

Topological Data Analysis:
the algebraic topology of very large data sets?

Timothy Porter
School of Informatics,
University of Wales Bangor,
Bangor,
Gwynedd, LL57 1UT,
Wales, U.K.
Email: tporter@bangor.ac.uk

Plan.

- Simplicial Complexes, Triangulations, etc.
- Computational Homology.
- Philosophical interlude and non-polyhedral spaces.
- Čech and Vietoris.
- Computational Substitutes.
- Rips, Witness and α .
- Persistent homology.
- Where to next?

Simplicial Complexes

A **simplicial complex** K is a set of objects, $V(K)$, called **vertices** and a set, $S(K)$, of finite non-empty subsets of $V(K)$, called **simplices** such that if $\sigma \in S(K)$ and $\tau \subset \sigma$, $\tau \neq \emptyset$, then $\tau \in S(K)$.

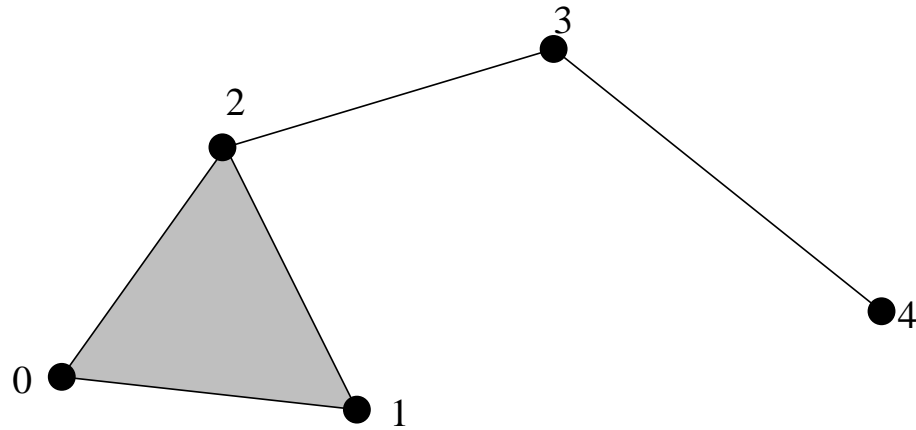
We say τ is a **face** of σ .

If $\sigma \in S(K)$ has $p + 1$ elements, it is said to be a **p -simplex**.

The set of p -simplices of K is denoted by K_p .

The dimension of K is the largest p such that K_p is non-empty.

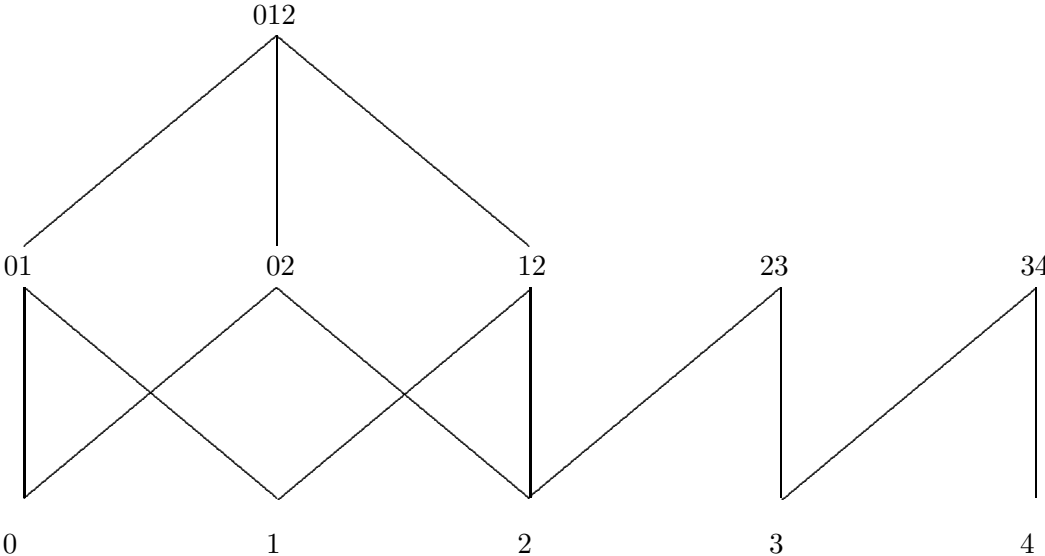
Here is an example of a simple simplicial complex:



