Applied Topology

Network Theory and Persistent Homology

Abstract

In these notes we will explore ideas in applied topology, with a focus on persistent homology. The goal of these notes is to quickly capture the main idea behind and how they are being used in practice. Proofs are largely avoided but not completely absent. This is done so someone without a math background is not deterred from following along. Most sections of the notes have an orange box associated with them which summarize the key ideas in a very high level manner. Pink boxes are used to highlight examples which largely focus on biological applications. These notes are still a work in progress and will be updated periodically.

1 Graphs

In biology we often like to study relationships between different kinds of things such as proteins, cells or species. Over the years network theory has played a central role in studying questions that we may have about how things are related. A graph or network, is simply a collection of points connected by lines. In biological networks, the object of interest tends to be molecules like DNA, RNA, proteins, cells, or species and we draw a line between these points if they have some kind of relationship.What makes networks so incredibly useful is their simplicity and flexibility to take on almost any situation. Before looking at a formal definition of network, consider some of how networks are used to solve problems in biology.

Example 1.1. ***Protein-Protein Interaction Networks*** *consist of proteins as the vertices and the edges denote some kind of interaction between the proteins in the physical environment.*

Example 1.2. ***Gene Regulatory Networks*** *consist of genes as the nodes. An edge from gene A to gene B indicates that the protein product of A regulates the expression of gene B.*

Example 1.3. ***Sequence Similarity Networks*** *captures the sequence similarity between proteins or genes. Here, the vertices are the genes or proteins and the edges denote the similarity between two given nodes.*

Example 1.4. ***Gene Co-expression Networks*** *consist of genes as nodes and the edges between them denote co-expression.*

Definition 1.5. A graph is a pair $G = (V, E)$, where V is a set of vertices and E is a set of paired vertices called edges. A graph is connected if there is path from any point to any other point. A graph is complete if every pair of distinct vertices is connected by a unique edge.

But simply being able to represent a problem as a network is not useful in itself, we want to be able to gain insight into our problem. For example, are certain vertices more important than others? Are some points closer to each other than others? Are there any clusters? Graphs have some general properties that are helpful in answering questions we might have.

Definition 1.6. The degree of vertex i, denoted d_i, is given by the number of edges adjacent to it. In a directed graph, we can further divide degrees into indegree and outdegree. The indegree is the number of edges pointing to the said vertex, whereas the out degree refers to the number of edges leading away from the vertex.

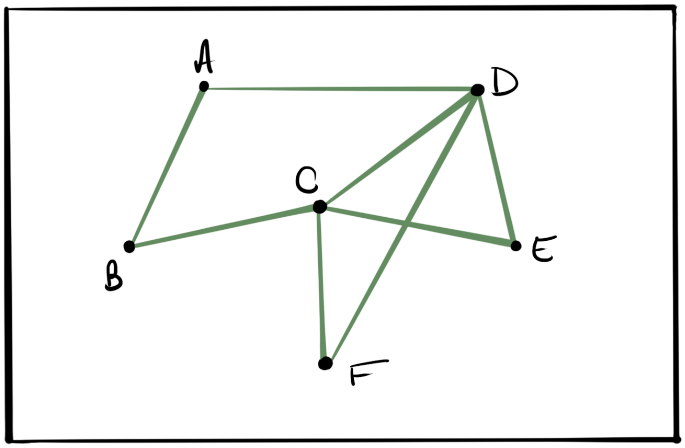
Definition 1.7. Density is the ratio between the number of edges in a graph to the the number of possible edges in the graph. A graph is called dense if $E \simeq V^k$, where $1 < k < 2$ and sparse if $k \leq 1$.

Definition 1.8. The clustering coefficient measures the ability of a vertex to form tight communities or clusters. The clustering coefficient is defined as $C = \frac{2E}{k(k-1)}$, where k is the degree of the vertex and E is the number of edges between the k neighbors.

Definition 1.9. The distance between two nodes is the length of the shortest path between them.

Certain vertices in our graph maybe more connected and thus can play an oversized influence on the network. These types of nodes, help us answer questions that we may have like: Are there any central hubs in the network? Which node connects different communities in our network? To answer these questions we can use the notion of centralities.

Definition 1.10. The degree centrality of vertex is simply the degree for the said node. The closeness centrality is defined as $C_c = \frac{1}{\sum d_i}$. The betweenness centrality is defined as $C_b = \frac{a_{xy}(i)}{a_{xy}}$, where a_{xy} is the total number shortest nodes and $a_{xy}(i)$ is the number of those paths which pass through the node i. The eccentricity centrality is geven by $C_e = \frac{1}{\max(d(i,j))}$.



