VRDI TDA Breakout Session: Topological Data Analysis on Geospatial Data

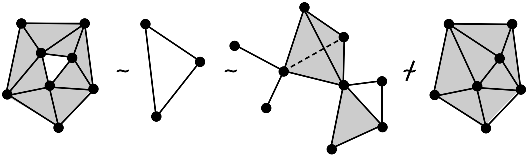
Moon Duchin, Tom Needham, Thomas Weighill

Voting Rights Data Institute
June 27, 2019

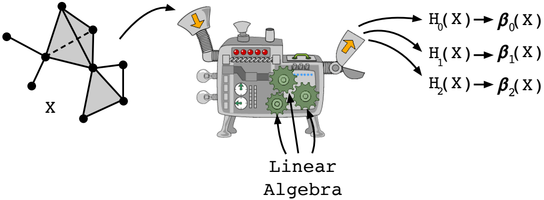
Quick Review of Algebraic Topology

Topology studies geometrical objects (called spaces) up to a loose notion of equivalence.

We will deal with special types of spaces called simplicial complexes.

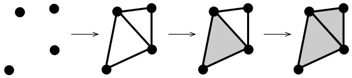


Algebraic topology distinguishes simplicial complexes by computing invariants; e.g., Betti numbers $\beta_k(X)$ count k-dimensional holes in a space X.

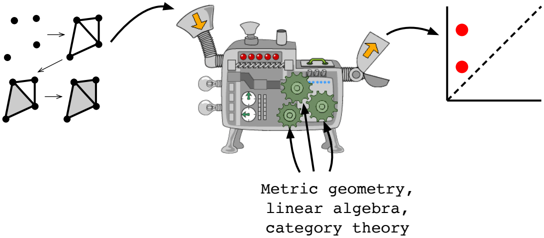


Quick Review of TDA

Persistent Homology computes topological invariants of families of simplicial complexes called filtered simplicial complexes.

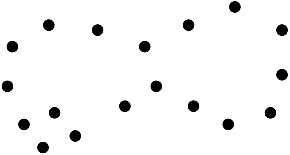


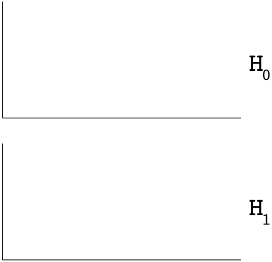
The invariants (Persistence Diagrams) describe topological features (holes) which appear and disappear in the family.



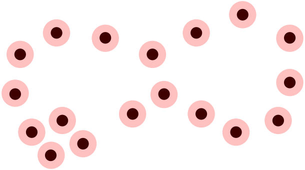
Given a dataset (e.g. a point cloud in $\mathbb{R}^d$), Topological Data Analysis explores its shape by turning it into a filtered simplicial complex.

