TED Talk Tag Topology

Brandon Garcia

February 19, 2018

Abstract

In this paper, the structure of the big ideas of two decades of TED talks is explored via topological data analysis. TED talks are catalogued with conceptual tags and these tags provide a way to analyze the closeness and connectivity of TED talks. This dataset is explored through three different modeling approaches of the persistent homology of these TED talks and TED talk tags.

Contents

1 Introduction

1.1 What are TED Talks?

TED Talks began in 1984 with the aim to share ideas from experts in Technology, Engineering, and Design, but have vastly expanded to include presentations on some of the most interesting ideas that humanity is currently puzzling about.

Each TED talk range in length from a few minutes up to a half hour, and present to a general audience an innovation, a discovery, or an idea that, as the TED.com website declares, are perhaps "ideas worth spreading." The filming of TED talks began in 1994, and the video of these talks is published online for broader public consumption. Through the archived recordings of these talks, we do preserve a partial history of some of the most interesting ideas of the late 20th century and early 21st century.

1.2 Why Analyze TED Talks?

Because each TED talk connects new ideas with a web of existing ideas, the collection of all TED talks present a window into the structure of the ideas important to discovery and innovation. Excitingly, this analysis is made readily accessible by the cataloguing of all TED Talks available online with conceptual tags. Over 416 different concepts are currently used in the tagging of TED talks, and include tags like: "Africa", "Justice", "Internet", "Vaccines", "language", "decision-making", "feminism", "jazz", and "pain".

Moreover, a diverse set of analyses are made possible via the dataset made available on kaggle.com (https://www.kaggle.com/rounakbanik/ted-data-analysis/data) of all TED talks published online from 1994 up to 2017 [Banik, 2017]. This dataset records, among various things: the date of filming, the audience ratings, the number of views, the occupation of the speaker, the number of comments, and the collection of tags for each of the 2550 talks in the collection. While this dataset provides many interesting avenues of research, this study focused on exploring the tags attributed to each talk.

1.3 Why use TDA?

This study designs a series of examinations of the topological shape of the connections between ideas as presented by the collection of tags attributed to all TED Talks from 1994 to the present.

Because each of the 2550 talks map to a subset of the 416 different tags, this dataset represents a very high dimensional dataset. While high dimensionality is a major obstacle to traditional data analysis techniques, the tools of topological data analysis presents an opportunity to peak into structure of very high dimensional datasets. Thus, with the tools of topological data analysis, this study sets out to discover the shape of humanity's big ideas.

1.4 What is TDA?

Persistent Homology is a type of topological data analysis that utilizes techniques which extract homological features on point cloud data. Thus, by analyzing point cloud data with respect to homology, it is possible to speak of the existence of topological invariant properties typically attributed to hyper-surfaces, like: components, cycles, and other higher dimension "separations" or "voids".

These homological features arise from the point cloud data, by constructing higher dimensional constructions from the points. Such higher dimensional constructions commonly include simplicial complexes or cellular complexes. By imposing filtrations on the construction of these complexes, the separation or dissimilarity between points can be accounted for by observing the corresponding changes in the homology throughout the filtration. The features in the homology that persistence over significant intervals of filtration values represent dominant topological features that can be said to be present in the point cloud data itself.

1.5 Overview of Paper

Sections 2 introduces some of the terminology used throughout the paper. Sections 3 through 5 provide a a summary of three experiments performed using the kaggle TED talk dataset. Each section introduces and provides results for one of the three data models utilized. The paper then concludes with a few insights from the aforementioned topological data analysis experiments.

First, section 3 introduces a vector space model for the data, where each TED talk is represented by a point in the 416 dimensional space. Next, section 4 explores the results from utilizing a complex of cells (instead of simplices) as the underlying topological space for the persistent homology. Finally, section 5 returns to the vector space model, except provides the results for the homology of TED talk tags as the dual to the space of TED talks.

2 Topological Data Analysis (TDA) Reference

The following provides a brief introduction to the TDA jargon used throughout this paper.

2.1 Simplicial Complexes

Definition: A simplicial complex $\mathcal{K}$ is a set of simplices statisfying the following conditions:

- Any face of a simplex from $\mathcal{K}$ is also in $\mathcal{K}$.
- The intersection of any two simplices $\sigma_1, \sigma_2 \in \mathcal{K}$ is a face of both σ_1 and σ_2.

Definition: An abstract simplicial complex if, for every set X in Δ, and every non-empty subset $Y \subset X, Y$ also belongs to Δ.

