Persistent Homology for Mobile Phone Data Analysis

William Fedus, Mike Gartner, Alex Georges, David A. Meyer, David Rideout[1]

[1] *Department of Mathematics, University of California at San Diego, La Jolla, CA 92093*

(Dated: January 2015)

Topological data analysis is a new 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 two independent questions: whether the introduction of the Dakar motorway has any significant impact on the topological structure of the data, and how communities can be constructed using this method. We consider three independent metrics to compute the persistent homology. In two of these metrics, we see no topological change in the data given the introduction of the motorway; in the remaining metric, we see a possible indication of topological change. The behavior of clustering using the persistent homology calculation is sensitive to the choice of metric, and is similar in one case to the communities computed using modularity maximization.

INTRODUCTION

Most probability theory relies upon geometrical methods for analyzing data. For instance, a statistical distance must be defined so that two statistical objects can be quantified as being either close or far apart in some statistical measure. So, probability theory fundamentally encodes some type of length information. But, what if you want to concern yourself with a more fundamental property of the statistical objects: how are they structured? Topological data analysis, specifically the persistent homology method, accomplishes this. It determines the global structure of a set of data rather than its metric properties.

Topological data analysis is a new approach to analyzing the structure of high dimensional datasets. Persistent homology, specifically, generalizes hierarchical clustering methods to identify significant higher dimensional properties, which are out of reach of any other approach. It has been used to discover interesting and useful properties of data from systems ranging from natural images [1] through the visual cortex [2] to RNA folding [3].

We use persistent homology to study and analyze Senegalese anonymized mobile network data provided by Sonatel and Orange.

HOMOLOGY

In its broadest form, homology is a mathematical prescription that calculates algebraic properties of objects called chain complexes. When these chain complexes consist of objects called simplices, the homology that is calculated is a topological invariant of the space. It is thus a way to talk about isomorphisms of groups rather than homeomorphisms of spaces. This turns out to simplify the question of whether two spaces are fundamentally put together the same way or not. Formally, simplicial homology is defined as follows.

A *simplicial k-chain* (c_k) is a sum of *k-simplices* (σ_k):

$$c_k = \sum_i \alpha_i \sigma_k^i, \quad \alpha_i \in \mathbb{F} \tag{1}$$

where $\mathbb{F}$ is some field. Each k-simplex can be thought of as a k-dimensional polytope. Thus, a 2-simplex represents a triangle; a 3-simplex represents a tetrahedron, etc.

Thus, various k-chains define a free Abelian group which is denoted as C_k - i.e. $c_k \in C_k$. The *boundary operator* $\partial_k : C_k \to C_{k-1}$, is a linear homomorphism defined to act on $\sigma_k = [v_0, v_1, \ldots, v_k]$

$$\partial_k \sigma_k = \sum_i (-1)^i [v_0, v_1, \ldots, \hat{v}_i, \ldots, v_k] \in C_{k-1}$$

where "$\hat{v}_i$" means this element is removed from the simplex. This definition forces the condition used to compute homology: $\partial^2 \equiv 0$. This definition allows a flow of information in the various chain groups:

$$\ldots \to C_{k+1} \xrightarrow{\partial_{k+1}} C_k \xrightarrow{\partial_k} C_{k-1} \to \ldots$$

Various subgroups of this map can be defined. In particular, the *cycle group* $Z_k \equiv \ker \partial_k$ and the *boundary group* $B_k \equiv \text{im}\, \partial_{k+1}$. Because $\partial^2 \equiv 0$, this implies $B_k \subseteq Z_k \subseteq C_k$. This condition is necessary so the homology group can be defined as the quotient group:

$$H_k \equiv Z_k / B_k = \ker \partial_k / \text{im}\, \partial_{k+1}$$

Each homology group, H_k, contains information about the existence of k-dimensional holes in the space. For instance, the torus has $H_0 = \mathbb{Z}$, $H_1 = \mathbb{Z}^2$, $H_2 = \mathbb{Z}$ and all the remaining homology groups vanish. Refer to Hatcher's text [4] for a full treatment of the subject.

PERSISTENT HOMOLOGY

The previous discussion on homology requires the spaces to be triangulable, that is able to be thought of as a sum of k-simplices. For an arbitrary data set, there is no fundamental procedure to triangulate this space. Various ways do however exist, each with their own distinct set of rules, that can be used to construct simplices from data. For each of these procedures, we choose the coefficients in equation 1 to be in $\mathbb{Z}_2$.

We use the terms *point cloud* and *data set* interchangeably. Let $d(a, b)$ denoted the distance in a metric space between points a and b. Let Z denote the point cloud. We refer to ϵ as the *filtration value*, or simply the *filtration*. Note that for a large enough filtration, the complex will become one connected component. For a small enough filtration, there will be as many connected components as there are vertices. A vertex set consists of the base set of points used to construct higher dimensional simplices. Refer to [5], [6], [7] for overviews of persistent homology.

Vietoris-Rips Complex

Given a point cloud, the *Vietoris-Rips Complex* (R_ϵ) defines k-simplices as being determined by (k+1)-tuples of points whose balls of radius $\epsilon/2$ pairwise intersect [5]. The balls are drawn around each point in the point cloud, and the radius can be computed with an arbitrary metric. Specifically, to construct $R(Z, \epsilon)$:

1. The vertex set is Z
2. Edge [a,b] is in $R(Z, \epsilon)$ iff $d(a, b) \leq \epsilon$
3. Higher dimensional simplices are in $R(Z, \epsilon)$ if all of its edges are in $R(Z, \epsilon)$

One of the motivating reasons for this construction is that the union of the balls, which we interpret as being fundamentally representative of whatever topology the points came from, has a homotopy type that is closely related to the homotopy type of $R(Z, \epsilon)$ (see [8]).