Example 1.11. *The above graph can be represented by the following adjacency matrix*

$$G = \begin{bmatrix} 0 & 1 & 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 1 & 1 & 1 \\ 1 & 0 & 1 & 0 & 1 & 1 \\ 0 & 0 & 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 & 0 & 0 \end{bmatrix}$$

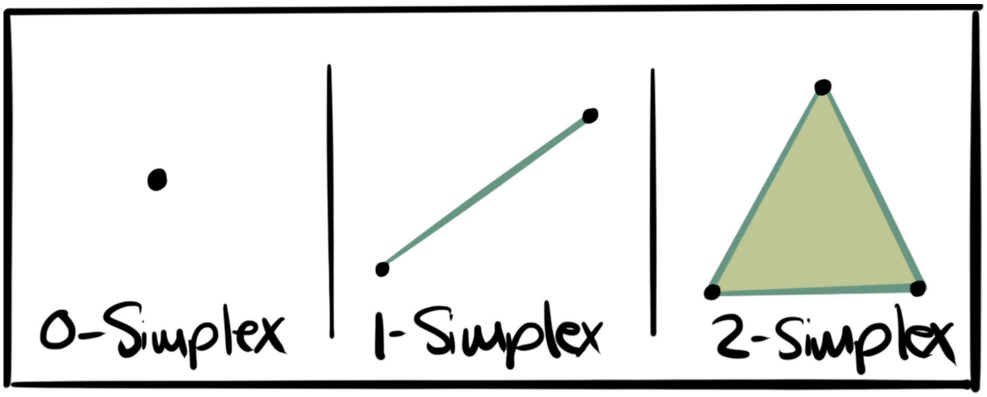
In this graph, vertex C has a degree of $d_C = 4$, where as $d_A = 2$. The distance between these two points is the shortest path between them and hence $d(A, C) = 2$. If we focus on C, it has 4 neighbors (as indicated by the degree earlier) and so the maximun number of edges between its neigher would be $\frac{4(4-1)}{2} = 6$ but only two exist (between D and E, D and F). Therefore, the clustering coefficent is $\frac{2}{6} = .33$.

2 Spaces and Complexes

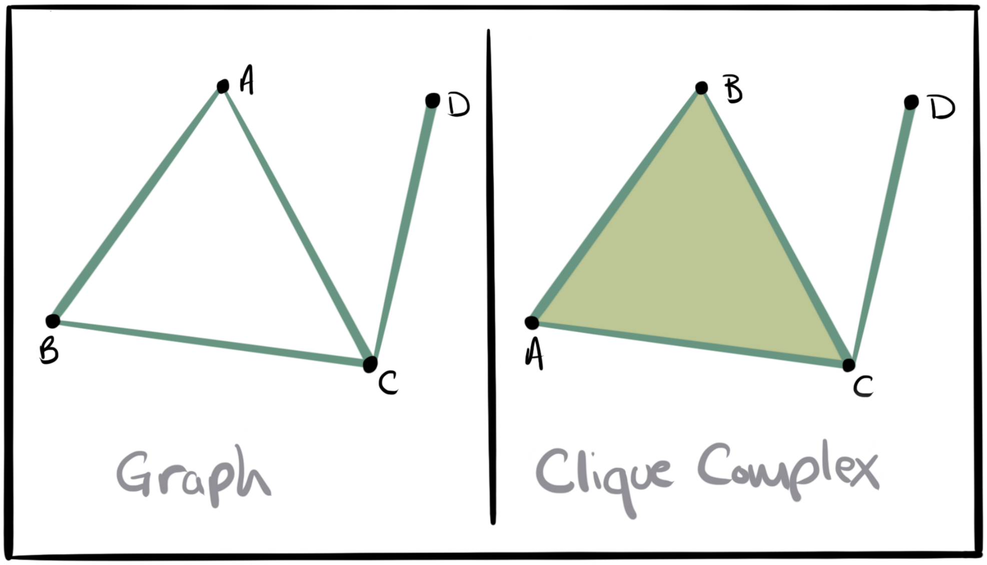
Although graphs are useful in many applications, we often times want to capture more general relationships. For example, consider a scenario where three friends go to lunch. If we were to represent this by a graph and each friend went to lunch with the other two, the best we could do is the following: $G = \{A, B, C, AB, BC, AC\}$. But, what if they all went to lunch together? This graph is not able to capture that relationship but if we could simply fill the triangle in to denote that relationship, our problem would be resolved. In other words, $G = \{A, B, C, AB, BC, AC, ABC\}$. One way to do this is through simplicial complexes.