$$V(K) = \{0, 1, 2, 3, 4\}$$

$$S(K) = \{\{0\}, \{1\}, \{2\}, \{3\}, \{4\}, \{0, 1\}, \{0, 2\}, \{1, 2\}, \{2, 3\}, \{3, 4\}, \{0, 1, 2\}\}$$

The faces of a simplicial complex form a partially ordered set under inclusion. This face poset for the above example is:



Triangulation.

A **triangulation** (K, f) of a space X consists of a simplicial complex K and a homeomorphism $f : |K| \rightarrow X$.

We will usually confuse $|K|$ with X and so will call X , itself, a **polyhedron** in this case.

We use triangulations to ‘control’ spaces, but they are something ‘imposed on the space’ or built from the observations of the ‘space’? You find papers that start ‘Let X be a polyhedron ’ or ‘We assume the data is sampled from a smooth manifold’, yet never ask how the data might verify if that is the case. Any sample of data points will give a polyhedron in various ways, but might that polyhedron be strongly dependent on the sample? (**and what exact meaning can be given to that question?**)

Homology

Homology is an algebraic invariant that counts the topological attributes of a ‘shape’ X in terms of its **Betti numbers**, β_i .

- β_0 counts the number of components of X ;
- β_1 is the rank of a basis for the *tunnels* through X ;
- β_2 counts the number of *voids* in X , that is, the number of enclosed 3-dimensional holes.
- and so on

Homology allows precise meanings to be given to notions of ‘connectedness’ in higher dimensions. It is easier to work with computationally than the notion of homotopy, so is more often used in this area even though it does not give the ‘fine detail’ of homotopy.

The details:

Simplicial homology - measures the homology of a simplicial complex:

Given a simplicial complex K .

Preprocessing: pick a total order on the vertices of K and represent each simplex in K as an ordered list of vertices: $\sigma = \langle v_0, \dots, v_p \rangle$, with $v_0 \leq \dots \leq v_p$ in the chosen order.

(This allows us to speak of the ' k^{th} ' face $d_k\sigma$ of σ . The zeroth face is $\langle v_1, \dots, v_n \rangle$, the first is $\langle v_0, v_2, \dots, v_n \rangle$ and so on.)

Now for each $p \geq 0$ form the vector space:

$C_p(K)$ = the vector space of all formal sums $\sum_{\sigma \in K_p} a_\sigma \sigma$, with $a_\sigma \in \mathbb{R}$.

Now let

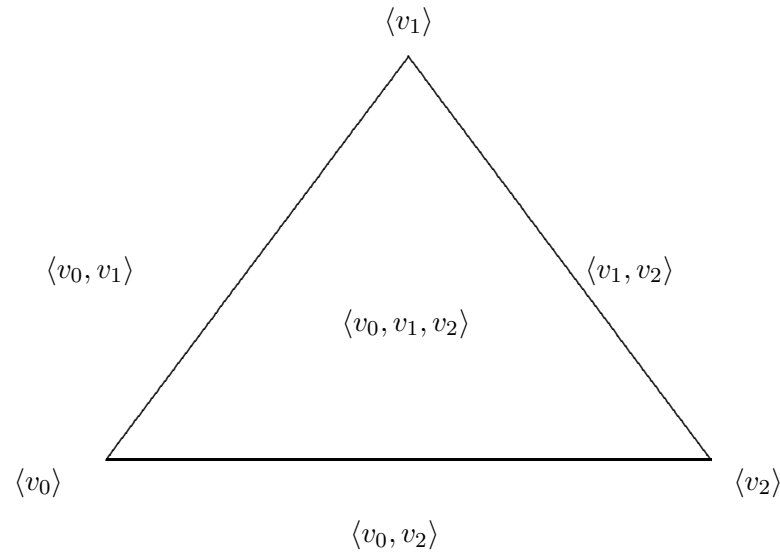
$$\partial = \partial_p : C_p(K) \rightarrow C_{p-1}(K)$$

be defined on p -simplices by

$$\partial_p \sigma = \sum_{i=0}^p (-1)^i d_i \sigma,$$

and then extended linearly to all of $C_p(K)$. This linear mapping is called the boundary map. The reason is easy to see in low dimensions.