Lazy Witness Complex

The construction of $R(Z, \epsilon)$ is computationally expensive because the entire point cloud is included as the vertex set. Choosing the vertex set as a subset of Z reduces the computation necessary to construct the simplex set over the range of all filtration values. $L \subset Z$ is called the *landmark set*, and points in it are chosen in one of two ways by selecting from Z [9]:

- *random* point selection: select points randomly from Z, the resulting set being L
- *maxmin* point selection: first, choose a random point in Z to serve as the first point in L. Each additional point in L is inductively chosen from Z by maximizing $d(z, l_i)\ \forall\ l_i \in L,\ \ z \in Z$

The size of L is variable depending on how large a vertex set is needed. Specifically, to construct $LW(Z, L, \epsilon, \nu)$:

1. The vertex set is L
2. Edge [a,b] is in $LW(Z, L, \epsilon, \nu)$ iff $\exists\ z \in Z$ such that $max\{d(a, z), d(b, z)\} \leq D_\nu(z) + \epsilon$
3. All higher dimensional simplices are in $LW(Z, L, \epsilon, \nu)$ if all of its edges are in $LW(Z, L, \epsilon, \nu)$

$D_\nu(z)$ is defined to be the distance from z to its νth closest neighbor. A feature of the lazy witness complex is that it behaves like a Delauney triangulation of the space when $\nu = 1$; for $\nu = 0$, the complex behaves similarly to $R(Z, \epsilon)$ [9].

Core Subsetting

Core subsetting is a procedure that helps uncover statistically significant topological structure in data. Since the data in this analysis does not come from a pure topological structure, but rather from data in a real world process, *a priori* we do not expect the entire data set to have an interesting topological structure. Rather, we expect subsets of the data to have the interesting structure. The procedure of core subsetting follows. First, start with an abritrary $n \times n$ metric space A:

$$A = \begin{pmatrix} a_{1,1} & a_{1,2} & \cdots & a_{1,n} \\ a_{2,1} & a_{2,2} & \cdots & a_{2,n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{n,1} & a_{n,2} & \cdots & a_{n,n} \end{pmatrix}$$

The second step is to produce the *density* vector:

$$\Delta_k = \begin{pmatrix} \delta_1^k \\ \vdots \\ \delta_n^k \end{pmatrix}$$

Where δ_j^k is defined to be the inverse of the distance from the jth point in the metric space (i.e. the jth row of A) to the kth closest neighbor. Hence, k is a parameter we scan over. A large k can be thought of as giving a more global estimate of the topology; similarly, a small k gives a more local estimate. The final step is to select a percentage (later on referred to as "P") of the densest points in Δ_k. The points from the metric space that give these densest points are then chosen to form a smaller metric space:

$$\widetilde{A} = \begin{pmatrix} a_{1,1} & a_{1,2} & \cdots & a_{1,m} \\ a_{2,1} & a_{2,2} & \cdots & a_{2,m} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m,1} & a_{m,2} & \cdots & a_{m,m} \end{pmatrix}, \; 10^2 * \frac{m}{n} \equiv P$$

Analytical Tools

There are various quantities we examine when determining the fundamental topological structure of the point cloud. These quantities are computed with javaPlex - a software built to construct persistent homology from an arbitrary point cloud [10]. The quantities we examine are:

- **Barcode Plots:** These depict the various generators of the different $LW(Z, L, \epsilon, \nu)$ or $R(Z, \epsilon)$. The x-axis represents the filtration value; the y-axis represents, in no physically significant ordering, the different homology generators. A barcode exists for each H_n.
- **Betti Numbers:** Are integers that count how many generators of a specific dimension exist at a specific filtration. For example,

  $$|H_1(R(Z, \epsilon), \ \epsilon = 3)| = 2$$

  means the first dimensional homology group for a Vietoris-Rips complex at a filtration of $\epsilon = 3$ has Betti Number equal to 2. In other words, it has 2 one-dimensional holes at this filtration.
- **Relative Dominance [9]:** A few definitions are necessary:

  1. R_0 = The filtration value at which a certain topological structure appears
  2. R_1 = The filtration value at which a certain topological structure disappears
  3. K_0 = The filtration value at which the complex becomes one connected component

  Relative Dominance is defined as $\delta_R = \frac{R_1 - R_0}{K_0}$. By a topological structure appearing, we mean any combination of Betti numbers in the various dimensions. So, a large δ_R corresponds to the topological structure being physicially significant; a small δ_R corresponds to the topological structure potentially being noise or a statistical fluctuation.

METRICS FOR ANALYSIS

We consider three metric spaces which arise from initially constructing a complete weighted graph, or equivalently a complete weighted adjacency matrix W_{ij}, whose vertices correspond to the 1666 cell phone towers (or a subset thereof). In two of the three cases, this adjacency graph does not immediately define a metric space, as the triangle inequality is not satisfied. However, through the use of an algorithm determining the shortest path between any two points in a complete weighted graph, a genuine metric space can be constructed. This procedure was carried out to obtain two of the metrics defined below, while the third can be defined directly from its complete weighted graph.

The Floyd-Warshall Algorithm

Our computation of persistent homology requires an underlying metric space from which simplices and chain complexes can be defined. Given a complete weighted adjacency graph on a set of vertices $X = \{1, 2, 3, ..., n\}$ with edge weights w_{ij}, one can construct a metric space $(X, \ d)$ by the following construction, known as the *Floyd-Warshall Algorithm.* It is constructed recursively:

$$\text{Path}_0(i, j) = w_{ij},$$
$$\text{Path}_k(i, j) = \min(\sum_{\text{edges}(lm) \in \gamma} w_{lm})$$

where the minimum is taken over all paths γ in the adjacency graph from vertex i to vertex j, using only vertices in the set $\{1, 2, ..., k\}$ as intermediate vertices. We then define a metric space $(X, \ d)$ by:

$$d(i, i) = 0 \ \forall \ i,$$
$$d(i, j) = \text{Path}_n(i, j)$$

One can verify that this satisfies the axioms of a metric space, since we begin with a complete weighted adjacency graph. Note: there are other possible constructions in going from a weighted adjacency graph to a metric space.

Data Aggregation

For each of the metrics we consider, the explicit construction of the metric from the data depends on a choice of aggregation period from the data provided. Let T denote an arbitrary set of time intervals during the year; for example, T could be the entire month of July, or the set of time intervals corresponding to hour 5 from every day of the year. We define below three metrics which depend explicitly on the choice of aggregation period T.