Definition 2.1. Given $k+1$ points in $\mathbb{R}^d$, we say that $u_0, u_1, \cdots, u_k$ are affinely independent if and only if k vectors $u_i - u_0$ are linearly independent. A k-simplex is the convex hull of $k+1$ affinely independent points denoted $\sigma =$ conv$\{u_0, u_1, \cdots, \boldsymbol{u}_k\}$

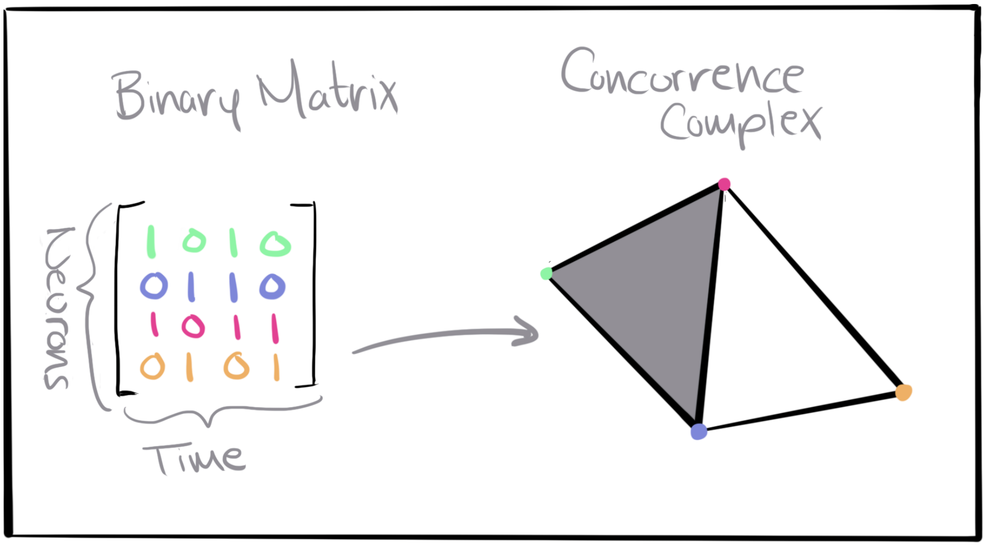
This is a bit of an abstract definition. One way to visualize what is happening is to think of our $k+1$ points as distinct pegs. Then the convex hull is like putting a rubber sheet over these pegs. What this allows us to do is to define points, lines, triangles, tetrahedron, and so on. A simplical complex is just a collection of these shapes glued together to form a more complex structure.



Definition 2.2. A simplicial complex is a finite collection of simplices k such that if $\sigma \in k$ and $\tau \in \sigma$ then $\tau \in k$. The dimension of a simplicial complex K is the the maximum of the dimensions of all of its simplices and we denote it $\dim(K)$. A graph is also called a 1-dimensional simplical complex.



Example 2.3. *One way to get a simplical complex is to first build a graph and then replace all complete subgraphs with a simplex. Consider the graph in the figure above, $G = \{A, B, C, D, AB, BA, CA, CD\}$. The vertices A, B, C form a complete graph and so we can replace this with the simplex ABC. Replacing all complete graphs forms a clique complex. In our example, G, turn into the clique complex $X(G) = \{A, B, C, D, AB, BA, CA, CD, ABC\}$*



$$\begin{bmatrix} 1 & 0 & 1 & 0 \\ 0 & 1 & 1 & 0 \\ 1 & 0 & 1 & 1 \\ 0 & 1 & 0 & 1 \end{bmatrix}$$

Example 2.4. *In neuroscience, we often times want to study coactivity. For example, we might ask questions like: What neurons are firing at the same time? What regions of the brains show similar activation patterns? A concurrence complex allows us to answer these questions by capturing the relationship between two variables. We can encode this relationship in a binary matrix where the row are one variable and the columns another. Non-zero entries correspond to the relationship between the two variables.*

Example 2.5. *Sometimes, the relationships we have between our objects of interest due not meet the full definition of a simplical complex that all σ is a simplex and $\tau \subset \sigma$, then τ is also a simplex. If this is a case, the structure is called hypergraph and are harder to work with computationally. To get around this, we can take the complement of the our original structure.*

Definition 2.6. The c-vector of K is the vector $c_K = (c_0, c_1, \cdots, c_i)$ where c_i is the number of simplices of dimension i for all $0 \leq i \leq \dim(K)$.