2.2 Cellular Complexes

Definition: A cellular complex, or a CW complex, is a Hausdorff space X together with a partition of X into open cells (of perhaps varying dimension) that satisfies two additional properties:

- For each n-dimensional open cell C in the partition of X, there exists a continuous map f from the n-dimensional closed ball to X such that the restriction of f to the interior of the closed ball is a homeomorphism onto the cell C, and the image of the boundary of the closed ball is contained in the union of a finite number of elements of the partition, each having cell dimension less than n.
- A subset of X is closed if and only if it meets the closure of each cell in a closed set.

2.3 Filtration

Definition: A filtration $\mathcal{F}$ is an indexed set S_i of subobjects of a given algebraic structure S, with the index i running over some index set I that is a totally ordered set, subject to the condition that

if $i \leq j$ in I, then $Si \subseteq Sj$.

Specifically, a filtration on a simplicial complex X is a collection of subcomplexes $\{X(t)|t \in \mathbb{R}\}$ of X such that $X(t) \subset X(t')$ whenever $t \leq t'$. The filtration value of a simplex $\sigma \in X$ is the smallest t such that $\sigma \in X(t)$.

2.4 Homology

Definition: The construction begins with an object such as a topological space X, on which one first defines a chain complex $C(X)$ encoding information about X. A chain complex is a sequence of abelian groups or modules $C_0, C_1, C_2, \cdots$ connected by homomorphisms $\partial_n : C_n \rightarrow C_{n-1}$, which are called boundary operators. That is, where 0 denotes the trivial group and $C_i \equiv 0$ for $i < 0$. It is also required that the composition of any two consecutive boundary operators be trivial. That is, for all n, $\partial_n \circ \partial_{n+1} = 0_{n+1,n-1}$ i.e., the constant map sending every element of C_{n+1} to the group identity in C_{n-1}. That the boundary of a boundary is trivial implies $\text{im}(\partial_{n+1}) \subseteq \ker(\partial_n)$, where $\text{im}(\partial_{n+1})$ denotes the image of the boundary operator and $\ker(\partial_n)$ its kernel. Elements of $B_n(X) = \text{im}(\partial_{n+1})$ are called boundaries and elements of $Z_n(X) = \ker(\partial_n)$ are called cycles.

Since each chain group C_n is abelian all its subgroups are normal. Then because $\ker(\partial_n)$ is a subgroup of C_n, $\ker(\partial_n)$ is abelian, and since $\text{im}(\partial_{n+1}) \leq \ker(\partial_n)$ therefore $\text{im}(\partial_{n+1})$ is a normal subgroup of $\ker(\partial_n)$. Then one can create the quotient group $H_n(X) := \ker(\partial_n)/\text{im}(\partial_{n+1}) = Z_n(X)/B_n(X)$, called the nth homology group of X [Kun, 2013]. The elements of $H_n(X)$ are called homology classes and gives the number of k-dimensional holes.

2.5 Persistent Homology and Barcodes

Definition 6: Persistent homology describes how the homology of a topological space $X(t)$ changes with filtration value t. The interval over which an element in the k-dimensional homology persists, with endpoints $[t_{start}, t_{end})$, corresponds roughly to a k-dimensional hole that appears at filtration value t_{start}, remains open until it closes at value t_{end}. Elements in the homology that persist for a long filtration range signify features in the space $X(t)$ not likely due to noise or or other artifacts.

The interval over which an element in the homology persists can be visualized as a bar. A collection of these bars visualized as a barcode graph provides a representation of homology groups of a filtered simplicial or cellular complex. Thus, these barcode graphs represent a visually intuitive way to conceptualize the structure of a point clouds persistent homology

3 Experiment 1

3.1 Data Model

In the first experiment conducted, the kaggle dataset was analyzed as a collection of TED talks and each TED talk was modeled as a "set-of-tags". For a given TED talk, its "set-of-tags" was all conceptual tags used by TED.com to catalogue the content of a talk at the time of the creation of the kaggle dataset. So, for example, the set of tags for Ken Robinson's "Do schools kill creativity?", one of the most popular TED talks of all time, consists of the tags: 'children', 'creativity', 'culture', 'dance', 'education', 'parenting', and 'teaching'.

A natural way to talk about similarity of sets is the Jaccard coefficient, or $(\frac{A \cap B}{A \cup B})$ the ratio of the intersection of a pair of sets to the union of the sets [Wikipedia]. The dual to the Jaccard coefficient, $1 - (\frac{A \cap B}{A \cup B})$, provides a measure of the dissimilarity of sets. While the dual to the Jaccard coefficient is not a metric, this measure can provide enough structure to closely approximate a metric space, assuming the the triangle inequality does not fail to often in practice. According to Adams and Tausz, it is still possible to define a Vietoris–Rips complex on top of a "metric space" that fails to meet the triangle inequality [Adams and Tausz, 2017].