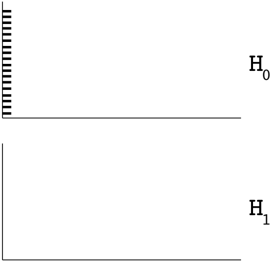
Persistent Homology - An Example



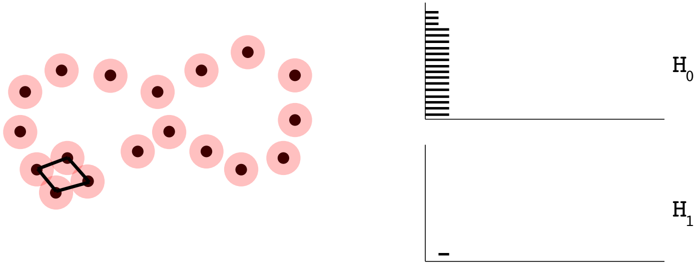


Persistent Homology - An Example

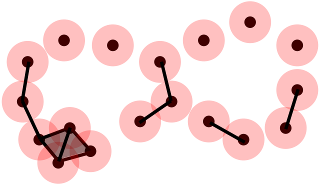


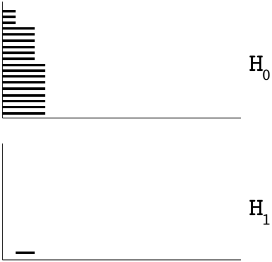


Persistent Homology - An Example

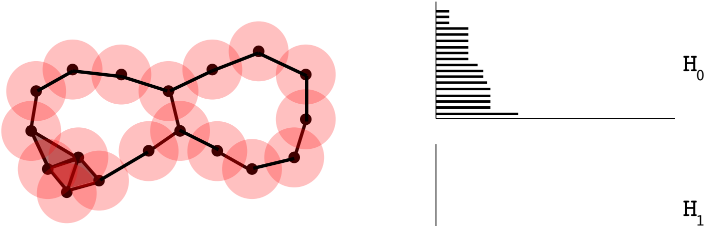


Persistent Homology - An Example

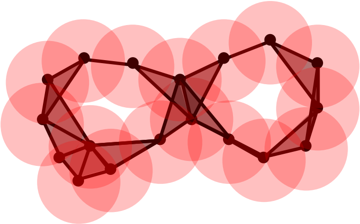


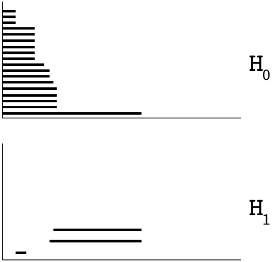


Persistent Homology - An Example

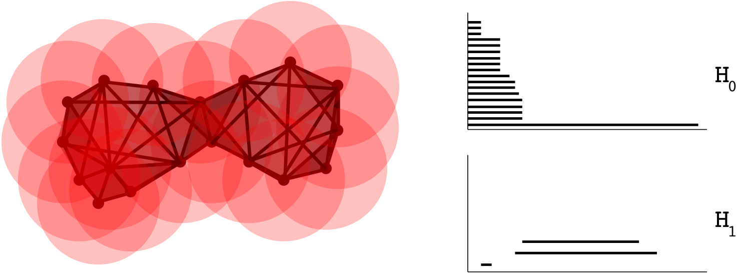


Persistent Homology - An Example



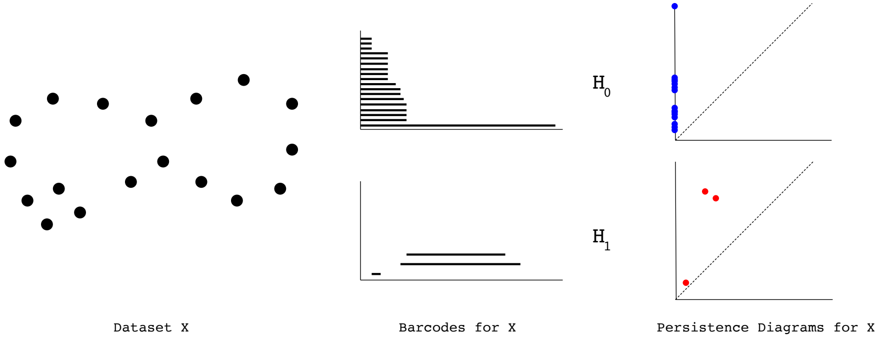


Persistent Homology - An Example



Terminology

We can record the "birth time" and "death time" of each topological feature to get a barcode or a persistence diagram.



Distance Between Diagrams

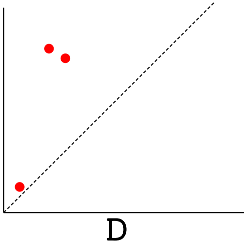
Important piece of the story: comparing persistence diagrams.

Each diagram D is a set[1] of points

$$D = \{(b_i, d_i)\}_{i=1}^{N}$$

with each $b_i < d_i$. Each point in D represents a topological feature of a dataset.

Let $\mathcal{D}$ denote the set of all diagrams.