Inverse Call Duration Metric

The first metric we consider is a metric on the set of the 1666 cell phone towers, or a subset thereof, which is determined solely by the call volumes between towers. For a given choice of aggregation period T, and choice of a subset of the towers, let $C(T)_{ij}$ be the total duration of calls made between tower i and tower j during the aggregation period T. Note that our definition includes contributions from both i to j, and from j to i, so that $C(T)_{ij} = C(T)_{ji}$. These quantities were obtained from the call data provided in SET1V. Using these quantities, we then define a complete weighted adjacency matrix $w(T)_{ij}$ by:

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

Finally, we take the complete weighted adjacency matrix obtained above, and run the Floyd-Warshall Algorithm on it to obtain the *Inverse Call Duration Metric (ICD Metric)* for the aggregation period T.

Gravity Model Call Metric

Given an aggregation period T and a choice of subset of the towers, let $C(T)_{ij}$ be as above, and let $C(T)_i$ be the total duration of calls made to or from tower i during the aggregation period T. Let d_{ij} be the geographical distance between the towers i and j. This distance matrix was computed using the (altered, and thus slightly inaccurate) latitudes and longitudes provided, via the length of a spherical geodesic. Consider the following model, which is often described as a gravity model:

$$\log(C(T)_{ij}) = b + \log(\frac{C(T)_i C(T)_j}{d_{ij}^a})$$

Here a and b are parameters of the model. Since the call data will of course not fit the model exactly for any particular choice of a, b and T, we performed a linear least squares fit. In doing so, we fit only the subset of the data aggregated over the period T for which $C(T)_{ij}$, $C(T)_i$, $C(T)_j$, and d_{ij} were nonzero. The questions of how well the data fits the model, and whether a linear least squares fit is the best methodology to be using, were both considered, but will not be addressed here.

We then constructed the complete weighted adjacency graph $w_{ij} = d_{ij}^{a_0}$, where a_0 is the parameter value a arising from a linear least squares fit. Finally, we ran the Floyd-Warshall algorithm on this adjacency matrix to produce a metric, which we will call the *Gravity Model Call metric (GMC metric).*

Dot Product Call Metric

Given an aggregation period T one can construct for each tower, with label i, the vector:

$$v_i = \begin{pmatrix} C(T)_{i1} \\ \vdots \\ \vdots \\ C(T)_{i1666} \end{pmatrix}$$

Then define the *Dot Product Call Metric (DPC metric)* r to be the metric

$$r_{ij} = \arccos(\frac{v_i \bullet v_j}{\|v_i\| \|v_j\|})$$

It is easy to check that the angle between two vectors in a Euclidean space does indeed define a metric space.

Illustrative Example

Visualization and verification of these techniques is easily done by considering the geometrical, and in this case geographical, structure of the most persistent generators of the first homology group H_1. One can consider the most trivial example of what such persistent generators might mean by using the metric space arising from the physical geographical distance between any two cell towers. The persistent generators of the first homology group arising from a Vietoris-Rips complex correspond in general to long-lived cycles in the set of towers for which certain constituents are not directly connected over a long range of the filtration parameter. In the case of using the geographical metric arising from a Vietoris-Rips complex on the set of towers, the persistent generators correspond to the largest geographical voids of cell phone towers.

To demonstrate the technique, we first inspect Figure 1 which maps the approximate tower locations within Senegal.

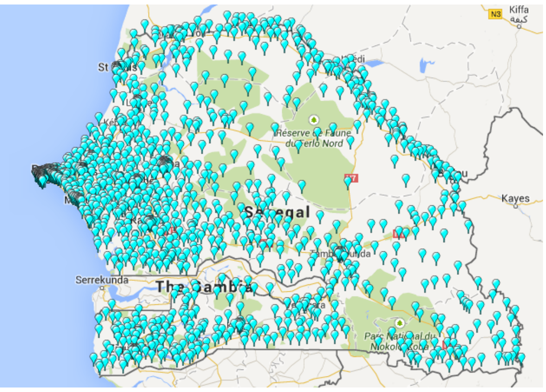

FIG. 1: The approximate location of all 1666 towers mapped onto Senegal where each tower is represented by a light-blue pin.

Visually, it is clear that a large void in the towers exists in the northeastern region of Senegal as a result of the

Ferlo Nord Wildlife Reserve. In order to more clearly see this void or cycle, we can consider subsets of these towers. In particular by choosing 550 towers via the sequential maxmin approach and then selecting the 475 densest points, we arrive at the following barcodes in Figure 2.

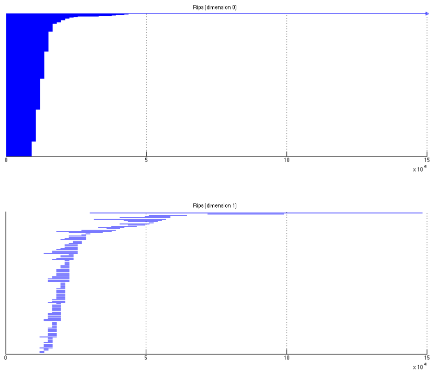

FIG. 2: Rips barcodes for the simplicial complexes created via the geographical distance metric. Note that the most persistent H_1 generator spans a very long filtration window of $[3 \times 10^4, 14.8 \times 10^4]$ before it closes. Other generators are short-lived, corresponding to noise in the determination of the topology.

Now note that one generator of the first homology group H_1 persists over the entire range of filtration values. If we then map the towers and overlay the generator of this persistent H_1 homology group, we find the following result in Figure 3. Additionally, the relative dominance of this generator is 2.72, indicating that this generator existed for a considerable duration after the entire point cloud was a single connected component. The substantial relative dominance indicates the significance of this feature and this helps guide intuition for the future search of true underlying structure.

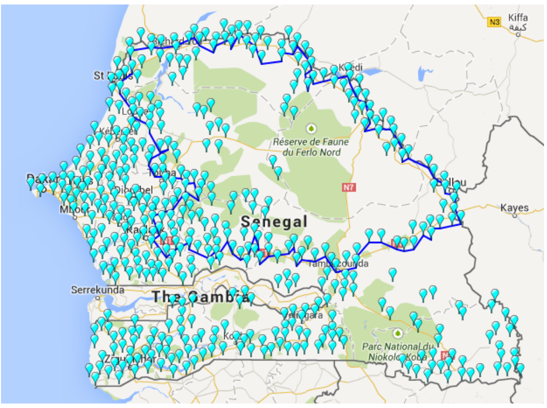

FIG. 3: The approximate location of the 475 towers selected via sequential maxmin and then subsetting based on $k = 1$ density are mapped with light blue pins and the most persistent H_1 homology group generator is mapped in dark blue.