Definition 2.7. Let $\tau, \alpha \in K$ with $\tau \subseteq \alpha$. Then we say that τ is the face of α and α is the coface of τ. Furthermore, we define $\dim(\alpha) - \dim(\tau)$ to be the codimension of τ with respect to α.

Suppose we are given some collection of points. How do we know if two points or a cluster of points is near each other? The simplest way for us to answer this question is by taking out a ruler and measuring the distance between them. In fact, we can endow the entire space where the points exist with a mathematical structure called a metric space.

Definition 2.8. A set M along with a metric function d is called a metric space is the following holds

1. $d(x, y) = 0$ if and only if $x = y$
2. $d(x, y) = d(y, x)$
3. $d(x, z) \leq d(x, y) + d(y, z)$
4. $d(x, y) > 0$ for all $x, y \in M$

Example 2.9. *The most important metric space is the Euclidean space $\mathbb{R}^n$ which is a collection of n-tuples $X = \{(x_1, x_2, \cdots, x_n) | x_i \in \mathbb{R}\}$ equipped with the standard distance metric*

$$d(X, Y) = \sqrt{\sum_{i=0}^{n} (x_i - y_i)^2}$$

Example 2.10. *Create an alphabet denoted Σ. We will call its elements letters. Then the Hamming distance between two words of length n is simply the number of positions that where the letters differ. The Hamming distance is an important example of metric spaces in coding theory and can also apply to biology. For example, let $\Sigma = \{A, C, T, G\}$, then $d_H = \{ACGT, ACAA\} = 2$*

Example 2.11. *Gene expression can be considered to be a point in $\mathbb{R}^n$ and a typical metric that is applied is the Pearson Correlation.*

Metric spaces are a very useful tool but they are too strict. To get around this rigidity, we can loosen our requirements for what it means to be near by completely getting rid of the notion of a distance. These two structures that we explore end up being central to many ideas in not only mathematics but also in statistics and biology.

Definition 2.12. Let X be a set and $\tau \in X$. Then the pair (X, τ) is a topological space if

1. $\phi, X \in \tau$
2. infinite unions of the elements of τ are contained within τ
3. finite intersections of the elements of τ are in τ

We call τ a topology and the elements of this topology are called open sets.

The notion of a metric space is much more down to earth but topological spaces will allow us the flexibility to study spaces in ways that the former will not. But we do not have to choose between the two and in fact the types of topological spaces we will be concerned with will be metrizable spaces. Theses are topological spaces which naturally arise from metric spaces.

A metric space measures nearness between points by measuring there distance. We abstract this to a **topological space** which is a more general structure that measures nearness for a collection of points. These points have some sort of connections which make them distinct from a different set of points that also have connections. Although topological spaces are a nice structure, they are not practical in an applied setting where we are chiefly concerned with computations. One way to work around this is to restrict ourselves to only discrete structures and glue them together to make larger structures. This is the idea behind a simplicial complex. A simplicial complex is collection of points, lines, and triangles glued together to represent some underlying topological space. These structures are called graphs when we restrict ourselves to only lines and points, and are a common tool in applications. Furthermore, simplicial complexes are a generalization of a graph which easily lends them to applications.

3 Simple Homotopy

We have defined several structures up to this point and how to create these structures from data along with some numbers that can be associated with them. The next question, we want to ask is how can we tell when two things are the same? One way to do this is to start with a simplicial complex and then morph into another complex. If this is possible, then these to simplicial complexes are the same. We fomralize this below.

Definition 3.1. Let K be simplicial complex and consider a pair of simplices $\{\sigma^{(p-1)}, \tau^p\}$ where $\sigma^{(p-1)}$ only has one coface, τ. Then the simplicial complex $K - \{\sigma^{(p-1)}, \tau^p\}$ is called an elementary collapse, denoted $K \searrow K - \{\sigma^{(p-1)}, \tau^p\}$. Similarly, an elementary expansion, denoted $K \nearrow K \cup \{\sigma^{(p-1)}, \tau^p\}$, is the addition of such a pair. For both cases, the pair $\{\sigma^{(p-1)}, \tau^p\}$ is called a free pair.