3.2 Persistence Homology Computation Methods

Using the "metric space" of the set of tags for each TED talk, it is possible to define various simplicial complexes over this underlying topological space. In this experiment, both Vietoris-Rips and lazy witness complex constructions were utilized.

A Vietoris-Rips complex builds the simiplicial complex according to the following definition [Adams and Tausz, 2017].

Definition: Let $d(\cdot, \cdot)$ denote the distance between two points in metric space Z. The Vietoris-

Rips complex $VR(Z,t)$ is defined as follows:

- the vertex set is Z.

- for vertices a and b, edge $[ab]$ is included in $VR(Z,t)$ if $d(a,b) \leq t$.

- a higher dimensional simplex is included in $VR(Z,t)$ if all of its edges are.

Unfortunately, because of the rapidity with which the Vietoris-Rips complex grows, the Vietoris-Rips complex is very computationally intensive for high dimensions and large filtration ranges. For such application contexts, the lazy witness constructions can be an effective approach to reduce computational load. A lazy witness complex builds the simiplicial complex according to the following definition [Adams and Tausz, 2017].

Definition: Suppose we are given a point cloud Z, landmark subset L, and parameter $\nu \in N$. If $\nu = 0$, let $m(z) = 0$ for all $z \in Z$. If $\nu > 0$, let $m(z)$ be the distance from z to the ν-th closest landmark point. The lazy witness complex $LW_\nu(Z,L,t)$ is defined as follows:

- the vertex set is L.

- for vertices a and b, edge $[ab]$ is in $LW_\nu(Z,L,t)$ if there exists a witness $z \in Z$ such that $maxd(a,z), d(b,z) \leq t + m(z)$.

- a higher dimensional simplex is in $LW_\nu(Z,L,t)$ if all of its edges are.

Thus, with the underlying space of TED talks with dissimilarity measure of the Jaccard Coefficient dual, the homology of chains of this space over $\mathbb{Z}/2\mathbb{Z}$ was computed up to dimension 7. The results of these computations follow.