This example, though trivial, demonstrates a successful application of topological data analysis as we have identified the large void in towers resulting from Senegal's extensive Wildlife Reserve. As we move to more useful metric spaces, the ability for simple visualization is lost, but the principles of the analysis remain the same.

RESULTS

We aim to use the formalism of persistent homology applied to the three particular metrics on the set of towers described above to address the thematic development issues outlined by the D4D Development team. Our hope is that this novel method of telecom data analysis will allow certain objectives to be accomplished in a relatively easy way which will provide new insight into the structure of the data.

Dakar Motorway: Preliminary Results

The first issue we intend to address is the facilitation of useful development of transportation and infrastructure systems in Senegal. The analytical framework we work with seems well suited to addressing the problem of identifying regions which would benefit the most from the development of new local transport methods, as well as identifying the effects of installment of new local transport methods which have already occurred. In particular, we considered 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.

For the three metrics described above, we considered the persistent homology of the Vietoris-Rips complex arising from the aggregation periods T_1 and T_2 corresponding to the entire months of July and August in 2013, respectively. Additionally, we limited the points in our metric space to be among the cell phone towers with labels between 1 and 500, which correspond to the western region of Senegal potentially impacted by the Dakar Motorway opening. These choices of aggregation periods and subset were made with the goal in mind of identifying a substantial change in the generators of the first homology between these months, indicating a signature of the construction of the section of the Dakar motorway between Pikine and Diamniadio, which opened on August 1st of 2013.

However, as we saw in the geographical metric example, using the entire point cloud can often obscure the underlying topology within a higher dimensional space. In order to determine the underlying topology and reduce noise, we may instead choose a representative subset of the points. For instance, towers for which there is no activity on either period T_1 or T_2 will be excluded since the core subset procedure will choose the densest towers; towers with no activity will necessarily be low-density within our choice of metrics.

To proceed, we again consider the DPC, ICD and GMC metrics. However, now we choose a smaller subset of the original 500 tower set near Dakar. To do so, we first choose 250 of the 500 towers via the sequential maxmin method and then further reduce this by choosing the top 100 densest towers (i.e. $P = 40$) as defined by a core subset with varying k. The goal is to reduce the noise sufficiently that any hidden structure and any changes in this structure from period T_1 to T_2 will become evident.

After choosing these subsets, the T_1 and T_2 barcodes of the 0th and 1st dimensional homology for the DPC, ICD and GMC metrics are displayed in Figure 4, 5 and 6.

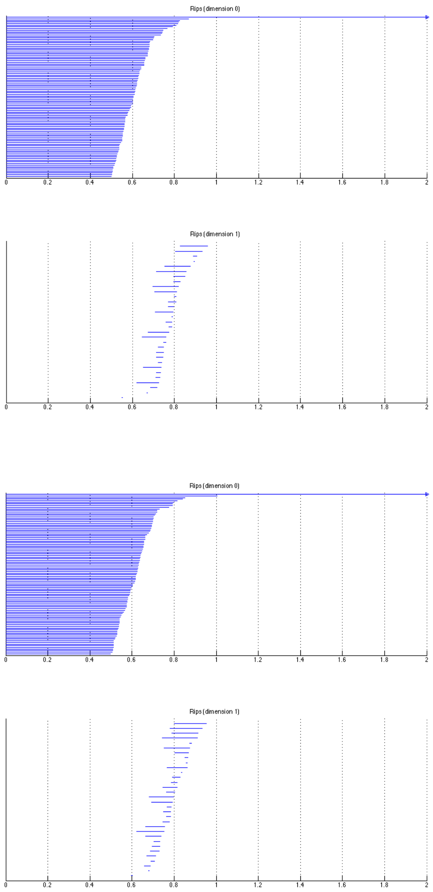

FIG. 4: Rips Barcodes for the simplicial complexes created from reduced point cloud and the DPC metric for the aggregation periods T_1 and T_2 respectively

For each of the barcode plots shown, we considered the relative dominance of the three most persistent generators of the one dimensional homology H_1. These relative dominance measures are listed in Table no 1-dimensional homological structure is evident nor is a significant change between period T_1 and T_2 apparent in the H_1 generators. To summarize these barcode plots, we can consider the relative dominance of the top three H_1 generators. We summarize the most persistent generators for the three metrics in Table I. For all of the metrics we detect no long generators, nor do we detect a significant change in the length of the three most persistent H_1 generators for this choice of parameters.

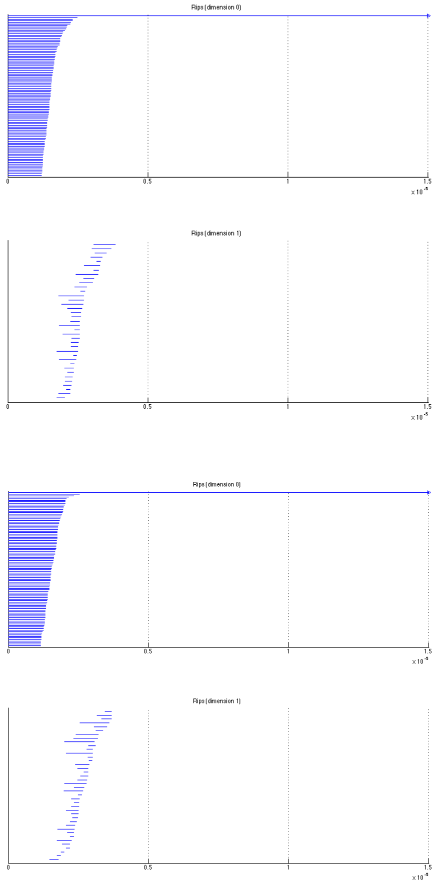

FIG. 5: Rips Barcodes for the simplicial complexes created from reduced point cloud and the ICD metric for the aggregation periods T_1 and T_2 respectively

Note that the relative dominance varies widely over the three metrics. While there is no precise lower bound on the relative dominance necessary to deem that a generator is indicative of a genuinely persistent topological structure, it is apparent that the ICD metric and the Dot Product Call metric give rise to simplicial complexes whose most persistent generators are more highly dominant than those of the GMC metric.

