Topological Data Analysis

Deepak Choudhary(11234) and Samarth Bansal(11630)

April 25, 2014

Contents

ABSTRACT

Topology is the branch of mathematics which concerns itself with the study of shapes and its properties. The first major work in this area was done by Leonhard Euler in his 1736 paper on the Seven Bridges of Knigsberg. Topology has seen applications in the field of data analaysis in the last century. We here try experimenting on a heart disease dataset[4] using an algorithm, which has algebraic topology at its core, called topological data analysis.

1 Introduction

A topological space is the most general notion of a mathematical space and it consists of a set of points, along with a set of neighbourhoods for each point, that satisfy a set of axioms that relate points and neighbourhoods.

Topological data analysis is used for robust analysis of scientific data.

Definition 1. *Given a finite dataset $S \subseteq Y$ of noisy points sampled from an unknown space X, topological data analysis recovers the topology of X, assuming both X and Y are topological spaces.*

We are using topological methods to analyze heart disease data set obtained from UCI Machine learning repository. Part I of the project is focussed on computing persistent homology of our data. Part II involves visualization of the data set with Mapper function.

2 Barcodes

Before moving on to the barcode analysis let us explain a few relevant basic definitions in topology.

2.1 Simplical Complexes

Definition 1. *Simplicial complex is a topological space of a certain kind, constructed by "gluing together" points, line segments, triangles, and their n-dimensional counterparts.*

A simplical complex is made up of simplexes that are subsets of set of vertices.

2.1.1 Representation of a k-simplex

	Name	Representation
0-simplex	vertex	$\{v\}$
1-simplex	line	$\{v_1, v_2\}$
2-simplex	triangle	$\{v_1, v_2, v_3\}$
n-simplex		$\{v_1, v_2, ..., v_n\}$

2.1.2 Creating simple simplical complexes

- $\Sigma \leftarrow \phi, V \leftarrow \phi$
- Start by adding 0-dimensional vertices to Σ and to V
- Add 1-dimensional edges to Σ. No loops are allowed and neither are multiple edges.
- Add 2-dimensional trainagles to Σ. Boundary of a triangle is a cycle consisting of 3 edges.
- Keep in mind the constraint that if $\sigma \in \Sigma$, then for any $\tau \subseteq \sigma \Rightarrow \tau \in \Sigma$

Then (V, Σ) is a simple simplicial complex of degree 2

2.1.3 Creating a simplical complex from point cloud

Definition 1. *Point Cloud: Finite metric space, that is a finite set of points, equipped with a notion of distance.*

Creating a simplicial complex from data

- Represent the data as a point cloud. These act as our 0-dimensional vertices.
- Add 1-dimensional edges. Add edge between data points that are close. By close, we mean any metric that measures closeness, which is largely dependent on the data set you are trying to study. eg It could be Euclidean norm etc. We need to define a certain threshold below which we will consider the points to be close.
- Then add higher dimensional simplicial complex.

How to add?
That depends on kind of simplicial complex we want to create, simplest one being Vietoris-Rips complex that adds all possible simplices of dimension > 1.

The choice of threshold determines our simplicial complex. We are not very clear on choice of the threshold, so we turn to understand persistent structure that retains itself for a large variation as we change threshold.

2.1.4 Persistent Homology

Persistent homology is a method for computing topological features of a space at different spatial resolutions. More persistent features are detected over a wide range of length and are deemed more likely to represent true features of the underlying space, rather than artifacts of sampling, noise, or particular choice of parameters.To find the persistent homology of a space, the space must first be represented as a simplicial complex, which we have discussed above.

2.1.5 Betti Numbers

Betti numbers are used to distinguish topological spaces based on the connectivity of n-dimensional simplicial complexes. Basically they record significant topological features of the shape.