3.3 Results

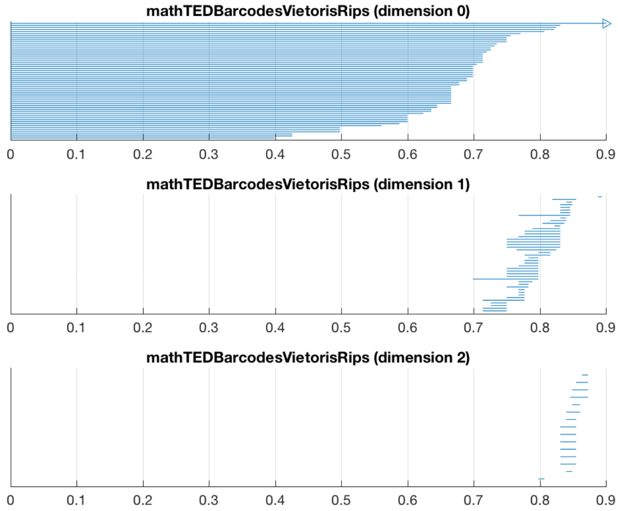

Figure 1: Observing Math-Related Talks using a Vietoris Rips Stream

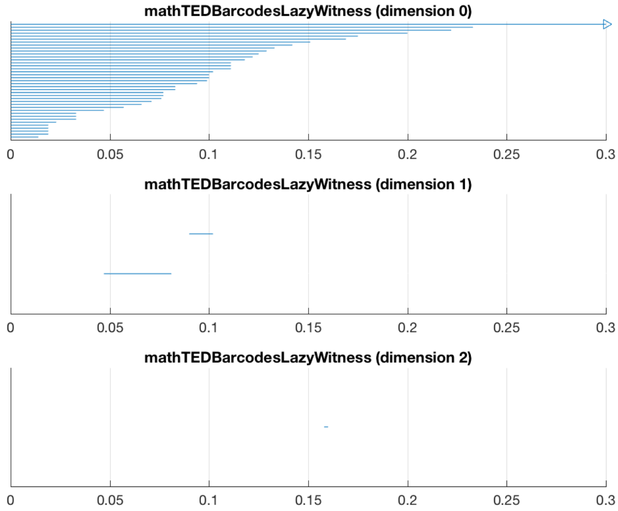

Figure 2: Restricting to Math-Related Talks with Lazy Witness

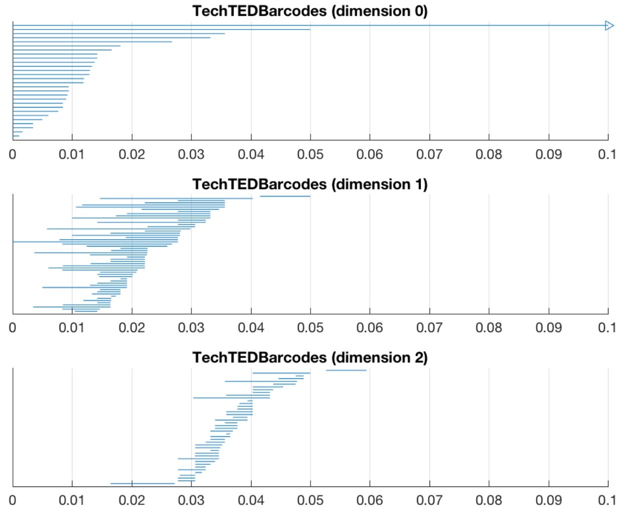

Figure 3: Restricting to Tech-Related Talks with Lazy Witness

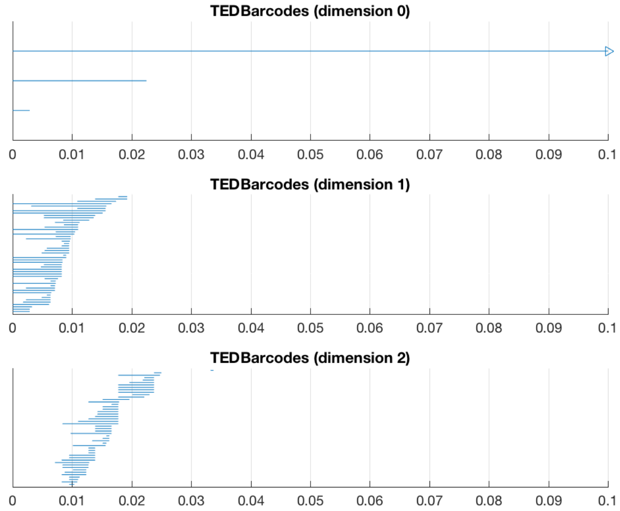

Figure 4: Returning to Full Collection of TED Talks with Lazy Witness

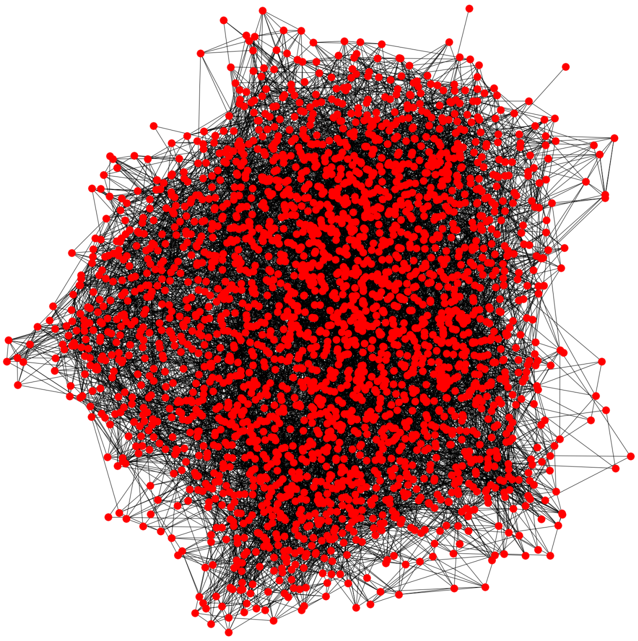

Figure 5: The network structure of "Related" Talks

4 Experiment 2

4.1 Data Model

In the second experiment conducted, the kaggle dataset was again analyzed as a collection of TED talks, but in this experiment each TED talk was modeled as a “polygon-of-tags”. That is, for a given TED talk, its set of tags was modeled as a 2-cell. This 2-cell was constructed by filling the “polygon” defined by all edges of a talk, where any two tags attributed to the talk form an “edge” or 1-cell.

By defining the topological space as that of a cellular complex, this experiment aimed to analyze the more explicit structure defining how each talk connects with other talks in the space. For example, with this modeling methodology, the shape of two talks connecting in the space is that of two polygons that share edges and/or vertices. The topology of the whole space thus provides how much these polygons fit together into larger assemblages possibly with a number of components, cycles, or voids.

4.2 Persistence Homology Computation Methods

In this experiment, a cellular complex was constructed from the cells of each talk via explicit programming. Specifically, the persistent homology of the explicitly constructed cellular complexes was computed using three different filtrations: a trivial filtration, a frequency based filtration, and a time based filtration.