In addition to the fact that the GMC metric has a low relative dominance for its most persistent generators as noted in Table I, each of the three metrics yields bar-

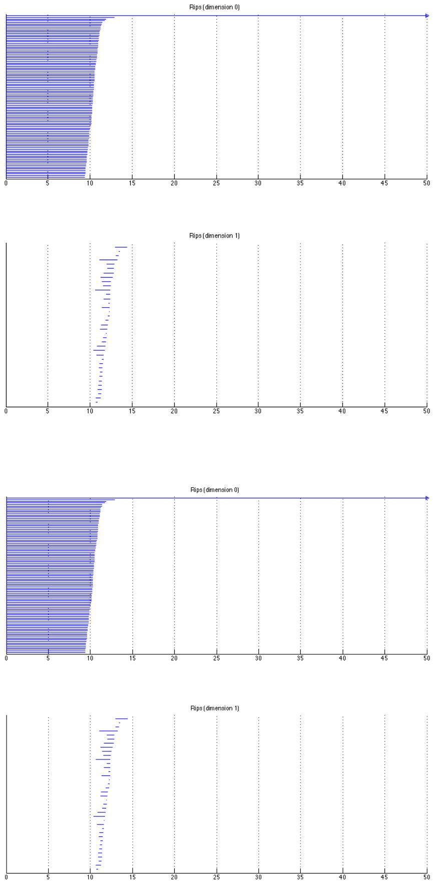

FIG. 6: Rips Barcodes for the simplicial complexes created from reduced point cloud and the GMC metric for the aggregation periods T_1 and T_2 respectively

ICD T_1	ICD T_2	DPC T_1	DPC T_2	GMC T_1	GMC T_2
0.323	0.420	0.162	0.166	0.166	0.166
0.280	0.377	0.148	0.153	0.132	0.132
0.273	0.338	0.147	0.152	0.108	0.107

TABLE I: The relative dominance of the three most persistent generators of H_1 for the ICD, DPC, and GMC metrics over the aggregation periods T_1 and T_2. Note all relative dominance values for this particular subset and parameter choice are less 0.5, indicating no significant topological structure.

codes which look similar in structure for the aggregation periods of July and August (see the following section for a quantitative analysis of this conclusion). In the case of the GMC metric, this similarity can be attributed to the fact that the least squares parameter a_0, which is the only dynamic aspect of that model across different data aggregation periods, varied only slightly between the two months. Thus in the context of that metric, the actual underlying topology, and not just its barcode, underwent very little change between July and August. This suggests that for the particular choice of subset and aggregation periods made here, the GMC metric is not a sensitive probe for changes in structure of the data. While our initial hope was that there might be some qualitatively obvious change in the structure of the 1st dimensional homology between July and August appearing readily in at least one of the barcodes, their appearance indicates that detection of any significant change will necessitate sensitive analysis.

In order to further analyze whether the topological structure of the data encoded in these metrics underwent a significant change between July and August of 2013, we consider geographical realizations of the most persistent 1-dimensional generators in each barcode. This has the double benefit of providing more information than is present in the barcode plots, which will aid in the potential recognition of a topology change, as well as giving insight into what such a change indicates in terms of the specific affects on local call activity near Dakar. In Figures 7, 8, 9 representative cycles of the the three generators with the highest relative dominance, or "the most persistent" generators, are overlayed for the two aggregation periods on a map with the locations of the actual cell tower locations.

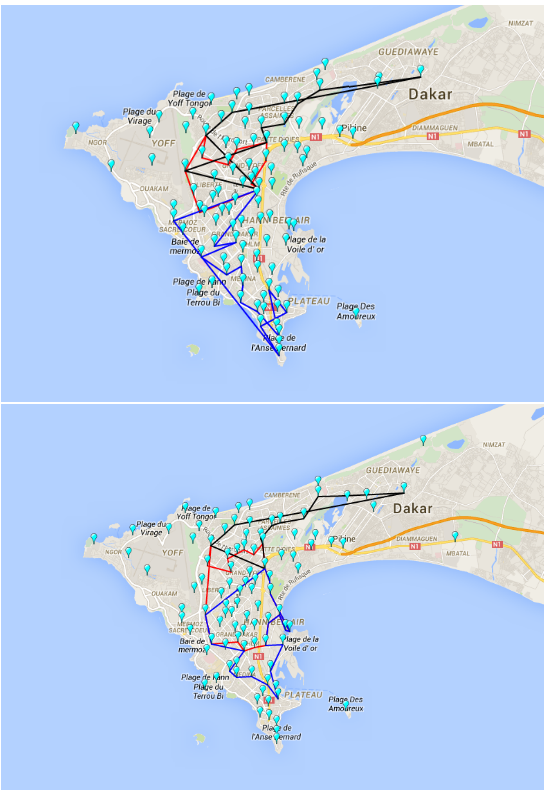

FIG. 7: Representatives of the three most persistent generators of H_1 for the DPC metric for aggregation periods T_1 and T_2, overlayed on a map with the approximate tower locations.

For the GMC and DPC metrics, these representative

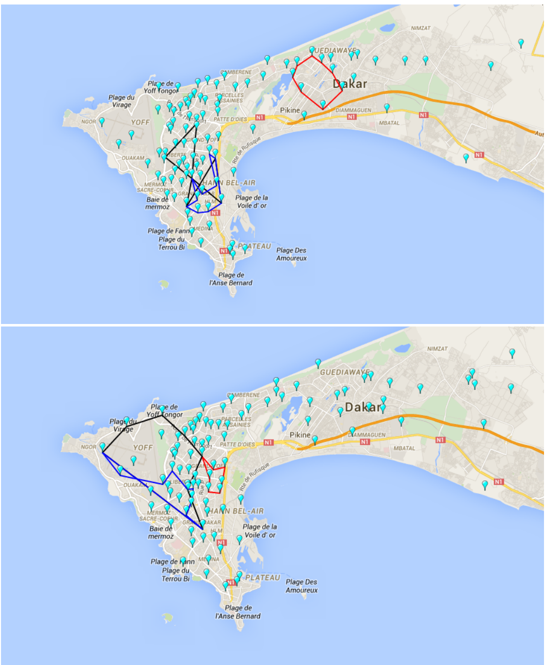

FIG. 8: Representatives of the three most persistent generators of H_1 for the ICD metric for aggregation periods T_1 and T_2, overlayed on a map with the actual tower locations.