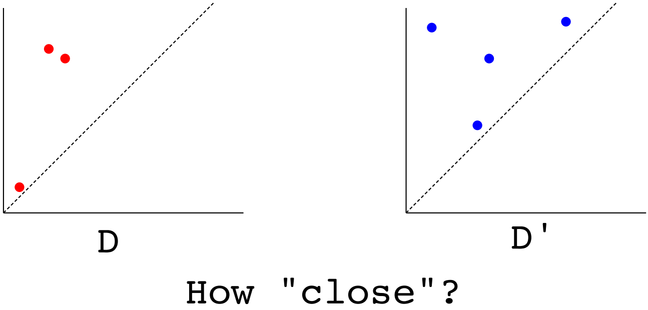
[1] Actually it's a multiset, but let's ignore that...

Metric on Diagrams

We wish to define a metric on $\mathcal{D}$.

This is a function $d : \mathcal{D} \times \mathcal{D} \to \mathbb{R}_{\geq 0}$ satisfying:

- (Positivity) $d(D, D') = 0 \Leftrightarrow D = D'$
- (Symmetry) $d(D, D') = d(D', D)$
- (Triangle Inequality) $d(D, D'') \leq d(D, D') + d(D', D'')$.



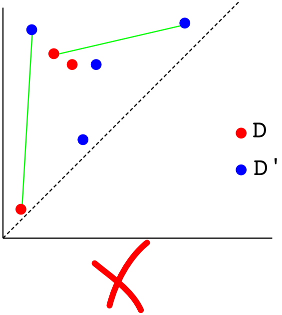
Bottleneck Distance

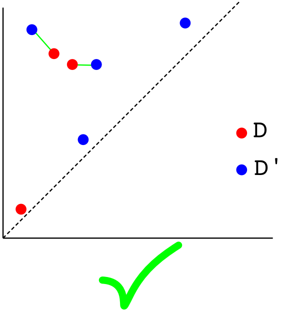
The bottleneck distance between persistence diagrams D and D' is

$$d_b(D, D') = \min_{\phi} \max \left\{ \max_{p \in A} c_m(p, \phi(p)), \max_{p \notin A} c_u(p), \max_{p' \notin A'} c_u(p') \right\}$$

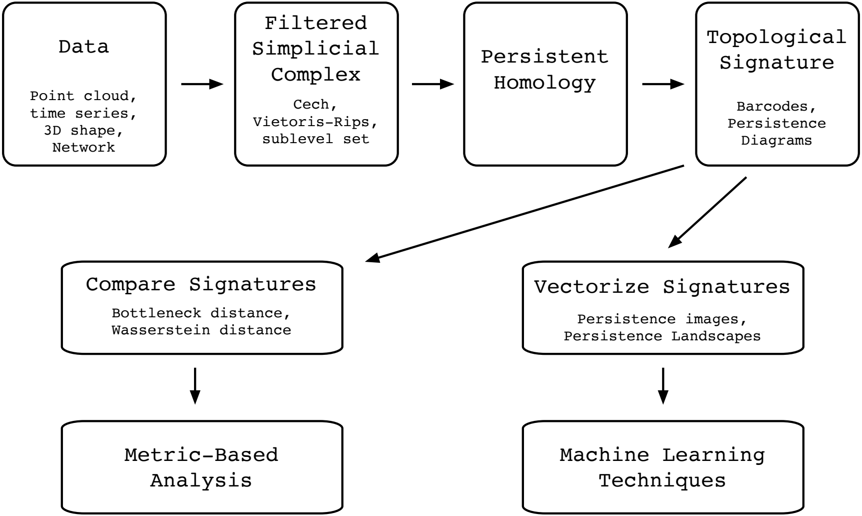
with min over partial bijections $\phi : A \to A'$, $A \subset D$, $A' \subset D'$ and

$$c_m(p, p') = \max\{|b' - b|, |d' - d|\}, \quad c_u(p) = \frac{d - b}{2}.$$

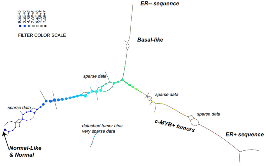




TDA Workflow



Applications



Nicolau et. al. (2011)