These filtrations were chosen to accentuate or embed additional structural features concerning the history of the conceptual tags attributed to each TED talk. First, the trivial filtration present a static viewpoint on the structure of of the "polygon-of-tags" space, by including all cells of the space at the same filtration value. Second, the frequency based filtration embeds the how frequent a tag or connection between two tags occurs, over the thirteen year period spanned by the dataset, into the 0-cells or 1-cells, respectively. Specifically, cells are included at filtration value of $\frac{1}{freq}$, where $freq$ is the frequency of the associated tag or co-occurance of two tags. Thus, for example, if a particular tag occurs in three separate talks throughout the dataset, the tag's 0-cell has a filtration value of $\frac{1}{3}$. Lastly, a additional filtration imposed on the tags and co-occurance of two tags was based on the earliest date that a TED talk associated with a tag or pair of tags was filmed. The filtration value used in this construction is standardized to the interval of $[0, 1]$, by substracting the TED talk whose filming date was the earliest and dividing by the last filmed talk in the collection. So, for example, consider Nicholas Negroponte's talk "5 predictions from 1984" filmed in 1984. Negroponte's TED talk is one of the earliest TED talks in the collection and hence the tags associated with the talk ('demo', 'design', 'entertainment', 'future', 'interface design', 'media', 'movies', and 'technology') all appear for the first time at roughly 0, as do the co-occurances of any two tags in this set.

The results of these computations follow.