For instance, if σ was the triangle / 2-simplex, $\langle v_0, v_1, v_2 \rangle$, we would want $\partial\sigma$ to be $\langle v_1, v_2 \rangle + \langle v_0, v_1 \rangle - \langle v_0, v_2 \rangle$, since going (clockwise) around the triangle, that cycle will be traced out:



Exercises for the audience! (If you haven't seen them before)

$\partial\partial = 0$, the zero linear mapping.

The number of components of the complex is the dimension of the vector space $H_0(K) = C_0(K)/\partial(C_1(K))$.

If K has dimension 1 and so is just a graph, $\text{Ker}\partial$ is generated by the cycles of that graph.

In general set

$$H_p(K) = \frac{\text{Ker}(\partial_p : C_p(K) \rightarrow C_{p-1}(K))}{\text{Image}(\partial_p : C_{p+1}(K) \rightarrow C_p(K))}.$$

This is the p^{th} *homology group of K with real coefficients* and the p^{th} Betti number $\beta_p(K)$ is its dimension.

(This is easily computed using basic linear algebra routines from, for instance, MATLAB, cf. PLEX¹, or CHOMP².)

¹PLEX is available from <http://math.stanford.edu/comptop/programs/plex/>.

²Computational Homology, recent book in Springer Applied Maths series.

Philosophical interlude.

Some questions and contentious contentions!:

What is the point of points?

Do spaces really have points or is that just a useful device for handling something else?

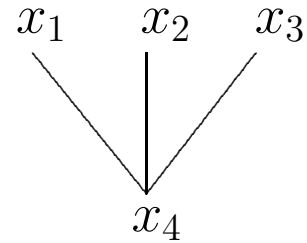
Spaces may correspond to some geometric object, but may be used just to organise data which may not be spatial in essence.

Example: Objects and attributes, formal context analysis, Chu spaces, (or just relations).

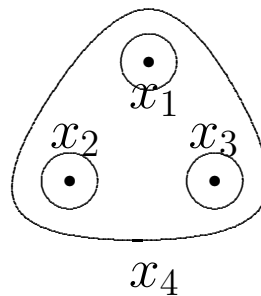
A Chu space, \mathcal{C} , is given as $\mathcal{C} = (C_o, \models_{\mathcal{C}}, C_a)$, where C_o and C_a are sets, respectively, here called the sets of *objects* and of *attributes* and $\models_{\mathcal{C}} \subseteq C_o \times C_a$ is a relation.

Example

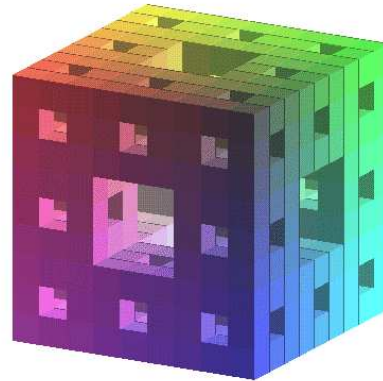
\mathcal{C}	a_1	a_2	a_3
x_1	1	0	0
x_2	0	1	0
x_3	0	0	1
x_4	1	1	1



The relation is the partial order as shown. It is the dual of the lattice of non-empty open sets of a three point discrete space:



Not all spaces are polyhedra. What would happen if we tried to analyse one of them computationally or ‘observationally’?



The Menger cube construction, shown after 2 iterations. The Menger cube is obtained by iterating this basic construction. It is not a polyhedron. Can we find some ‘homology’ for it?

