

VRDI TDA Breakout Session: Topological Data Analysis on Geospatial Data

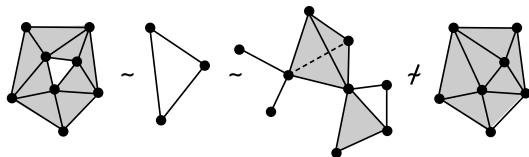
Moon Duchin, Tom Needham, Thomas Weighill

Voting Rights Data Institute
June 27, 2019

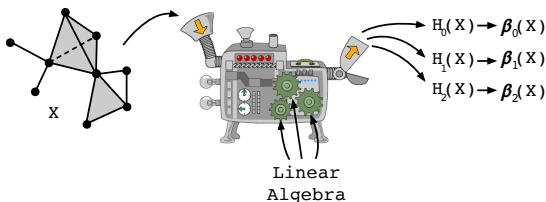
Quick Review of Algebraic Topology

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

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

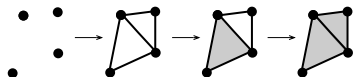


Algebraic topology distinguishes simplicial complexes by computing **invariants**; e.g., **Betti numbers** $\beta_k(X)$ count k -dimensional holes in a space X .

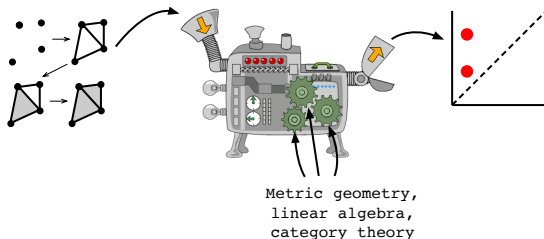


Quick Review of TDA

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

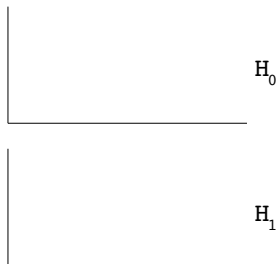
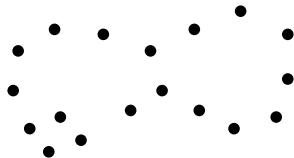


The invariants (**Persistence Diagrams**) describe topological features (holes) which appear and disappear in the family.

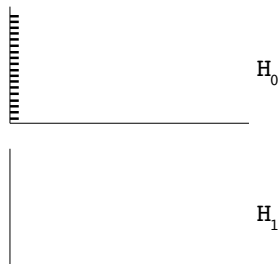
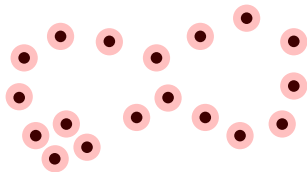


Given a dataset (e.g. a point cloud in \mathbb{R}^d), **Topological Data Analysis** explores its shape by turning it into a filtered simplicial complex.

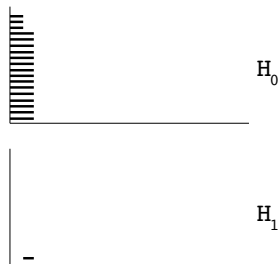
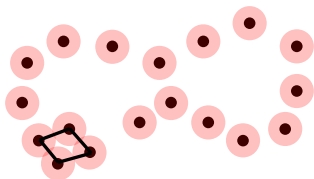
Persistent Homology - An Example



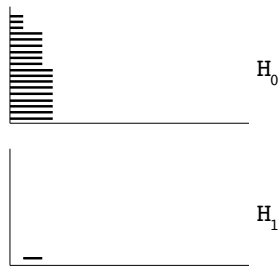
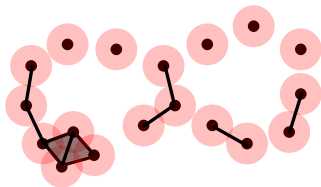
Persistent Homology - An Example



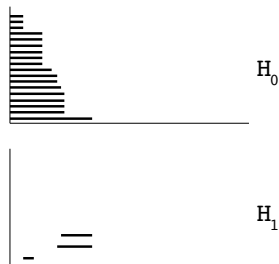
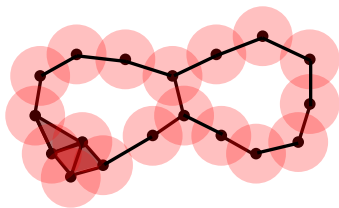
Persistent Homology - An Example



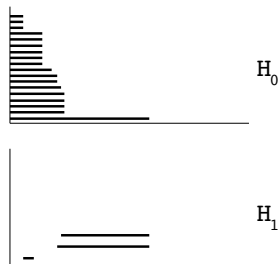
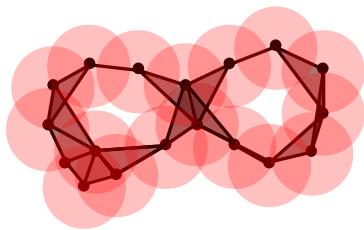
Persistent Homology - An Example