cycles remain virtually unchanged between July and August. This further validates the claim that the GMC and DPC metric and the underlying simplicial complex to which it gives rise remain unchanged between July and August, and provides more evidence than the barcode alone. For the ICD metric in Figure 8, we do see the largest change in the most persistent generators between periods T_1 and T_2, however, due to the relative shortness of both generators, we still do not attribute this to a genuine topological change. The ICD metric appears to be a higher variance distance metric than the other two we have chosen and we explore this further in the next section. However, given the shortness of all the generators, we do not conclude that a definitive change has been detected. Thus even considering the geographical representatives of the generators of the first homology group, it appears that none of our metrics were able to capture a significant topology change due to the introduction of a section of the Dakar Motorway. We hypothesize from these results that there is no detectable change in the topological structure of the call duration data with respect to the models here considered. This null result suggests that either the introduction of the Motorway had no real affect on call traffic, and social activity coupled to this traffic, or if there was an effect that its detection would require a different measure of call activity and distance between towers.

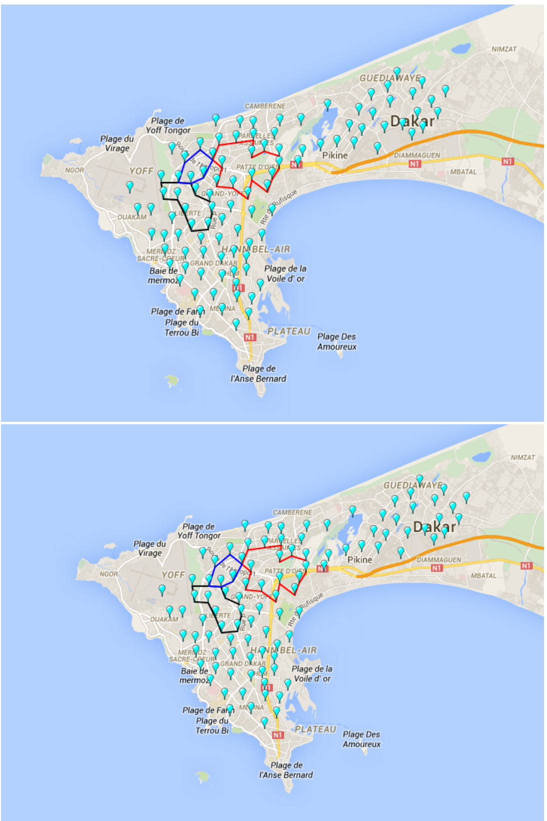

FIG. 9: Representatives of the three most persistent generators of H_1 for the GMC metric for aggregation periods T_1 and T_2, overlayed on a map with the actual tower locations.

Dakar Motorway: Statistical Analysis of Preliminary Results

To make this hypothesis about a lack of detectable 1-dimensional homology more rigorous, we consider repeated runs where a different subset of towers is randomly chosen each time. For each repeated run, we record the top three relative dominances and consider the distributions between periods T_1 and T_2. By repeatedly choosing different subsets of towers and by additionally scanning over k core subset parameters, we reduce the likelihood that we are missing underlying topological structure.

In order to do this, we again select 250 of the original 500 tower set, but instead of using sequential maxmin, we now randomly select the towers for each run. This ensures that each run is sufficiently different in order to get an independent sampling of the tower set. Then with these 250 towers randomly selected, we choose the 100 densest towers (i.e. $P = 40$) as defined by core subset choices for $k \in \{1,\ 26,\ 51\}$.

As one specific example of this procedure, the results of 250 randomized runs with the ICD metric for $k = 1$ core subset is shown in Figure 10.

To compare the frequency distributions of the relative

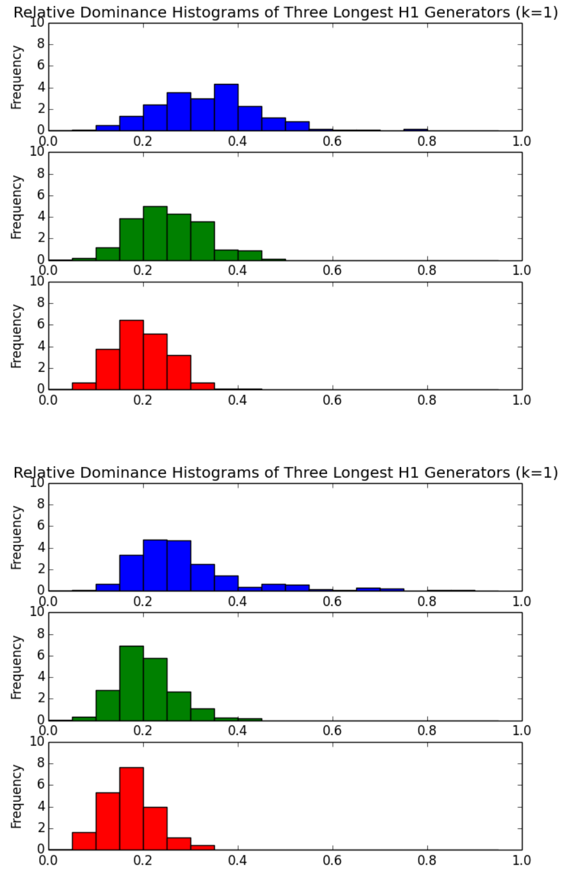

FIG. 10: Histograms of the three longest H_1 generators resulting from 250 runs of the ICD metric for $k = 1$ over the aggregation periods T_1 and T_2 respectively

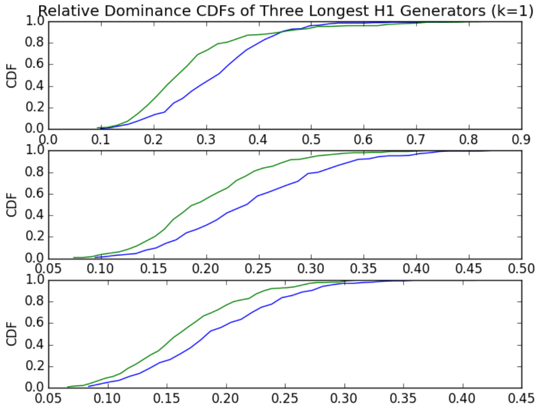

FIG. 11: CDFs of the three longest H_1 generators resulting from 250 runs of the ICD metric for $k = 1$ over the aggregation periods T_1 and T_2. T_1 is colored blue; T_2 is colored green.

dominance for a given metric, we use the Kolmogorov-Smirnov test ("KS" test). The KS test takes as an input a cumulative distribution function ("CDF") and outputs a number which is used to check whether two samples come from the same distribution.