Some non-polyhedral spaces occur in the study of orbits, etc. in dynamical systems theory:

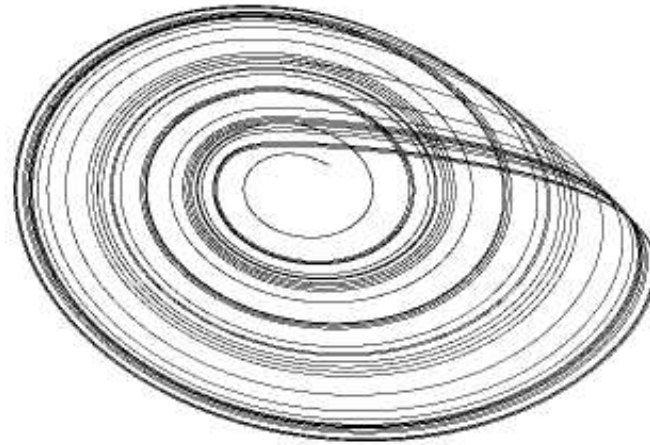
e.g. the Rössler band. This is defined by certain linked very simple differential equations, namely

$$\begin{aligned}\frac{dx}{dt} &= -(y + z) \\ \frac{dy}{dt} &= x + ay \\ \frac{dz}{dt} &= b + xz - cz \ ,\end{aligned}$$

where a , b and c are parameters. For some values of these parameters, nothing exceptional happens, but the graphics of the attractor for other values seem to show a Möbius band-like space having the interval replaced by a Cantor set.³

³ The values $a = 0.15$, $b = 0.2$, $c = 10.0$ will do.

A picture:



How can we obtain some 'homological information' from such spaces? We cannot use triangulations as they are not triangulable.

Čech and Vietoris: the observational solution from the 1920s and 1930s

Instead of a triangulation, assume we are given an open cover \mathcal{U} of our 'space' X , so \mathcal{U} is a family of open sets, U , of X and for any $x \in X$ there is some $U \in \mathcal{U}$ that contains it.

(The 'observational' idea is that we probe X and each probe can measure things in a small patch.)

To each such open covering we can attach two simplicial complexes one due to Vietoris (1927) the other to Čech (early 1930's).

Čech: the nerve of X denoted $N(X, \mathcal{U})$ or $N(\mathcal{U})$ if no confusion will arise.

- vertices, the open sets in \mathcal{U} ,

and

- simplices, those finite families of open sets in \mathcal{U} whose intersection is non-empty:

i.e. $\{U_0, \dots, U_n\} \subset \mathcal{U}$ is a simplex of $N(\mathcal{U})$ if and only if $\bigcap_{i=0}^n U_i \neq \emptyset$.

The Vietoris complex reverses the roles of points and open sets:

Vietoris complex of X denoted $V(X, \mathcal{U})$ or simply $V(\mathcal{U})$.

- vertices, the points of X itself

and

- simplices : $\langle x_0, \dots, x_n \rangle$ is an n simplex of $V(\mathcal{U})$ if there is a $U \in \mathcal{U}$ containing all the vertices $x_i, i = 0, \dots, n$.

In general, no 'spatial' context is needed (cf. extensive applications in **AI**) and any metric data is not used.

Dowker (1952) Let $R \subset X \times Y$ be a relation. (In our topological case, $X = X, Y = U$ and $xRU = x \in U$.) Any such relation determines two simplicial complexes:

(i) $K = K_R$: - the set of vertices is the set, Y ;

p -simplex of K is a set $\{y_0, \dots, y_p\} \subseteq Y$ such that there is some $x \in X$ with xRy_j for $j = 0, 1, \dots, p$.

(ii) $L = L_R$: - the set of vertices is the set X ;

- a p -simplex of K is a set $\{x_0, \dots, x_p\} \subseteq X$ such that there is some $y \in Y$ with x_iRy for $i = 0, 1, \dots, p$.

The homology of the two complexes is the same.

By refining the cover (so increasing the ‘sampling density’) you can ascribe limiting homology groups to an arbitrary X . If X is a polyhedron :

Information from triangulations



Information from open coverings

but for the case of non-polyhedra ‘inverse systems of simplicial complexes’ are used⁴. This leads on to Shape Theory, but that is not for today!