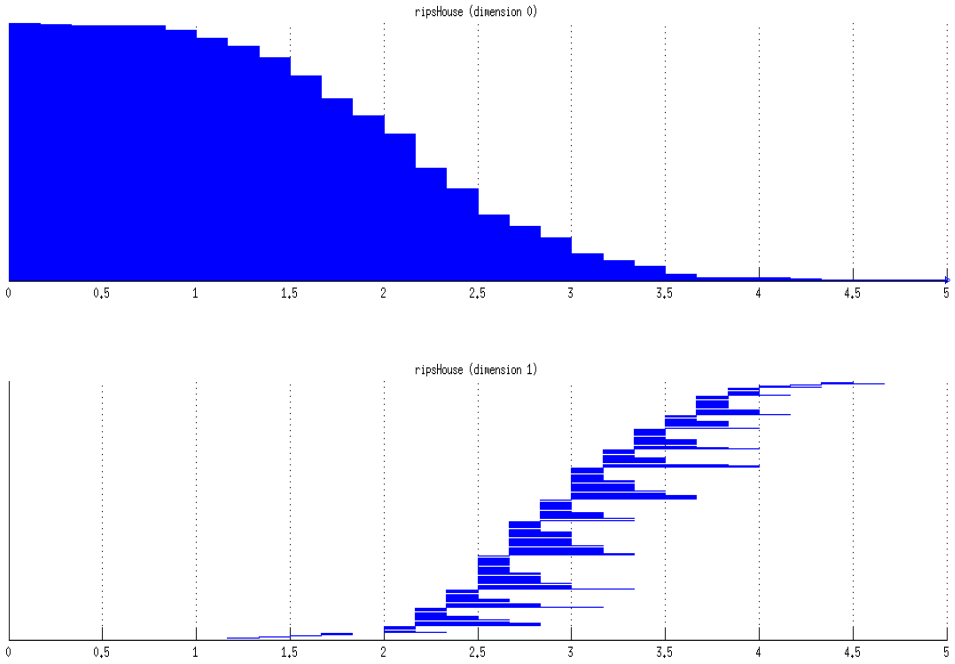
The kth Betti number refers to the number of k-dimensional holes on a topological surface.

- b_0 is the number of connected components
- b_1 is the number of one-dimensional or "circular" holes
- b_2 is the number of two-dimensional "voids" or "cavities"

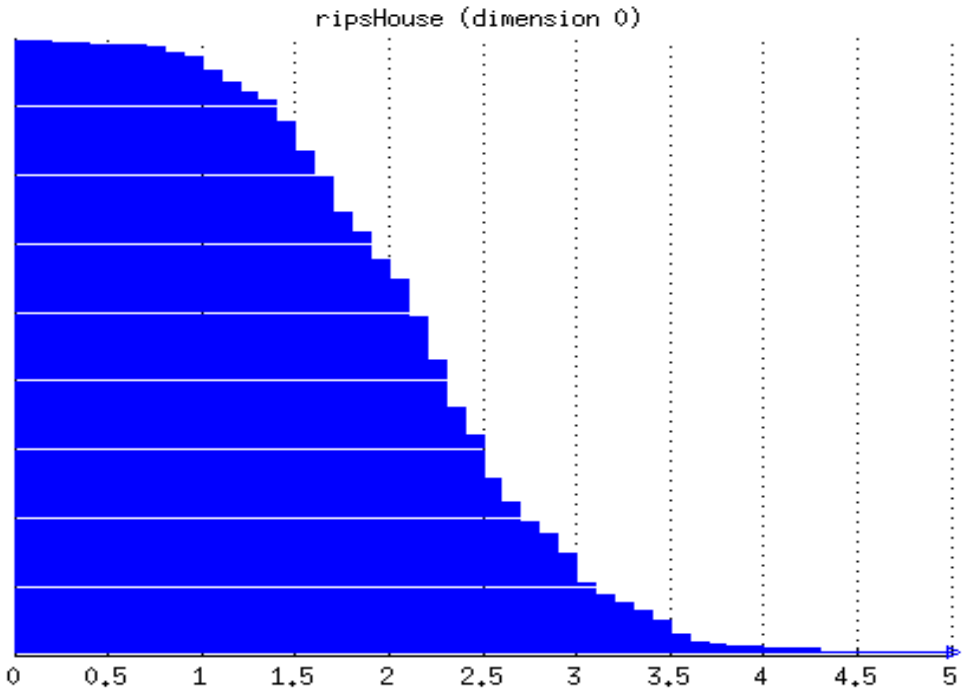
	b_0	b_1	b_2
Circle	1	0	0
Taurus	1	2	1
Sphere	1	0	1