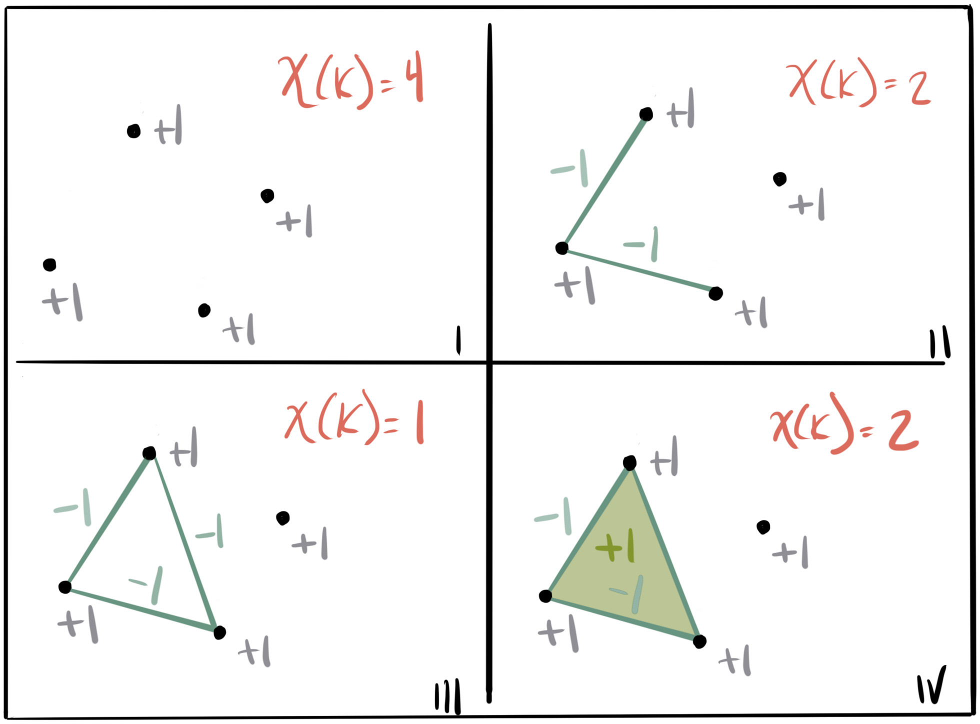
Definition 3.2. Let K and L be simplicial complexes. We say that K and L have the same simple homotopy type, denoted $K\ L$, if there is a series of elementary collapses and expansions that can take us between the two.

Having a notion of sameness is nice but it is not practical to do a series of collapses and expansions to determine this. So, the immediate question that we want to answer now is: What things do not change between two complexes that are simple homotopy equivalent?

Definition 3.3. Let K and L be simplicial complexes and let α be a function that associates a real value to the simplicial complexes. We call α an invariant if $\alpha(K) = \alpha(L)$ whenever $K\ L$.

Definition 3.4. Let K be a simplicial complex and $c_i(K)$ be the c-vector. Then we define the Euler characteristic of K to be

$$\chi(K) = \sum_{i=0}^{n} (-1)^i c_i(K)$$



Example 3.5. *A classic example of an invariant is the Euler characteristic. To understand this consider a collection of points. We begin with n things but if we add an edge, we are left we only $n-1$ things. So it seems fair to say that, if we have one dimensional points, we count that as $+1$ but soon as we connect two points by bringing in an edge, we count it as -1. But soon as we create a triangle with a hole, notice that we do not lose any components. If we persevere on, we see that soon as the hole is filled, we can give it a weight of $+1$, which undoes our previous 'mistake'. This process is shown in the figure above, where we start with four points and so have 4 components. When we add add two edges, we are left with 2 components but when we complete the triangle by adding a third edge, by our rule, we have only 1 component. But we can clearly see that there are two components. This discrepancy is remedied in the fourth panel where the triangle is filled in and our result is updated to 2. This is the Euler characteristic, $\chi(K) = 4 - 3 + 1 = 2$*

Theorem 3.6. *If $K \sim L$, then $\chi(K) = \chi(L)$.*

Definition 3.7. A simplicial complex, K, is collapsible if there exists a series of only collapse such that

$$K = K_0 \searrow K_1 \searrow K_2 \searrow \cdots \searrow K_{n-1} \searrow K_n = \{v\}$$

Simple homotopy is a way to go from one simplicial complex to another through a series of collapses and expansions of the simplices. But this still leaves the question of what remains the same under these operations. It is not practical to find a series of collapses and expansion everytime we want to show that two complexes are equivalent. We call such things invariants and generalize counting to the Euler Characteristic.