⁴These are infinite, non-computational tools but are related to the hierarchical triangulation techniques used in **GIS**.

Computational substitutes.

We have these two complexes from ‘classical algebraic topology’. They are not computationally feasible as such. Various replacements are used. They exploit the metric structure of much data.

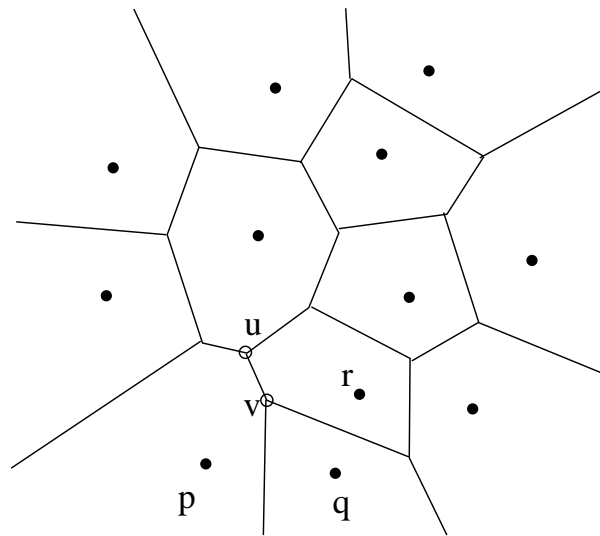
Usual assumption: *The data is sampled from some ‘idealised’ subspace X of some \mathbb{R}^n , (but both the ambient and intrinsic metrics may be used).*

We need to recall Voronoi diagrams and the related Delaunay triangulations given by the sample.

Voronoi diagrams. Let P be a set of data points in \mathbb{R}^n .

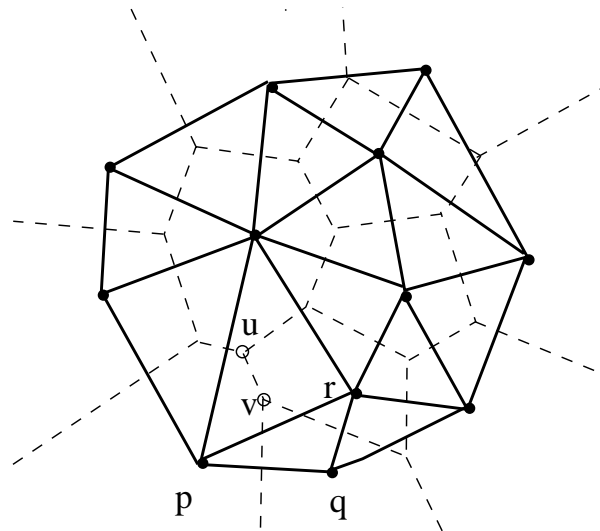
The *Voronoi diagram* of P denoted V_P is a collection of *Voronoi cells* V_p , one for each point $p \in P$, where

$V_p = \{x \in \mathbb{R}^n \mid \|x - p\| \leq \|x - q\| \text{ for any } q \in P\}$. In other words, V_p is the set of all points in \mathbb{R}^n that are closer or at least equidistant to p than to any other point in P .



Delaunay triangulation. There is an associated *dual* structure to Voronoi diagram V_P , called the *Delaunay triangulation* denoted D_P . Formally, we define D_P as a simplicial complex where

$$D_P = \{\sigma \mid \bigcap V_p \neq \emptyset \text{ where } p \text{ is any vertex of } \sigma\}.$$



Note the similarity with the classical nerve.

Simplicial Complex Approximations: Overview of the problem

Let $X \subset \mathbb{R}^n$ be a subspace and let $Z \subset X$ be a finite set of sample points. By a *simplicial complex approximation* to X , we are referring to the following situation:

1. A construction $\mathcal{S} = \mathcal{S}(Z)$ of a simplicial complex depending on Z and possibly on additional parameters, but not depending on X itself;
2. A *similarity result* or heuristic comparing X with $\mathcal{S}(Z)$ (possibly homeomorphism or some weaker, but well behaved / well controlled, notion of comparison), under reasonable conditions on Z as a sample of X , and for some choice of values for the additional parameters.