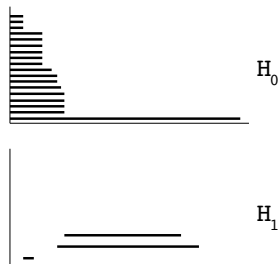
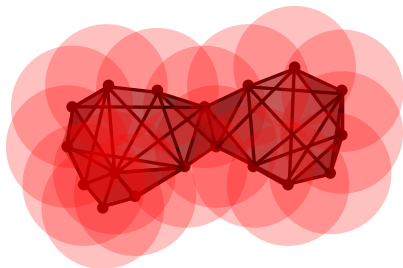
Persistent Homology - An Example



Persistent Homology - An Example

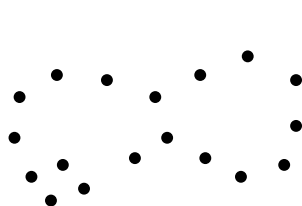


Persistent Homology - An Example

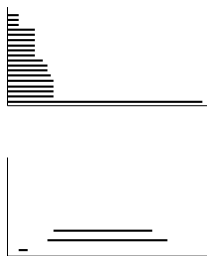


Terminology

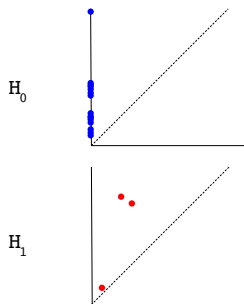
We can record the "birth time" and "death time" of each topological feature to get a **barcode** or a **persistence diagram**.



Dataset X



Barcodes for X



Persistence Diagrams for X

Distance Between Diagrams

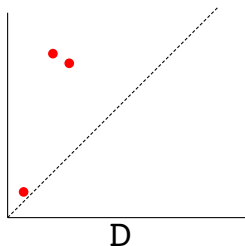
Important piece of the story: **comparing** persistence diagrams.

Each diagram D is a set¹ 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**.



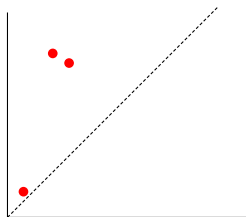
¹ Actually it's a **multiset**, but let's ignore that...

Metric on Diagrams

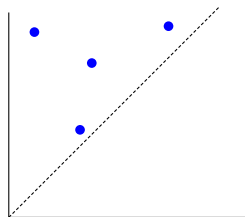
We wish to define a **metric** on \mathcal{D} .

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

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



D



D'

How "close"?

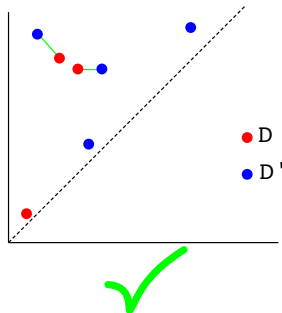
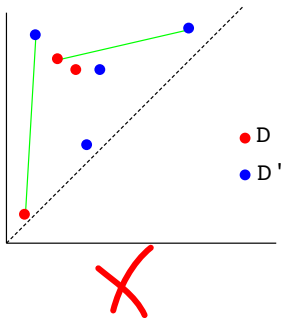
Bottleneck Distance

The **bottleneck distance** between persistence diagrams D and D' is

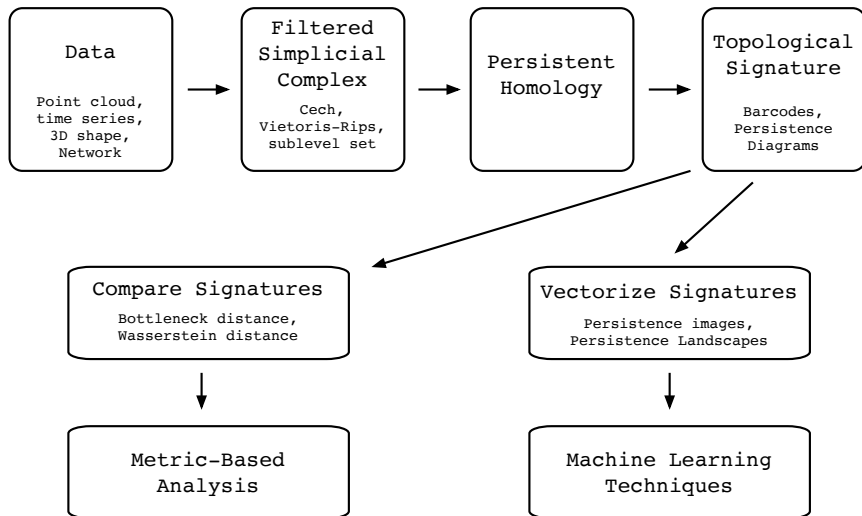
$$d_b(D, D') = \min_{\phi} \max \left\{ \max_{p \in A} c_m(p, \phi(p)), \max_{p \notin A} c_u(p), \max_{p' \notin A'} c_u(p') \right\}$$