4.3 Results

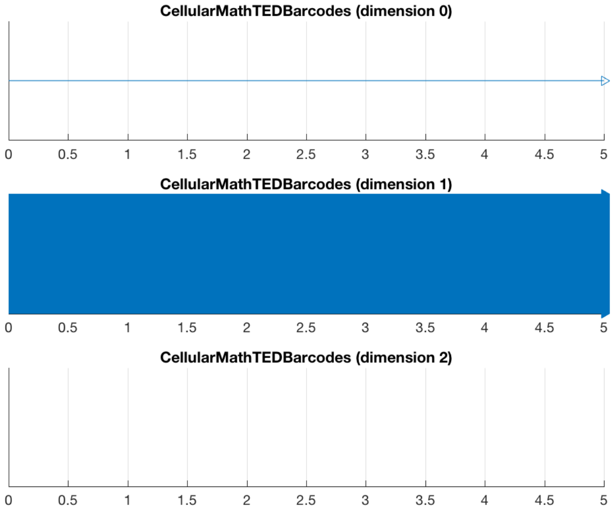

Figure 6: Results for Cellular Complexes Over Math-Related Talks

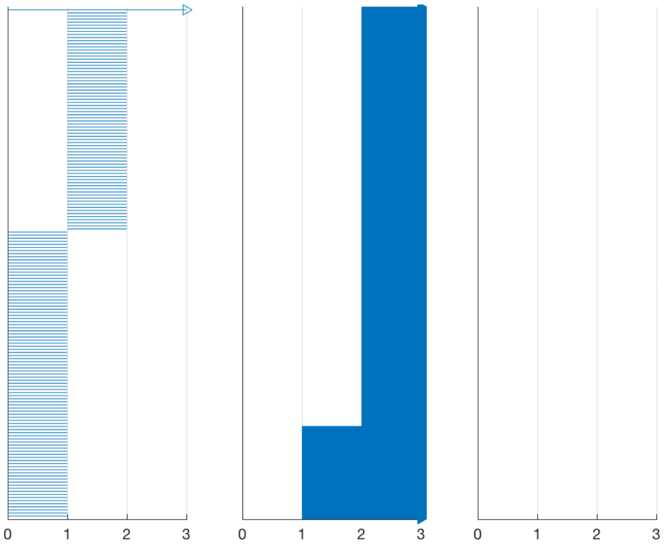

Figure 7: Results for Math-Related Talks with Frequency Filter

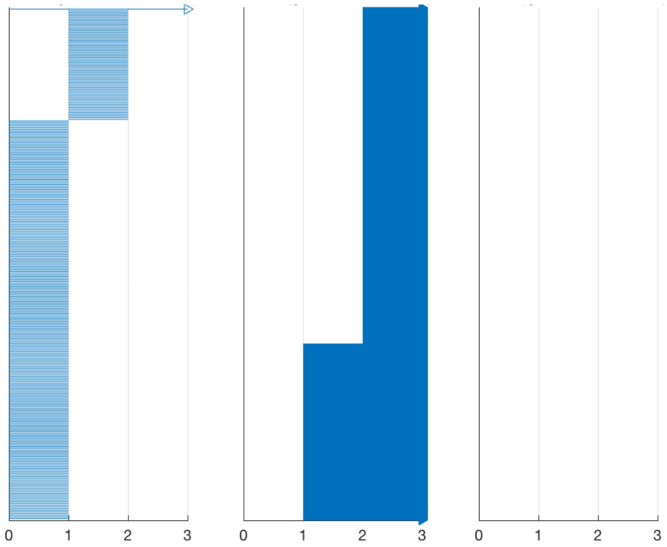

Figure 8: Results for Design-Related Talks with Frequency Filter

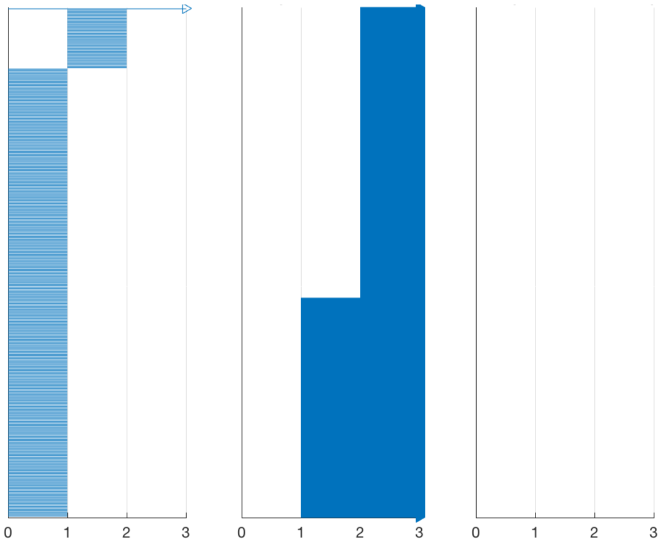

Figure 9: Results for Tech-Related Talks with Frequency Filter

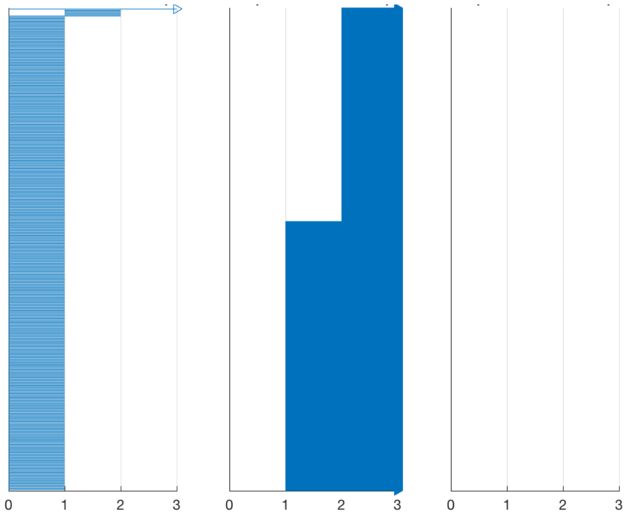

Figure 10: Results for all TED Talks with Frequency Filter

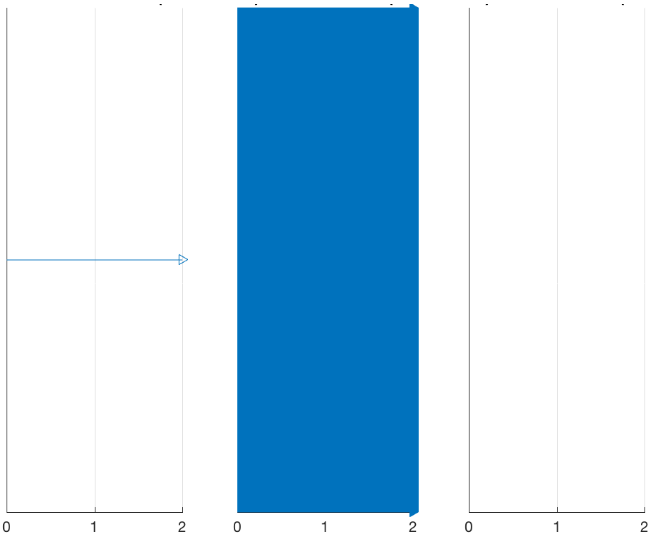

Figure 11: Results with Time-based Filtration

5 Experiment 3

5.1 Data Model

The final experiment analyzed the kaggle dataset with respect to the tags associated with the talks so that each tag was modeled as a “set-of-talks”. That is, a given conceptual tag is represented as a set of all talks in the kaggle dataset that are catalogued with this tag.

Again, utilizing the natural dissimilarity measure given by the dual of the Jaccard coefficient, a “pseudo” metric space can be constructed from the tags. By returning to the vector space modelling of the first experiment, simplicial complex constructions provide an intuitive way to envision the shape of connections between the ideas that have been drawn together by the speakers at TED. The topology of this space more clearly indicates the opportunities for interesting new talks in the nexus of ideas never before connected.