2.1.6 Barcode Analysis

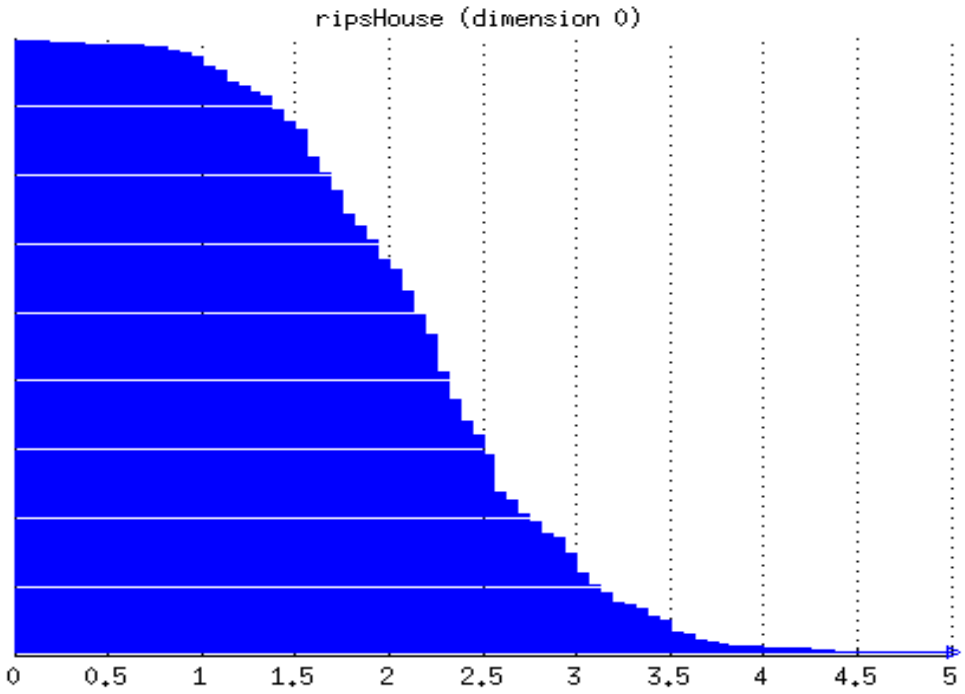
Our herat disease data set has 303 data points, each having 14 features. We plotted the barcode to compute the betti numbers and hence to understand the persistant features in this data. Following was the plot obtained:



From the figure, $b_0 = 1$ and $b_1 = 0$. The barcode plots show the connectedness of our data set.



This figure has been plotted only for 1 dimension at num-of-div=50



This figure has been plotted only for 1 dimension at num-of-div=80

3 Mapper

Mapper is an algorithm for describing high-dimensional datasets in terms of simple geometric objects. To be particular, representing our data as simplicial complexes, as discussed in the previous section. Mapper clusters the data partially in accordance with set of functions dened on the data. The method bear similarities to density clustering trees, disconnectivity graphs and reeb graphs but is a more generalised approach. Our goal remains to recover a low-dimensional representation of a point cloud, that is, to create a good plot.

3.1 Visualization Techniques

Multidimensional scaling (MDS) is a means of visualizing the level of similarity of individual cases of a dataset. It refers to a set of related ordination techniques used in information visualization, in particular to display the information contained in a distance matrix. Related developments are the Isomap and locally linear embedding.

The result in a point cloud in R^2 or R^3 , which can then be visualized by scatter plots and other techniques. But we can visualize our data by a geometric representation. One such possibility is representation as a graph or as a higher dimensional simplicial complex. Mapper is suited for this purpose.

Mapper has the following properties:

- **Metric Independent** : Our imaging method should not be sensitive to the notion of distance metric, because given the variety of data available to us, there is no one suitable distance measure, and is highly dependent on type of data.
- **Changes with parameter values** : Our algorithm should not be restricted in terms of parameter choices, and we should have the flexibility of changing parameters to understand different behaviors.
- **Multiscale Representations** : Features that are represented in multiple scale resolution are more interesting and convey useful information about our data, than features which change upon changing resolution.

3.2 Algorithm

1. We start with a function f : X $\rightarrow$ R whose value is known for the N data points. We call this function a filter. The function should convey some in- teresting geometric or other, for the task at hand relevant, properties of the data.
2. Finding the range (I) of the filter f restricted to the set X and creating a cover of X by dividing I into a set of smaller intervals (S) which overlap. This gives us two parameters which can be used to control resolution namely the length of the smaller intervals (l) and the percentage overlap between successive intervals (p). [1].
3. "Now, for each interval $I_j \in S$, we find the set $X_j = \{x|f(x) \in I_j\}$ of points which form its domain. The set $\{X_j\}$ forms a cover of X, and $X \subseteq \cup_j X_j$." [1].
4. Choosing a metric $d(.,.)$ to get the set of all interpoint distances $D_j = \{d(x_a, x_b)|x_a, x_b \in X_j\}$

5. For each X_j together with the set of distances D_j we find clusters $\{X_{jk}\}$

6. Each cluster then becomes a vertex in our complex and an edge is created between vertices if $X_{jk} \cap X_{lm} = \emptyset$ meaning that two clusters share a common point.