4 Homology

From looking at our examples from simple homotopy, it may have occurred to some that there seems to be some preservation of the holes under collapses and expansions. Is this true? If so can we distinguish different spaces by counting the number of holes in a simplicial complex? This is the idea behind homology where we are able to count not only the number of holes but also the type of hole. To be able to do this we need the notion of a vector space.

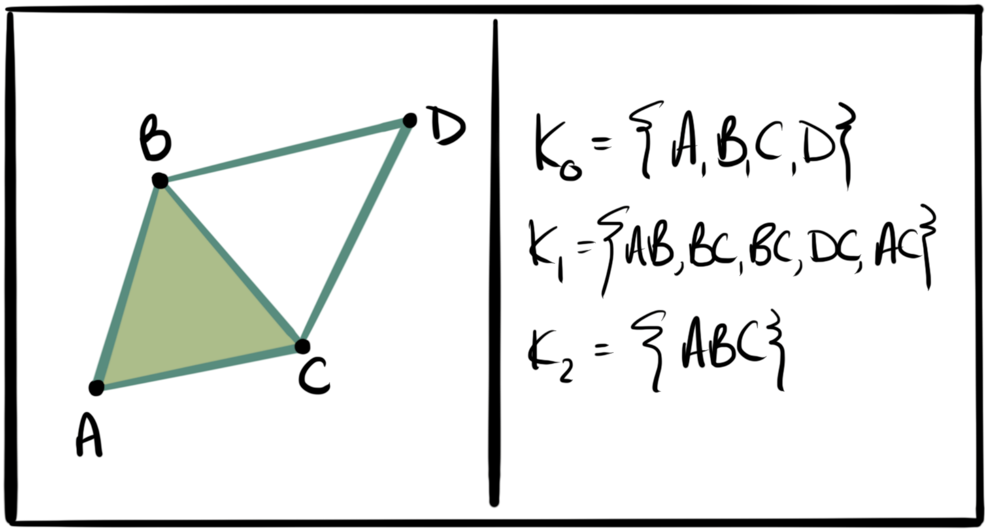
Definition 4.1. A field is a set $\mathbb{F}$ of numbers such that if $a, b \in \mathbb{F}$, then $a + b, a - b, ab, a/b$ are also in $\mathbb{F}$.

Definition 4.2. A vector space is a set V over a field $\mathbb{F}$ paired with the operations of vector addition and scalar multiplications such that the following axioms are satified:

1. $(u + v) + w = u + (v + w)$ for all $u, v, w \in V$
2. There is a vector denoted $\vec{0}$ such that $u + \vec{0} = u$. We call this the zero vector.
3. For all vectors $v \in V$, there is another vector in V such that $v + (-v) = \vec{0}$.
4. $(ab)u = a(bu)$ for all $a, b \in \mathbb{F}$ and $u \in V$.
5. $(a + b)u = au + bu$ and $a(u + v) = au + av$ for all $a, b \in \mathbb{F}$ and $u, v \in V$.
6. $1u = u$ for all $u \in V$.

It may not be immediately obvious why we need vector spaces to characterize holes in topological spaces. Consider any shape with a single hole. In each of these, it seems like the sequence of edges and vertices seem to define the hole. But if we were to consider any sequence of simplices, they should detect exactly one hole but this is not necessarily the case. Some of these sequences may be duplicates of other sequences and others may not detect holes at all. The notion of a vector space over allows us to work around these issues.

Definition 4.3. Let $X = \{e_1, e_2, \cdots, e_3, e_n\}$ be a set of n distinct elements. The vector space, denoted $\mathbb{K}^n$ over the field $\mathbb{F}_2$ generated by X is given by the linear combination $\mathbb{K}^n = \{c_1 e_1 + c_2 e_2 + \cdots + c_n e_n : c_i \in \{0, 1\}\}$. The elements of set X are called the basis elements and n is the dimension of the vector space.



Example 4.4. *Consider the simplicial complex shown below in figure 4. The c_i vectors listed in the image, generate the following three vector spaces:* $\mathbb{K}^4 = \{A, B, C, D, A + B, A + C, A + D, B + C, B + D, C + D, A + B + C, A + B + D, B + C + D, A + C + D, A + B + C + D\}$

Definition 4.5. Any element of a vector space generated by a collection of simplices is called a chain.

