Persistent Homology for Mobile Phone Data Analysis

William Fedus[1], Mike Gartner[1], Alex Georges[1], David A. Meyer[1], David Rideout[1]
Department of Mathematics, University of California, San Diego[1]

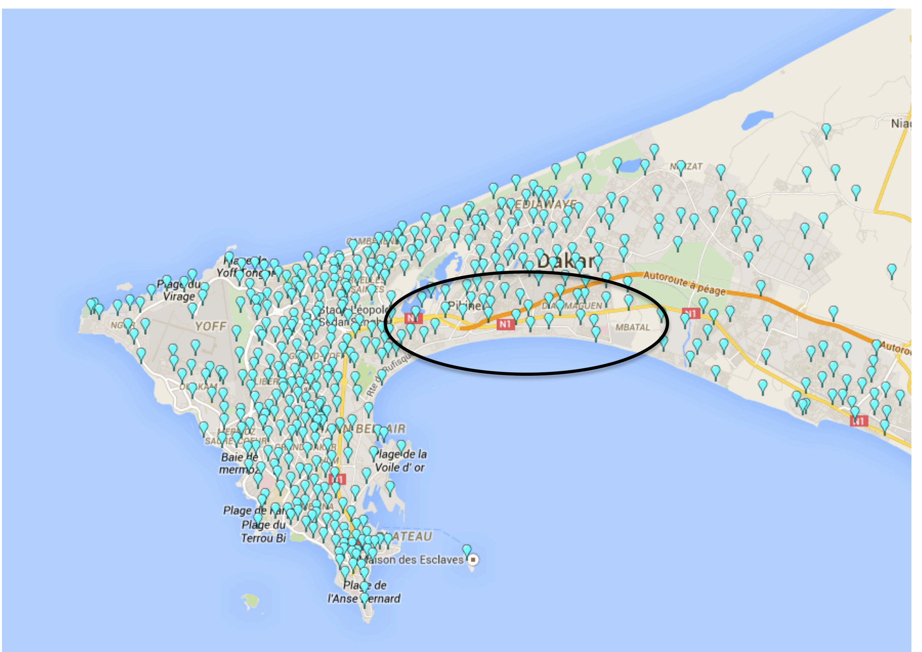
Abstract

Topological data analysis is an approach to analyzing the structure of high dimensional datasets. Persistent homology, specifically, generalizes hierarchical clustering methods to identify significant higher dimensional properties. In this project, we analyze mobile network data from Senegal to determine whether significant topological structure is present. We investigate whether the introduction of the Dakar motorway has any significant impact on the topological structure of the data. We see a possible indication of topological change, however, further exploration is continuing.

Dakar Motorway

Analyze the opening of the Pakine to Diamniadio section of the Dakar Motorway in 2013, with the goal of identifying signatures and impact of the new section of road via local changes in the persistent homology in the regions most directly affected.

Consider only the 500 towers in and surrounding Senegal to isolate the impact of the motorway.



500 towers located in the vicinity of the new section of the Dakar Motorway circled in black.