A typical additional parameter might be some notion of *feature scale* $R \geq 0$. This can sometimes be interpreted as an amount of blurring or ‘fuzziness’ applied to Z . The complex $\mathcal{S}(Z, R)$ is expected to capture large-scale geometrical features and ignore small-scale features, where ‘large’ and ‘small’ are defined in terms of R . Varying R and / or the sample Z , we hope to capture ‘qualitative’ information on the idealised X .

Often for two values $R \leq R'$, the constructions will give **nested** simplicial complexes,

$$\mathcal{S}(Z, R) \subseteq \mathcal{S}(Z, R').$$

The Čech complex. This replaces the arbitrary open cover of the nerve construction, by little open balls around data points. The radius used gives a nesting parameter, R :

- Vertex set : all data points in Z ;
- Parameter: $R > 0$, nested;
- Definition: the p -simplex $\sigma = [z_0, z_1, \dots, z_p]$ belongs to $\check{C}ech(Z, R)$ if and only if the closed Euclidean balls $B(z_j, R/2)$, $j = 0, 1, \dots, p$ have non-empty common intersection.

The significant drawback with this is its inefficiency. If k points form a cluster of diameter at most R , there is a corresponding $(k - 1)$ -dimensional simplex in $\check{C}ech(Z, R)$. This leads to large complexes when R is large, even when the underlying topological information is very simple.

The Rips complex. This is a variant of the Čech complex which is easier to calculate, but slightly more prone to clustering-related inefficiency.

- Vertex set : all data points in Z ;
- Parameter: $R > 0$, nested;
- Definition: the p -simplex $\sigma = [z_0, z_1, \dots, z_p]$ belongs to $\text{Rips}(Z, R)$ if and only if for every edge $[z_j, z_k]$, $0 \leq j < k \leq p$, we have $\|z_j - z_k\| \leq R$.

Unfortunately, in general, this may not record the homology correctly. it is also difficult to know how to chose the parameter.

The α -shape complex, $A(Z, R)$: This is due to Edelsbrunner, and is based on the Delaunay triangulation. It gives a family of complexes similar to the Čech complexes but which are considerably smaller in size.

- Vertex set : all data points in Z ;
- Parameter: $R > 0$, nested;
- If V_Z is the Voronoi diagram for Z and V_z is the closed Voronoi cell for the sample point z , then define the α -cell for z to be the convex set $\alpha(z, R) = B(z, R) \cap V_z$;
- Definition: the p -simplex $\sigma = [z_0, z_1, \dots, z_p]$ belongs to $A(Z, R)$ if and only if the α -cells, $\alpha(z_j, R/2)$, $j = 0, 1, \dots, p$ have non-empty common intersection.

Standard implementations of the α -shape complex are based on a global calculation of the full Delaunay triangulation of Z . It thus does suffer from a ‘curse of dimensionality’ with respect to the dimension n of the ambient space \mathbb{R}^n .

Persistent homology

Any of these can give a homology and Betti numbers (for that construction and that value of R): $\beta_i(Z, R) = \text{rank} H_i(\mathcal{S}(Z, R))$.

If the construction is a 'nested' one then if $R \leq R'$, we have complexes,

$$\mathcal{S}(Z, R) \subseteq \mathcal{S}(Z, R')$$

and induced maps

$$H_i(\mathcal{S}(Z, R)) \rightarrow H_i(\mathcal{S}(Z, R')).$$

Algebraically we can compute **persistent Betti numbers** $\beta_i(R, R')$ for every pair $R \leq R'$.

Interpretation

Intuitively $\beta_i(R, R')$ counts the number of i -dimensional holes in $\mathcal{S}(Z, R)$ which remain open when we thicken the complex to $\mathcal{S}(Z, R')$.

Computation available using algorithm by Edelsbrunner, Letscher and Zomorodian (2000).

Produce **bar codes** or interval graphs.

For each dimension i get a set of closed intervals above an axis parametrised by R .