We perform a KS test between each CDF over the aggregation of July and August for a given metric and generator length for H_1. In Tables II, III, IV we give the probability that each pair of distributions were randomly sampled from the same distribution. So a sufficiently low probability is necessary to reject the null hypothesis.

Generators $k = 1$ **Metric**	1st	2nd	3rd
ICD	0	0	0
DPC	52	33	19
GMC	45	67	3.0

TABLE II: $k = 1$ P-values for KS test comparing the the aggregation periods T_1 and T_2. We compare the 1st, 2nd, and 3rd most persistent generators for each of the three metrics. P-values given in percentages.

Generators $k = 26$ **Metric**	1st	2nd	3rd
ICD	0	8.0	33
DPC	0.28	10	45
GMC	38	27	5.0

TABLE III: $k = 26$ P-values for KS test comparing the the aggregation periods T_1 and T_2. We compare the 1st, 2nd, and 3rd most persistent generators for each of the three metrics. P-values given in percentages.

Generators $k = 51$ **Metric**	1st	2nd	3rd
ICD	0	0	1.3
DPC	74	67	0.72
GMC	38	96	52

TABLE IV: $k = 51$ P-values for KS test comparing the the aggregation periods T_1 and T_2. We compare the 1st, 2nd, and 3rd most persistent generators for each of the three metrics. P-values given in percentages.

Communities via 0th Homology and Comparison with Modularity

Using Vietoris-Rips to construct simplices, which is equivalent to single-linkage clustering, we may identify connected components at any desired filtration value. With our three metrics, we compare the clustering results to that of another approach based on modularity. Modularity Q is a quality index for decompositions of a network into communities, which measures the fraction of edges that fall within the given communities minus the expected fraction if those edges were distributed at random, and has a value between -1 and 1. For a particular weighted graph $G = (V, E)$, with edge weights A_{ij}, the modularity of a decomposition into communities $V = \cup_i C_i$, $C_i \cap C_j = \emptyset \ \forall i \neq j$, may be defined as

$$Q(C) = \sum_{i,j} \left(\frac{A_{ij}}{A} - \frac{k_i k_j}{A^2} \right) \delta(c_i, c_j) \qquad (2)$$

where $A = \sum_{i,j} A_{ij}$, $k_i = \sum_j A_{ij}$, and c_i is the community label of vertex i. The optimal community decomposition for a given weighted graph is defined to be that which maximizes the modularity Q.

In practice, for any reasonably sized network, it is not computationally practical to compute an exactly optimal decomposition into communities. We therefore make use of the hierarchical algorithm, and accompanying software, introduced by Blondel et. al. [11]. We find that it is very effective in finding near optimal community decompositions for mobile phone networks of this sort [12].

Employing this algorithm, we compute communities for two weighted mobile phone networks, one weighted by the total volume of voice communication between each pair of towers in July, and the other weighted by the total voice volume in August. The results are depicted in Figures 12 and 13. We note in particular that, as above, the Dakar motorway does not have any appreciable effect on the modularity communities.

To compare these with communities identified as generators of H_0, we show in Figures 14 and 15 a clustering resulting from the ICD and DPC metrics. We note that the zeroth homology generators arising from the ICD metric are not terribly dissimilar from the communities detected by the modularity optimizing algorithm. This suggests that the ICD metric may be capturing similar information as the edge weights which go into the modularity optimization, and that therefore the higher dimensional homology generators arising from that metric may be exploring richer structures present in the graph weighted by call volume, than is available from the modularity analysis alone.

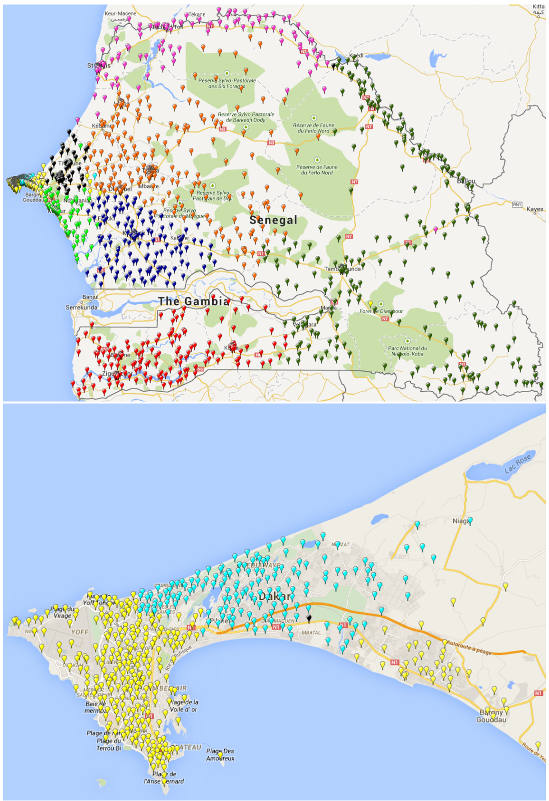

FIG. 12: Map of the ten communities within Senegal (above) and within Dakar (below) detected via modularity, using total voice communication volume in July. Communities with only a single tower are ignored.

CONCLUSIONS

We chose both to construct a reduced metric space from core subsetting, and to analyze the towers most likely to be affected by the introduction of the Dakar motorway. By comparing the barcode plots and the relative dominances for the Rips complex for the months of July and August, we determine no statistically significant changes in the homology from month to month, given these choices. We also detect no significant topological change in the Lazy Witness complex construction given these same choices.