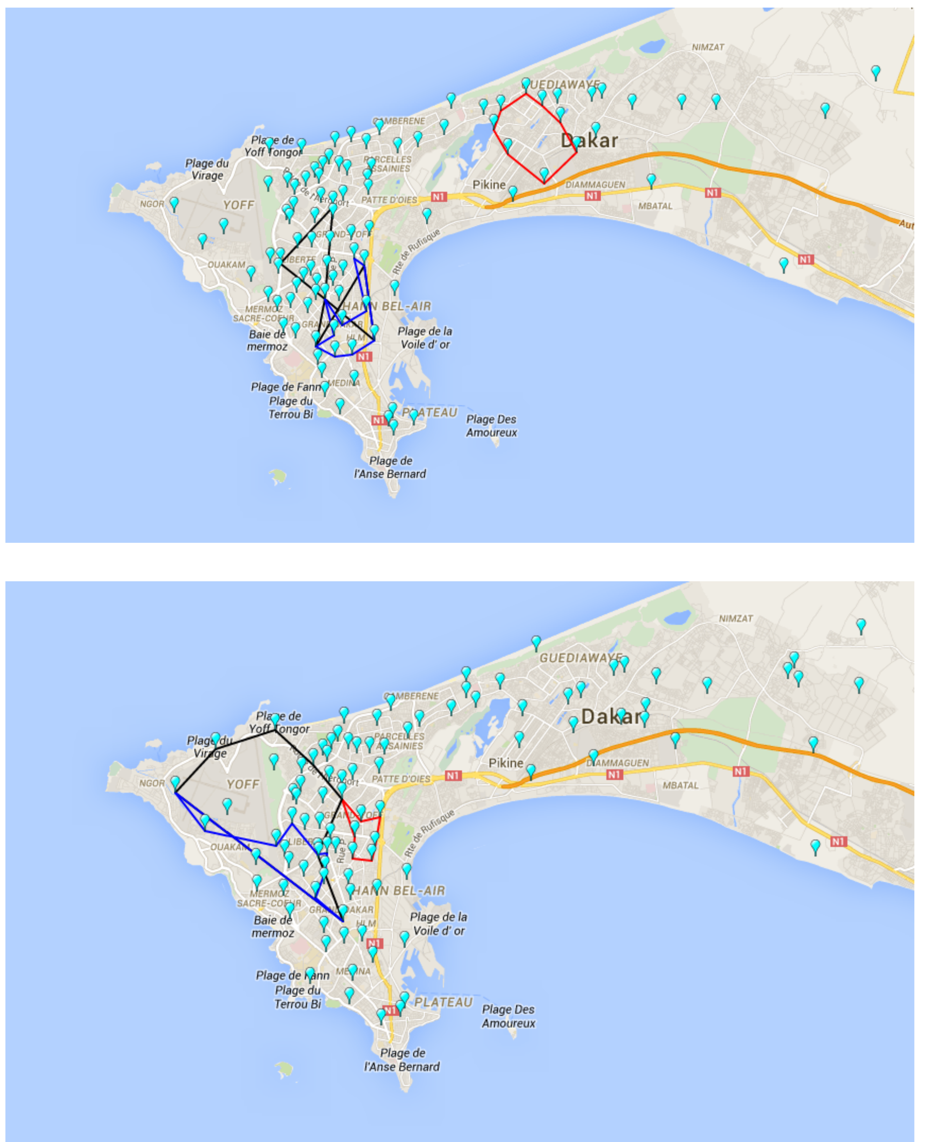
We considered three metrics to define a distance between towers, however, only one (defined below), the Inverse Call Duration metric indicated potential change to topology.

$$w(T)_{ij} = \begin{cases} C(T)_{ij}^{-1} & C(T)_{ij} \neq 0 \\ 1 & else \end{cases}$$

where $C(T)_{ij}^{-1}$ is call duration between towers i and j. Floyd-Warshall algorithm then ensures a valid metric.

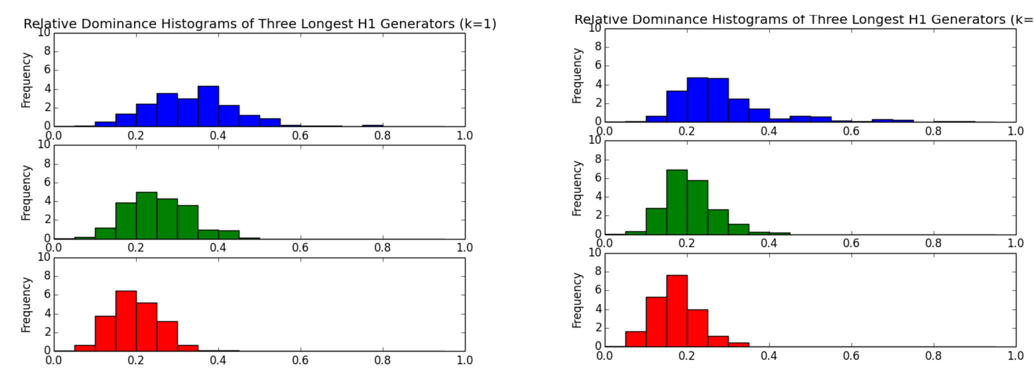
Dakar Motorway

We randomly select a subset of these 500 towers to perform our analysis on. For one iteration, we plot the resulting three longest H_1 generators before and after the opening of the Dakar Motorway.