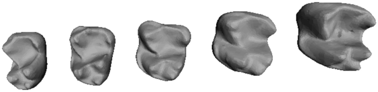
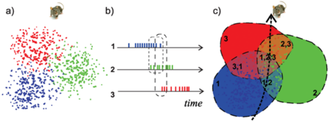
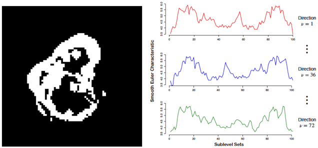


Figure 1: Images of the meshes of five teeth. A common problem in morphology is to measure distances between these five teeth.

Turner et. al. (2013)



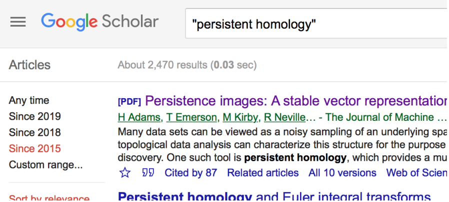
Dabagian et. al. (2012)



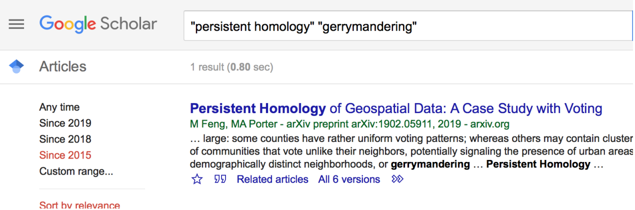
Crawford et. al. (2019)

Applications

There are many more recent applications...



...not too many applications to districting so far.



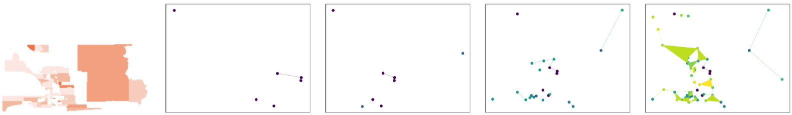
Feng-Porter Adjacency Networks

Create filtered simplicial complex for precincts in a county:

- Full simplicial complex is adjacency graph with all triangles filled.
- Filter by the function

$$\delta_{b,r}(p) := \frac{|V_b(p) - V_r(p)|}{V_b(p) + V_r(p)},$$

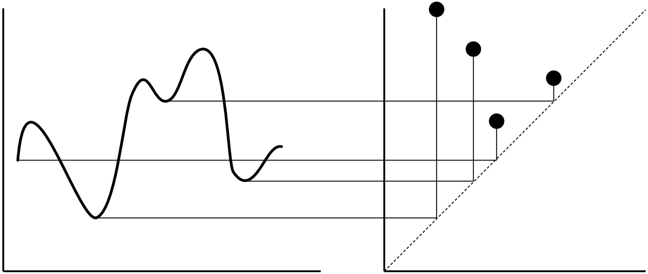
  with $V_b(p)$ the number of Clinton voters in precinct p, and $V_r(p)$ the number of Trump voters.
- Vertex for precinct p is included when filtration parameter is above $\delta_{b,r}(p)$.
- Edges/triangles are born at earliest possible time.



(a) (b) (c) (d) (e)

Level Set Filtrations

This is an example of a sublevel set filtration.



Suggestions for Future Directions

Ideas for how to push applications of TDA to districting problems:

- Consider other types of adjacency networks; e.g., the network of a districting plan.
- Filter adjacency networks for districting data by other functions, based on demographic or geometric data.
- Compare adjacency networks quantitatively using bottleneck distance. Can we determine that the shape of a particular districting plan makes it an outlier with respect to this metric?

These are explored in the accompanying Jupyter notebook.