3.2.1 Choices to be Made

We have to make various parameter choices whie implementing mapper function. We are listing down the choices we made while performing analysis on heart disease data set. Our approach was to compute using various choices and figure out the parameters that give us interesting results.

1. Choose how to cluster, and for that, we need to define a distance metric. As mentioned earlier, any measure of closeness can be chosen. We normalized our data using standard score and chose two metrics 1) Modified Euclidean Norm as a distance metric with a factor parameter for each feature. 2) Pearson correlation cofficients.

2. Choosing our filter function
   Filter function is mapping of n-dimensional data set to a lower dimesnions. e.g. f(x, y, z) $\rightarrow$ x
   Other possibilities :
   b. L_∞ centrality
   c. Vector Norm

   In our analysis, we experimented with following functions:
   a. L_∞ centrality [1]
   b. On the basis of individual data features eg Age, Cholestrol.
   c. On the basis of distance from a given fixed node.

3. Binning
   Put data into overlapping bins. We have to choose the length and the percentage overlap. We took 25, 50 and 75 percentage overlap and number of bins to be 3, 5 and 10 as different cases.

Different choice of parameters may generate a network with a different shape and thus allowing one to explore data from a different perspective.

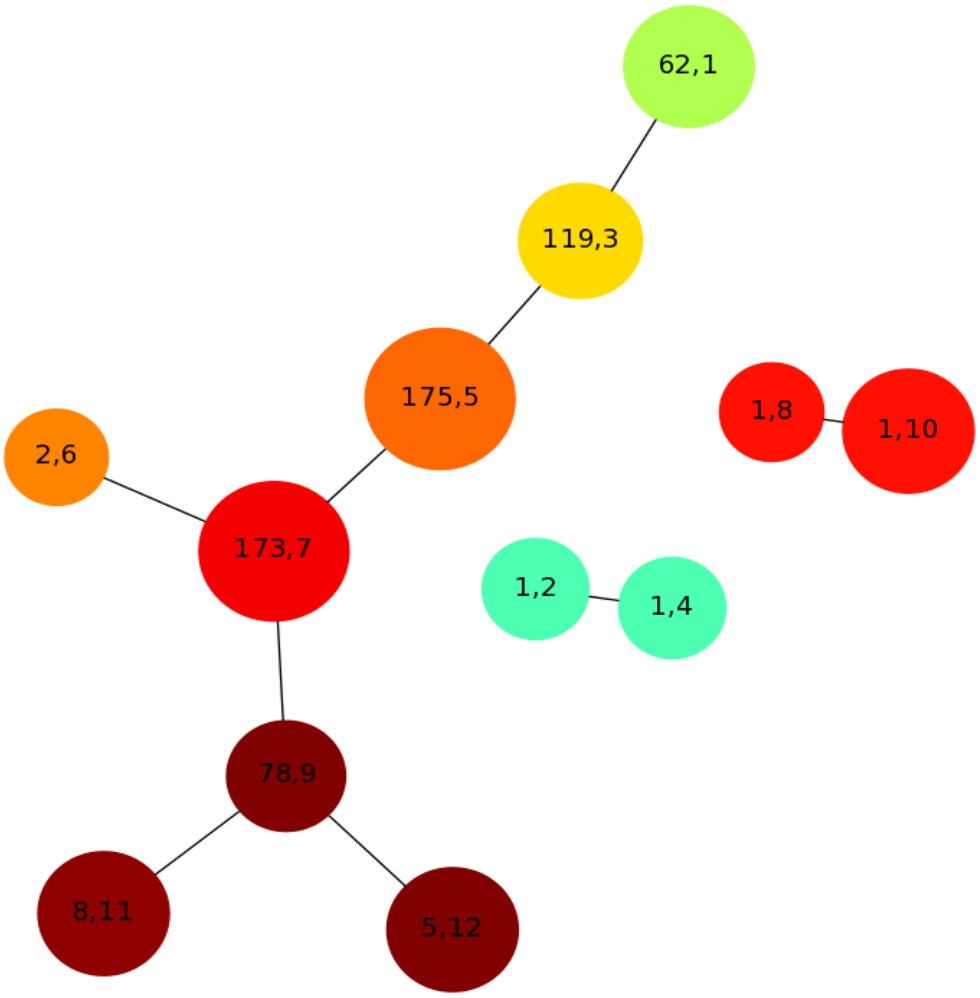
We performed analysis on heart disease data set. Data set was taken from [4]
The data has 303 data points, each having 14 features.