In both the Vietoris-Rips and Lazy Witness constructions, we scan over various parameters: the choice in metric space (either ICD, GMC, or DPC); the percentage of points that are used in the core subset ("P"); the density neighbourhood parameter ("k"); and the choice in the determination of the Landmark Set L (either random or maxim).

Thus, the null result suggests that either the introduction of the Motorway had no real affect on call traffic, and social activity coupled to this traffic, or if there was an effect then its detection would require a different measure of call activity and distance between towers.

In a future analysis, large scale averaging relative dominance over multiple trials will improve any claim to the statistical relevance of a topological feature. In the current analysis, we do no large scale averaging, and instead average over just a few trials.

With regard to modularity-based community detection, we find that, for the ICD metric, the H_0 genera-

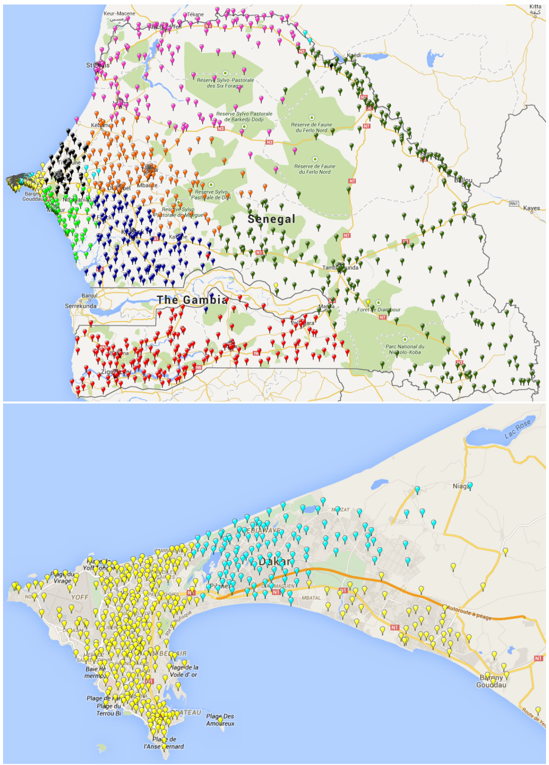

FIG. 13: Map of the ten communities within Senegal (above) and within Dakar (below) detected via modularity, using total voice communication volume in August. Communities with only a single tower are ignored.

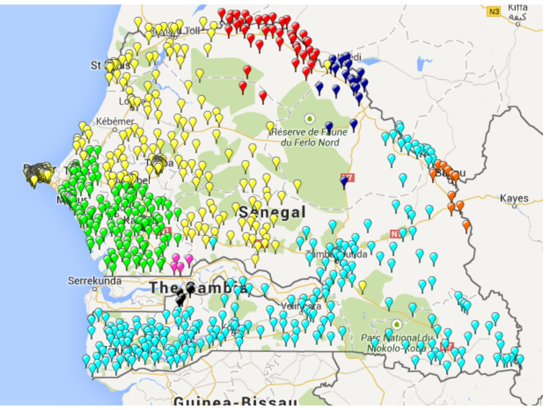

FIG. 14: Map of the nine largest detected communities within Senegal and within Dakar via the most persistent H_0 groups found with the ICD metric.

tors provide a qualitatively similar decomposition. Given that constructing simplices via Vietoris-Rips and building the clusters is mathematically equivalent to single-linkage clustering, a rudimentary approach, it is promising that the results can mimic that of the modularity communities.

Acknowledgements We would like to thank the developers of javaPlex, and specifically to Henry Adams for all the insight he provided into the world of persistent homology.

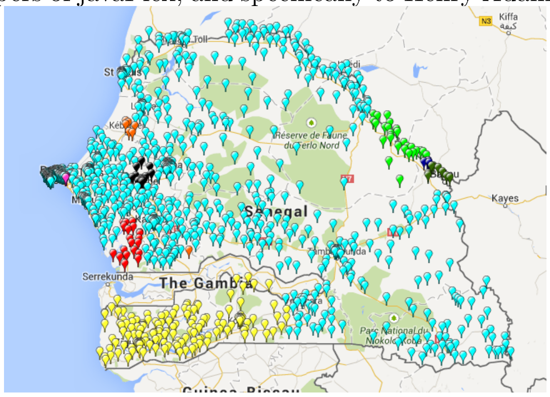

FIG. 15: Map of the nine largest detected communities within Senegal and within Dakar via the most persistent H_0 groups found with the DPC metric.

REFERENCES

[1] G. Carlsson, A. Zomorodian, et al, "On the Local Behavior of Spaces of Natural Images," Int J Comput Vis (2008) 76: 1-12, DOI 10.1007/s11263-007-0056-x, 2007.

[2] G. Carlsson, G. Sapiro, et al, "Topological Structure of Population Activity in Primary Visual Cortex," 2007.

[3] G. Carlsson, et al, "Structural Insight into RNA Hairpin Folding Intermediates," J. AM. CHEM. SOC. 2008, 130, 9676-9678 2008.

[4] A. Hatcher, "Algebraic Topology: A First Course," 2002.

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

[6] A. Zomorodian and G. Carlsson, "Computing Persistent Homology," 2004.

[7] A. de Silva and G. Carlsson, "Topological Approximation by Small Simplicial Complexes," 2003.

[8] G. Carlsson, "Topology and Data," Bull. Amer. Math. Soc. (N.S.) 46 (2009), no. 2, 255308.

[9] V. Silva, G. Carlsson, "Topological estimation using witness complexes," Eurographics Symposium on Point-Based Graphics 2004.

[10] A. Tausz, H. Adams, M. Vejdemo-Johansson "JavaPlex: A research software package for persistent (co)homology," 2011, Software available at http://javaplex.github.io/

[11] Vincent D Blondel, Jean-Loup Guillaume, Renaud Lambiotte, and Etienne Lefebvre, "Fast unfolding of communities in large networks", *Journal of Statistical Mechanics* P10008 (2008).

[12] O. Bucicovschi, R. Douglass, D. Meyer, M. Ram, D. Rideout, and D. Song, "Analyzing Social Divisions Using Cell Phone Data", NetMob 2013, Orange D4D Challenge Best Scientific Prize.