5.2 Persistence Homology Computation Methods

Using the “metric space” of the set of talks for each tag, it is possible to define various simplicial complexes over this underlying topological space. Because there were a maximum of 416 points in the space, Vietoris-Rips constructions provide a computationally accessible method to compute the persistent homology of the data, with the added benefit of being straight-forwardly interpretable.

Thus, with the underlying space of TED talk tags, the homology of chains of this space over $\mathbb{Z}/2\mathbb{Z}$ was computed up to dimension 2. The results of these computations follow.

5.3 Results

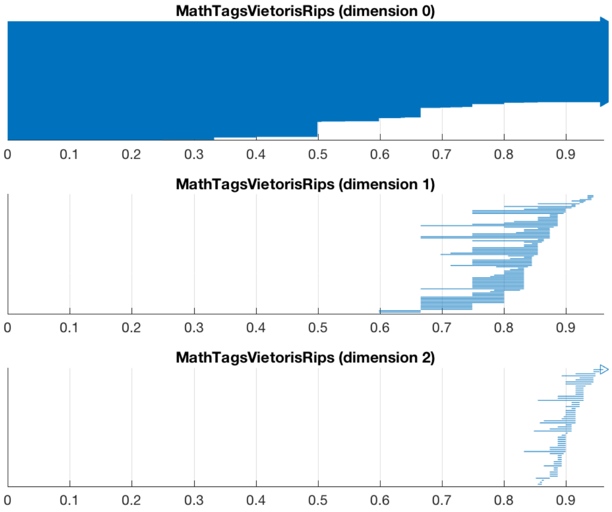

Figure 12: Results for Math-Related Tags

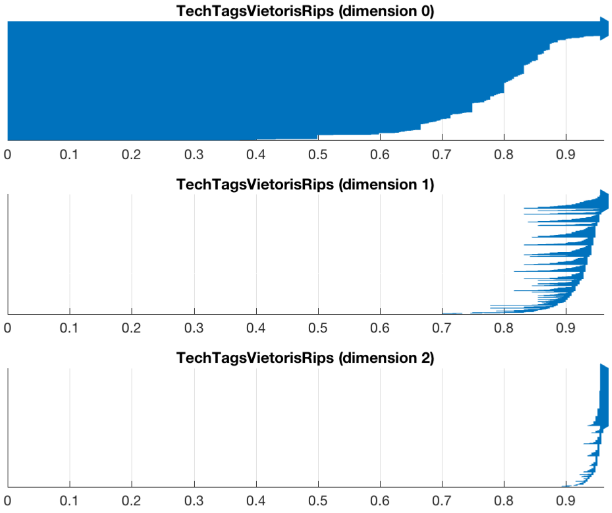

Figure 13: Results for Tech-Related Tags

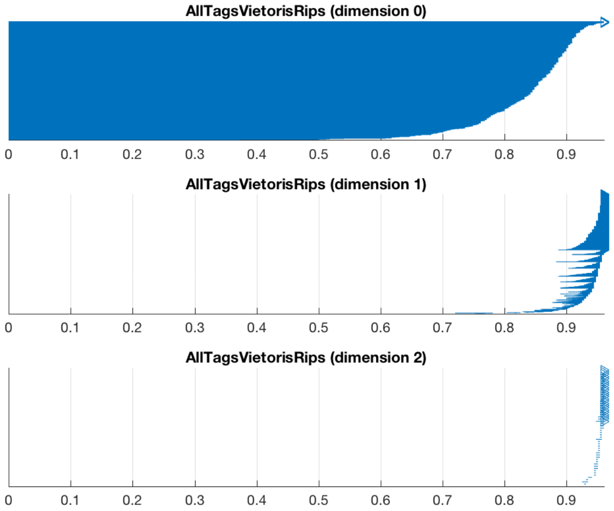

Figure 14: Results for all TED Talks Tags

6 Analysis and Conclusions

The set of all TED talks from 1984 to the present is both sufficiently long and wide to pose computational difficulties for analysis, but, in the first experiment, lazy witness complexes provide a sufficient reduction in computational load for results to be computed on a consumer-level personal computer.

Using the lazy witness construction to explore the persistence homology of TED talks, does reveal that the set of all TED talks is very densely connected. Under the pseudo-metric of the dual of the Jaccard complex, most points in the space are separated by the same "distance". Using a ν parameter of 2 in the lazy witness construction only presented a single barcode representing a single connected component. By setting the ν parameter to value less than two, additional barcodes are generated, but, as illustrated by figures 3 and 4, the space is still compacted to very small intervals of filtration values.