An example of a single iteration's most persistent H_1 generators before (top) and after (bottom) the introduction of the new Dakar motorway section.

We conduct this random subsampling 250 times and then plot a histogram of the relative dominance of the three longest H_1 generators. The relative dominance describes how long the generator lasts relative to the filtration time where all data points become one connected component.



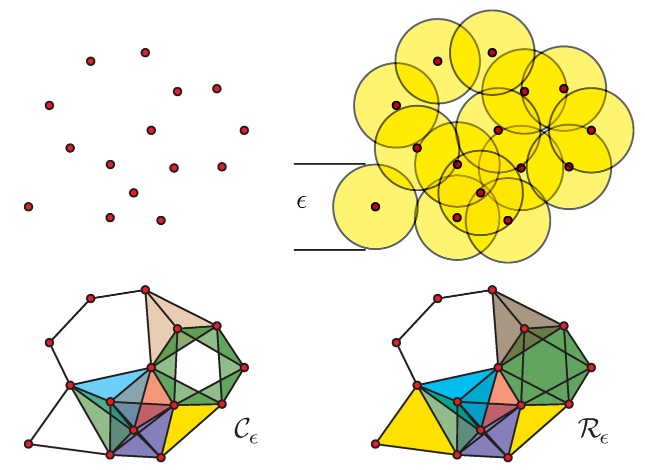
Histograms from before and after the motorway opening of the three longest H1 generators resulting from 250 runs over the inverse call duration metric.

Kolmogorov-Smirnov tests indicate that the distributions are likely different, implying a change in topology. We are conducting further work to ensure this change is unusual vs prior months.

Analysis Framework

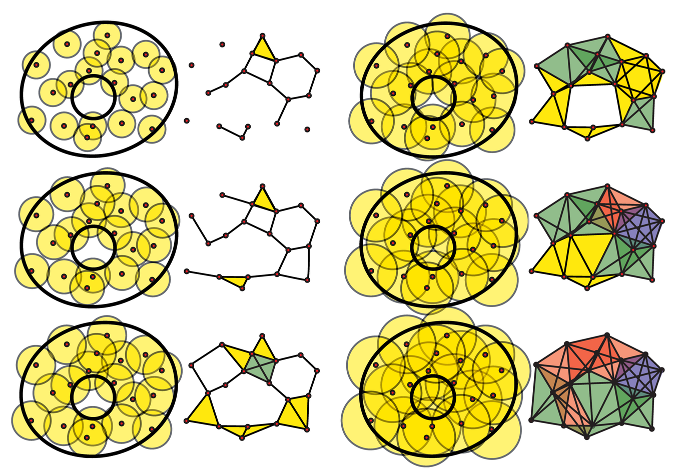
Select a metric, or distance measure, relevant to phenomenon to detect. Select subset of data points in order reduce noise and make structure apparent

II. Replace data points with a family of simplicial complexes defined by a distance parameter, ε



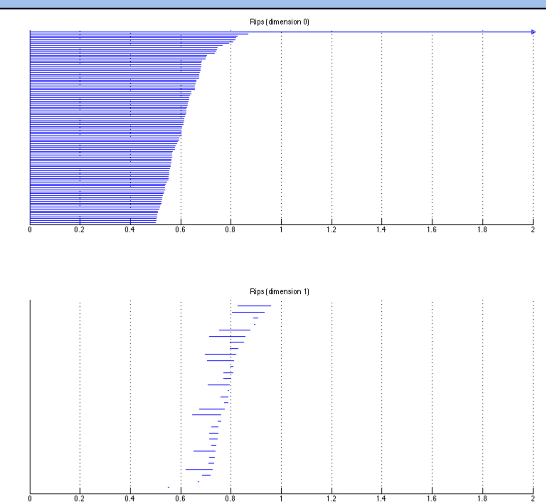
Schematic of how to construct a Cech complex (lower left) or Rips complex (lower right) from underlying high dimensional data. Note that the complex construction results in different homotopy type. [1]

III. Consider the underlying algebraic topology for particular epsilon (ε) parameters



Example of a sequence of Rips complexes on a data point cloud representing an annulus in a certain dimensional slice. [1]