Definition 4.6. A linear transformation between two vector spaces is a map $A : \mathbb{K}^n \longrightarrow \mathbb{K}^m$ such that for any two elements $v, v' \in \mathbb{K}^n$ the following conditions hold:

1. $A(v + v') = A(v) + A(v')$
2. $T(\alpha v) = \alpha A(v)$ for any scalar α.

Definition 4.7. The kernal of a linear transformation is defined by all elements that maps to the zero vector or in other words,

$$\ker(A) = \{x \in \mathbb{K}^n : A(v) = 0\}$$

. The dimension of of the kernal is called the nullity and denoted null(A). The image of the linear transformation A between two vector spaces is given by

$$\text{Im}(A) = \{y \in \mathbb{K}^m : \exists x \in \mathbb{K}^n \text{ such that } A(x) = y\}$$

. The dimension of the image is called the rank of A and denoted rank(A).

Theorem 4.8. *Let A be a linear transformation between a vector space of size n to one of size m. Then rank(A) + null$(A) = n$.*

Since all linear transformations can be represented as matrices, putting A into row echelon form, does not change the rank. In fact, if A is in row reduced form, then the rank of A is the number of non-zero rows.

But notice that in our example, we only have one hole, yet our vector space would count two. To work around this we need to develop the notion of a boundary operator which will allow us to take a simplex and compute it boundary.

Definition 4.9. Let K be a simplicial complex. Then partition K into sets of simplices of size i, denoted by K_i. The sequence of vector spaces generated by these partitions along with a linear transformation between them is called a chain complex.

The linear transformation is called a boundary operator and defined as follows.

Definition 4.10. Let $\sigma \in K_m$ and $\sigma = \sigma_{i_0}\sigma_{i_1}\cdots\sigma_{i_m}$. For $m \geq 1$, the boundary operator $\partial_m : \mathbb{K}^{c_m} \longrightarrow \mathbb{K}^{c_{m-1}}$ is given by $\partial_m(\sigma) = \sum_{0\leq j\leq m} \sigma - \sigma_{i_j} = \sum_{0\leq j\leq m} \sigma_{i_1}\sigma_{i_1}\cdots\hat{\sigma_{i_j}}\cdots\sigma_{i_m}$, where $\hat{\sigma_{i_j}}$ is the simplex being excluded.

Definition 4.11. The i-th unreduced homology of K is the vector space

$$H_i(K; \mathbb{F}_2) = \mathbb{K}^{\text{null}(\partial_i) - \text{rank}(\partial_{i+1})}$$

The i-th Betti number is given by

$$b_i(K; \mathbb{F}_2) = \text{null}(\partial_i) - \text{rank}(\partial_{i+1})$$

Often time we will shorten Betti numbers to just $b_i(K)$. The betti number is going to play an important role as an invariance when we begin to consider applications. There is also a nice relationship between the Euler characteristic and Betti numbers.

Theorem 4.12. *Let K and L be simplicial complexes. If K L, then $b_i(K) = b_i(L)$.*

5 Persistence Homology

We have spent the last few sections on developing the notion of nearness and sameness but how can we apply this to problems in science? To answer this question, we need to first figure out how to build simplicial complexes from data.

In real world applications, we are often give a finite collection of points in $\mathbb{R}^n$ and so we can equip this with some distance function and get a finite metric space. If we want to be able to leverage the tools we discussed in earlier sections, we need to somehow abstract this to a topological space. This is formalized in the below.

Definition 5.1. Let $X \subset \mathbb{R}^n$ be finite subspace and fix $\epsilon > 0$. The union of balls is the union

$$\bigcup_{x \in X} B_\epsilon(x) \subset \mathbb{R}^n$$

As discussed earlier, this is not a useful structure for computation. We want to develop this idea of a union of balls in terms of a simplicial complexes.

Definition 5.2. Let $X \subset \mathbb{R}^n$ be finite subspace and fix $\epsilon > 0$. The Cech complex $C_\epsilon(X, d_X)$ is an abstract simplicial complex where the vertices are the points in X and a k-simplex is created when a subset of points of X satisfy

$$\bigcap_i B_\epsilon(v_i) \neq \emptyset$$

The cech complex allows us to assign a simplicial complex to a metric space but determining when an intersection of ϵ-balls is non empty is not an easy task for higher dimensions. To work around this, we can use the idea that a graph is a one dimensional simplicial complex to create a computational more efficient simplicial complex for a metric space called the Vietoris-Rips complex.

