles on the other way illustrate the effect of a more complex migran and remittance effect of a 3-country flow in the local setting. They find 61 distinct 1-cycle generators throughout the asian migration network, that where born late in the filtration, most of them (except for four) had developing countries as its origin and a developed country as its sink, and they found only one cycle was a circuit. While they find 2 perpetual 2-cycles in the asian migration network ((c) and (d) of 4.5), they are also able to characterize the entire asian migration network with two spheres. In the remittance network as in the migration network the most persistent 1-cycles ((a) and (b) of 4.5) where those with origin in a high income nation and sink a low income nation, they also find that one of the most persistent 1-cycles in the remittance barcode is a copy of the most persistent 1-cycle for the net migration network, showing the robustness of their result.

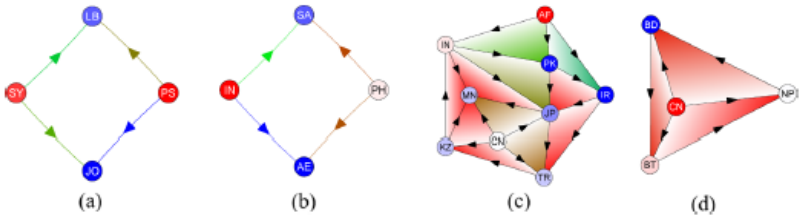


Figure 4.5: Persistent 1 cycle and perpetual 2 cycle from[11]

Chapter 5

Conclusion:

In this work it has been show how what is Persistent Homology and how it can be calculated. A little bit of its derivations has been exposed and an example done in a space with known geometry in order to compare it with homology. From this and the application the possibilities derived from this tool are evidence, particularly important for social sciences and complexity theory are the applications for networks and weighted networks. This can be seen in how Persistent Homology uncovers relationships that given the complexity of data where not easy to see before and allows a further analysis of the data an its behavior.

Bibliography

[1] F. Chazal and B. Michel, "An introduction to topological data analysis: Fundamental and practical aspects for data scientist." https://arxiv.org/abs/1710.04019v1, October 2017.

[2] N. Otter, M. A. porter, U. Tillman, P. Grindrod, and H. A. Harrington, "A roadmap for the computation of persistent homology," *EPJ Data Science*, vol. 6, no. 17, 2017.

[3] J. R. Munkres, *Elements of Algebraic Topology*. Perseus Publishing, 1984.

[4] H. Edelsbrunner and J. L. Harer, *Computational topology: An introduction*. American Mathematical Society, 2010.

[5] T. K. day, "Computational topology and data analysis," February 2017.

[6] B. T. Fasy, J. Kim, F. Lecci, C. Maria, D. L. Millman, and V. Rouvreau, *Introduction to the R package TDA*. Carnegi-Mellon University, Pittsburgh, Pensilvania, 2018.

[7] C. Brunson, "Sampling uniformly from an embedded klein bottle," 2019. Accessed: 2020-05-23.

[8] M. E. Aktas, E. Akbas, and A. E. Fatmoui, "Persistence homoloy of networks: Methods and applications," *Applied Network Science*, vol. 4, no. 61, 2019.

[9] S. Chowdhury and F. Mémoli, "Persistent path homology," in *Proceedings of the 2018 Annual ACM-SIAM Symposium on Discrete Algorithms* (A. Czumaj, ed.), 2018.

[10] G. Petri, M. Scolamiero, and F. Vaccarino, "Networks and cycles: Persistent homology approach to complex networks," in *Proceedings of the European Conference on Complex Systems 2012* (T. Gilbert, M. Kirkilionis, and G. Nicolis, eds.), pp. 93–99, Springer,Cham, 2013.

[11] P. S. Ignacio and I. K. Darcy, "Tracing patterns and shapes in remittance and migratin networks via persistent homology," *EPJ Data science*, vol. 8, no. 1, 2019.

[12] D. Jungnickel, *Graph, Networks and Algorithms*. Springer-Verlag Berlin Heidelberg, 2013.

[13] D. B. West, *Introduction to Graph Theory*. Prentice Hall, 2001.

[14] S. Wasserman and K. Faust, *Social Network Analysis: Methods and applications*. Cambridge University Press, 1994.

[15] G. Palla, I. J. Farkas, P. Pollner, I. Derenyi, and T. Vicsek, "Directed network modules," *New Journal of Physics*, vol. 9, no. 186, 2007.

[16] A. E. Sizemore, E. A. Karuza, C. Giusti, and D. S. Bassett, "Supplementary information to: Knowledge gaps in the early growth of semantic feature networks," *Nature Human Behavior*, vol. 2, 2018.

[17] F. Chazal, "Homology and topological persistence," January 2010.