IV. Determine the length of the longest generators. Persistence may indicate topological structure

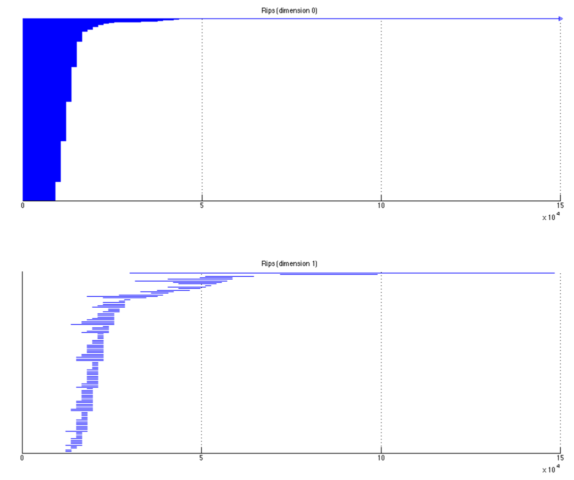


Example H_0 and H_1 barcodes for a particular point cloud. Notice that all H_1 generators are comparatively short and thus no significant topological structure would likely be evaluated from this data.

Illustrative Topology Example

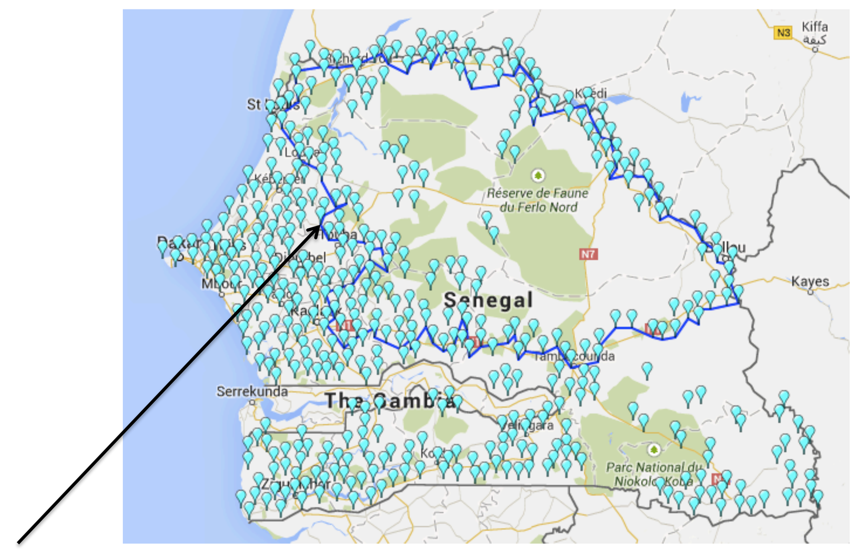
Visualization and verification of these topological techniques is easily done when considering the geographical structure of the most persistent generator of the first homology group, H_1 (i.e. the physical distance between towers).

Analysis finds one generator of the first homology group is very long, lasting for a 2.72 times the filtration window that the point cloud becomes a single connected component, indicating it's significant.



Rips barcodes for the simplicial complexes created via the geographical distance metric. Note that the most persistent H_1 generator spans a very long filtration window. Other generators are short-lived, likely corresponding to noise.

If we then map the generator on the towers of Senegal, we find that this lengthy generator represents the large void around the Wildlife Reserve. To better isolate this dynamic, we have chosen 550 towers via sequential max-min approach and then selected the 475 densest towers.



The approximate location of the 475 towers selected via sequential maxmin and then subsetting based on k=1 density are mapped with light blue pins and the most persistent H_1 homology group generator is mapped in dark blue.

Persistent Homology has efficiently and successfully found the the structure to this simple point cloud.

Contact Information: William Fedus, University of California, San Diego - Phone: 207.751.9036 – Email: liam.fedus@gmail.com
Acknowledgements: This work performed while the author held a UCSD Interdisciplinary Fellowship, Frontiers of Innovation Scholars Program

References: [1] R. Ghrist. 'Barcodes.: The Persistent Topology of Data", Bull. Amer. Math Soc., 45(1) 61-75 2008.