Restricting the dataset of Model 1 to only talks that used the tag of "Math", does reveal two 1-cycles of significant length. These barcodes are also reflected in the lazy witness barcodes for math-related talks, further justifying that there are two 1-cycles in the data. These cycles represent a collection of talks that are all sufficiently different from one another that there is a conceptual space than none of them all share.

Looking at the results for experiment 2, it appears that there are perhaps many conceptual cycles that no math-related talks explictly inhabit. That is, the high number of 1-cyles present when using the "polygon-of-tags" model point to the fact that tags of math-related talks link together into a very porous structure. While it is not surprising that the the conceptual web of ideas for math-related talks should be tangled, this high porosity indicates that TED catalogue of tags for each talk is sufficiently detailed enough to result in this sort of web complexity.

Using a frequency filtration on the cellular complex constructions of experiment 2, mainly reveals that math-related talks are highly explorative and innovative. This is indicated both by the few number of talks related to math (49) but also by the high number of tags that have a filtration value of 1 and co-occurrence of tags with filtration value 2. These features are not as striking present in either the design-related talks or the whole TED talk corpus.

Finally, experiment 3's construction of a space of tags provides a very intuitive way to conceptualize the connectivity of ideas. Looking at figures 12 through 14, many significant 1-cycles are very readily apparent. These conceptual voids were not as easily picked out through the modeling and analysis of experiment 1. Moreover, this dual structure to experiment 1, allows for Vietoris-Rips constructions that make direct interpretations of the basis of these 1-cycles easy. This accurate interpretability of both Vietoris-Rips and experiment 3's modeling approach makes figure 15 possible.

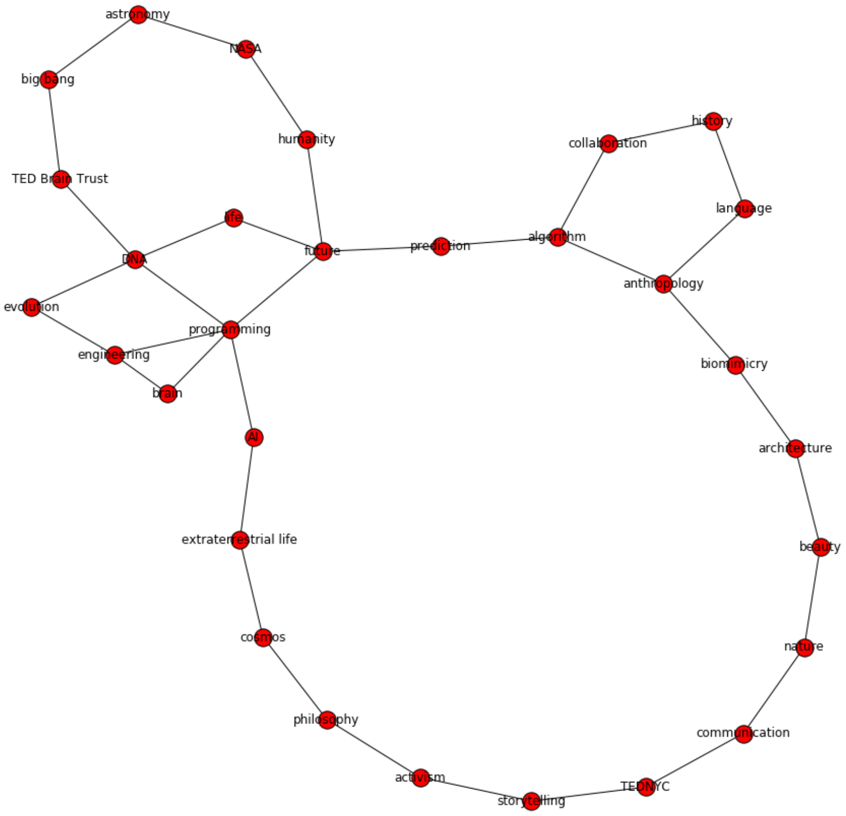

Figure 15: Conceptual network of 1-cycles in math tag “sets-of-talks” space

References

Rounak Banik. Ted data analysis, 2017. URL https://www.kaggle.com/rounakbanik/ted-data-analysis.

Jeremy Kun. Computing homology, 2013. URL https://jeremykun.com/2013/04/10/computing-homology/.

Wikipedia. Jaccard index. URL https://en.wikipedia.org/wiki/Jaccard_index.

H. Adams and A. Tausz. Javaplex tutorial, 2017. URL http://www.math.colostate.edu/~adams/research/javaplex_tutorial.pdf.