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

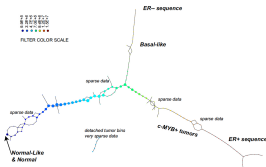
$$c_m(p, p') = \max\{|b' - b|, |d' - d|\}, \quad c_u(p) = \frac{d - b}{2}.$$



TDA Workflow



Applications



Nicolau et. al. (2011)

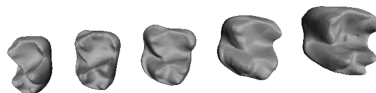
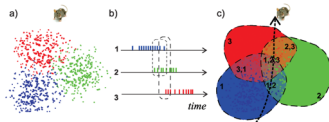
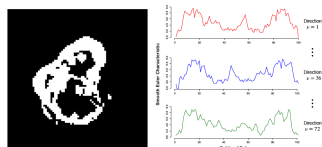


Figure 1: Images of the meshes of five teeth. A common problem in morphology is to measure distances between these five teeth.

Turner et. al. (2013)



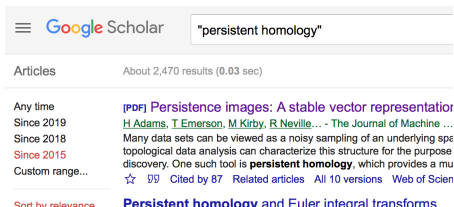
Dabagian et. al. (2012)



Crawford et. al. (2019)

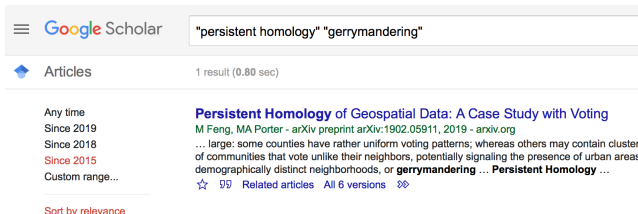
Applications

There are many more recent applications...



A screenshot of a Google Scholar search interface. The search bar contains the text "persistent homology". Below the search bar, it indicates "Articles" and "About 2,470 results (0.03 sec)". On the left side, there are filters for "Any time", "Since 2019", "Since 2018", "Since 2015", and "Custom range...". Below these filters is a red link "Sort by relevance". The main search results area shows a list of articles. The first article is titled "[PDF] Persistence images: A stable vector representation" by H Adams, T Emerson, M Kirby, R Neville... from The Journal of Machine ... The abstract mentions that many data sets can be viewed as a noisy sampling of an underlying space and that topological data analysis can characterize this structure for the purpose of discovery. It also mentions that one such tool is **persistent homology**, which provides a mu. Below the abstract are links for "☆", "99", "Cited by 87", "Related articles", "All 10 versions", and "Web of Scien". Below the first article is another article titled "Persistent homology and Euler integral transforms".

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



A screenshot of a Google Scholar search interface. The search bar contains the text "persistent homology" "gerrymandering". Below the search bar, it indicates "Articles" and "1 result (0.80 sec)". On the left side, there are filters for "Any time", "Since 2019", "Since 2018", "Since 2015", and "Custom range...". Below these filters is a red link "Sort by relevance". The main search results area shows a single article titled "Persistent Homology of Geospatial Data: A Case Study with Voting" by M Feng, MA Porter - arXiv preprint arXiv:1902.05911, 2019 - arxiv.org. The abstract mentions that some counties have rather uniform voting patterns; whereas others may contain clusters of communities that vote unlike their neighbors, potentially signaling the presence of urban areas or demographically distinct neighborhoods, or **gerrymandering**. Below the abstract are links for "☆", "99", "Related articles", "All 6 versions", and "99".

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)



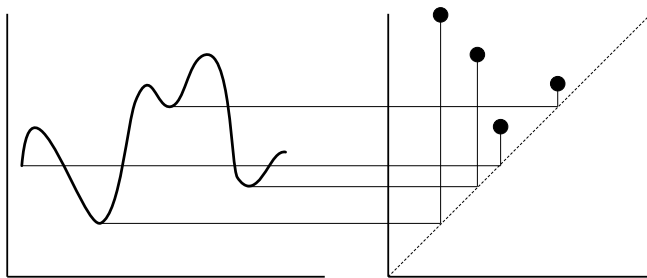
(d)



(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.