An interval $[R_0, R_1]$ indicates that a i -dimensional ‘hole’ appears for the first time at $R = R_0$ and persists until $R = R_1$, when it disappears (i.e. closes up). Long intervals correspond to large holes and thus to genuine features. Small intervals indicate are regarded as ‘noise’⁵.

⁵see Carlsson and da Silva, Topological Estimation using Witness Complexes, Eurographics Sym. Point Based Graphics, 2004.

Witness complexes.

The constructions above often give reasonable ‘geometric’ information (suitable for ‘reconstructing’ the ‘space’ e.g. for visualisation). For fast computation of homology and related invariants we don’t need geometric (e.g. metric based) information and not even topological information, only homotopy information, so can attempt to thin out the constructions with that in mind.

Assume given a set Z of N data points, and a subset $L \subset Z$ of n ‘landmark points’ again in some Euclidean space \mathbb{R}^m . Give Z a total order for convenience, with L the first n points. Let $D = n \times N$ matrix of distances from landmarks to general points.

Strict witness complex : $W_\infty(D)$

- Vertex set: L , the set of landmarks.
- Edges $\sigma = \langle a, b \rangle$ is in $W_\infty(D)$ if there is a data point z_i such that $D(a, z_i)$ and $D(b, z_i)$ are the smallest two entries in the i^{th} column of D ;
- by induction on p , $\sigma = \langle a_0, a_1, \dots, a_p \rangle$ is in $W_\infty(D)$ if all its faces are in $W_\infty(D)$ **and** there is a data point z_i such that the $D(a_j, z_i)$, $j = 0, 1, \dots, p$ are the smallest $p + 1$ entries in the i^{th} column of D , in some order.

Weak witness complex : $W_1(D)$ - the ‘Rips complex version of $W_\infty(D)$!

- Vertex set: L , the set of landmarks;
- $W_\infty(D)$ has the same 1-skeleton as $W_1(D)$;
- the p -simplex $\sigma = \langle a_0, a_1, \dots, a_p \rangle$ is in $W_1(D)$ if all its edges are in $W_1(D)$.

Computation of $W_\infty(D)$ is fussy, but $W_1(D)$ is much easier.

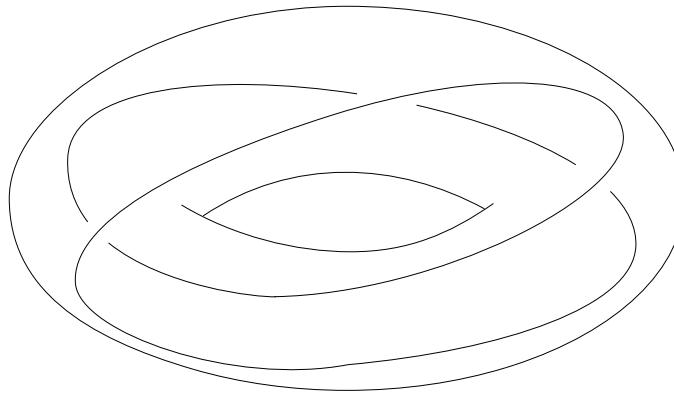
The Euclidean distance need not be used, instead one can use an intrinsic graph metric defined, see da Silva and Carlsson, Eurographics 2004.

Where to now?

Experiments and theory so far (mostly by the Stanford research team) have looked at feature detection and topological invariants from very large data sets, both artificial and 'real', but always with the assumption of a polyhedron underlying the data.

Plan: to see if it is possible to detect non-polyhedral behaviour in artificial data generated, initially, from the dyadic solenoid and the Menger cube. The idea is to vary not only the parameter R , but the other choices such as the landmark set, L , if working with the witness complexes, or the data set Z itself can also be varied taking different sample densities to detect features such as the double covering phenomena in this [dyadic solenoid](#).
What is that?

Here is a diagram showing how it is formed:



Inside the inner torus is another which goes four times around the 'central hole' of the original, and so on.

It is like a space swirling over a circle with fibre the (dyadic) Cantor set. How much about it could that be detected from a (noisy) sample? What theoretically, 'in the limit', could be detected?