Definition 5.3. Consider a finite metric space (X, d_X) and fix $\epsilon > 0$. The Vietoris-Rips complex $VR_\epsilon(X, d_X)$ is an abstract simplicial complex where the vertices ar the points in X and a k-simplex is created when $d_X(v_i, v_j) \leq 2\epsilon$ for all $i \leq i, j \leq k$.

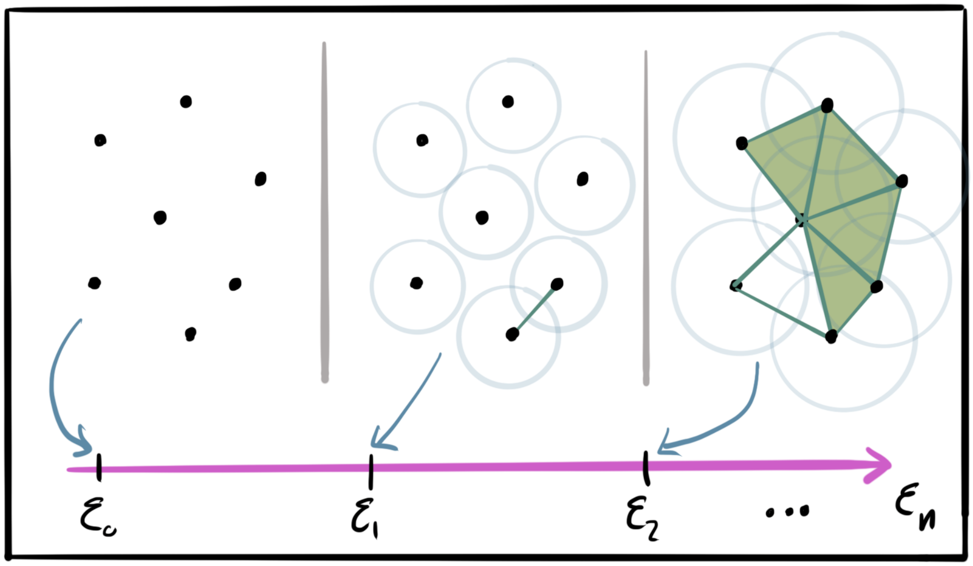
Lemma 5.4. *Let $X \subset \mathbb{R}^n$ be finite subspace and fix $\epsilon > 0$. The the following inclusion holds*

$$C_\epsilon(X, d_X) \subseteq VR_\epsilon(X, d_X) \subseteq C_{2\epsilon}(X, d_X)$$

This leads us naturally to another question. Given a collection of points, what is the correct choice of ϵ? The answer to this is that we do not have to choose a single value and this is where the first half of the name persistent homology comes from. The general idea is that we start with a collection of points and choose some interval of values $[\epsilon_0, \epsilon_n]$ and create a sequence of complexes called a filtration.

Definition 5.5. Let K be a simplical complex. Then a filtration of K is a sequence of complexes that begin with the empty complex and end with the completed complex,

$$\emptyset \subset K_1 \subset K_2 \subset \cdots K_n = K$$



The idea of a filtration is key to computing topological features in our data. Given this sequence of simplicial complexes, we can compute a the homology groups at each step.

Definition 5.6. A barcode is a multiset of intervals, $[x, y) \in \mathbb{R}$, which measures the the length a feature persists through each filtration.

Example 5.7. *Can we detect holes in logic? A paper by Tymochko et. al tries to do just that by using persistence homology to compare three statements: A valid statement, A circular invalid statement, and a random statement where the structure was not obvious. This text was embedded into a word vector point cloud. This was was turned into a one dimensional time series by computing the dot product of each word vector with a fixed random vector. To capture the dynamics of this time series, time delayed embedding is performed which can then be used to calculate the persistent homology. The resulting persistence diagram identifies two circular features for the valid and invalid argument which could not be identified using traditional methods.*

We can generalize the concept of persistence to a setting where the inclusion of a filtration do not go in one direction. For example, consider a metric space where the points have some sort of order. Let this be denoted by $X = x_1, x_2, x_3, \cdots$ and let X_k denote the subset of X consisting of the first k points. If we were too measure the distance between the subsets and X, then the subset X_{i+1} will be at least as close to the original set as X_i.

Persistence homology is the process of taking a collection of points and defining a way to create a sequence of simplicial complexes and computing the homology at each step.

6 Topological Descriptors and Statistics

7 Machine Learning