3.3 Mapper Analysis

We are including 3 images, which we found to be signifcant that represent the structure of data. They correspond to following choices:

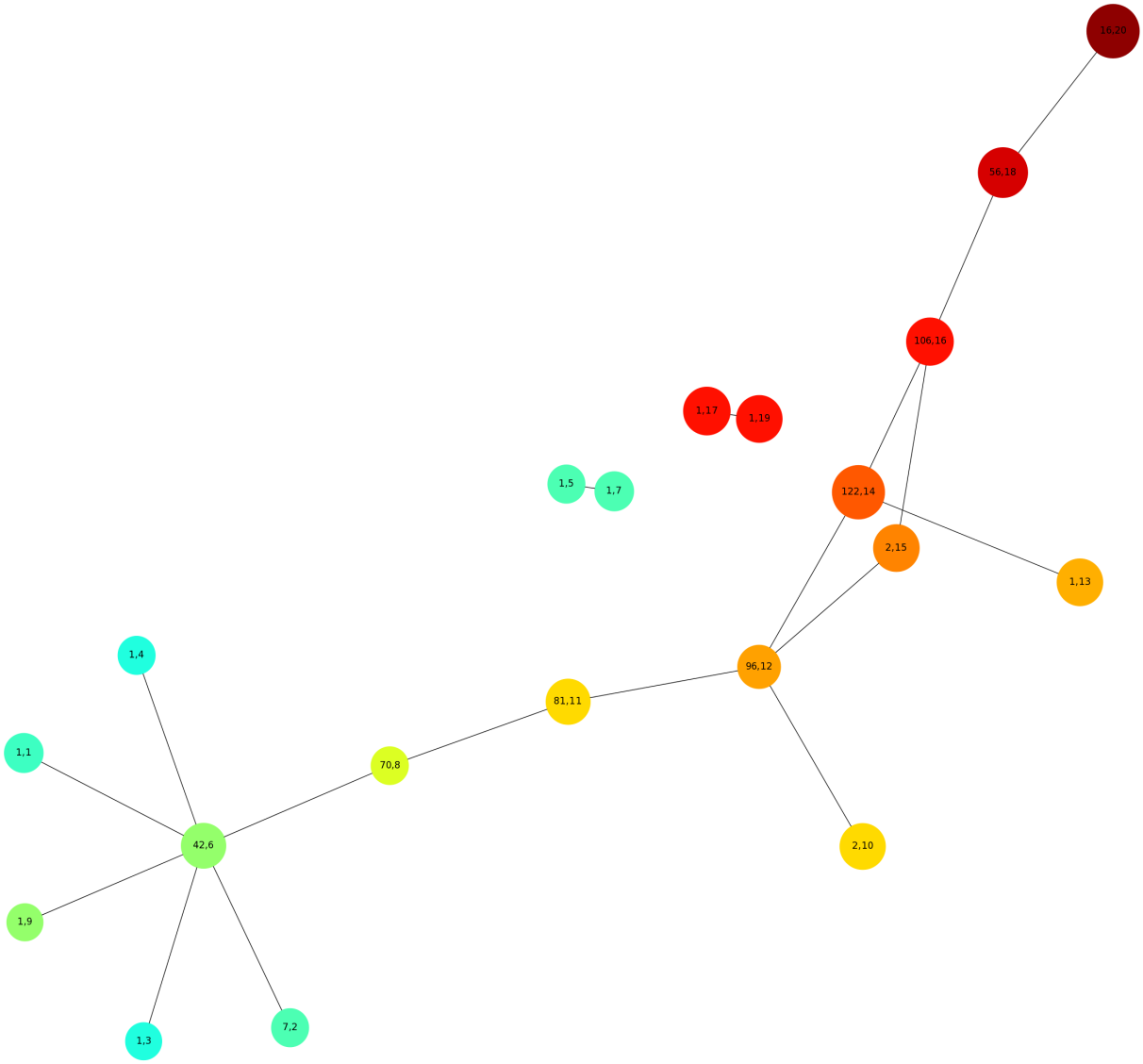
a. Filtered in accordance with age parameter. That is, age increases as color ranges from red to blue.
b. Euclidean Norm
c. Percentage overlap is 50

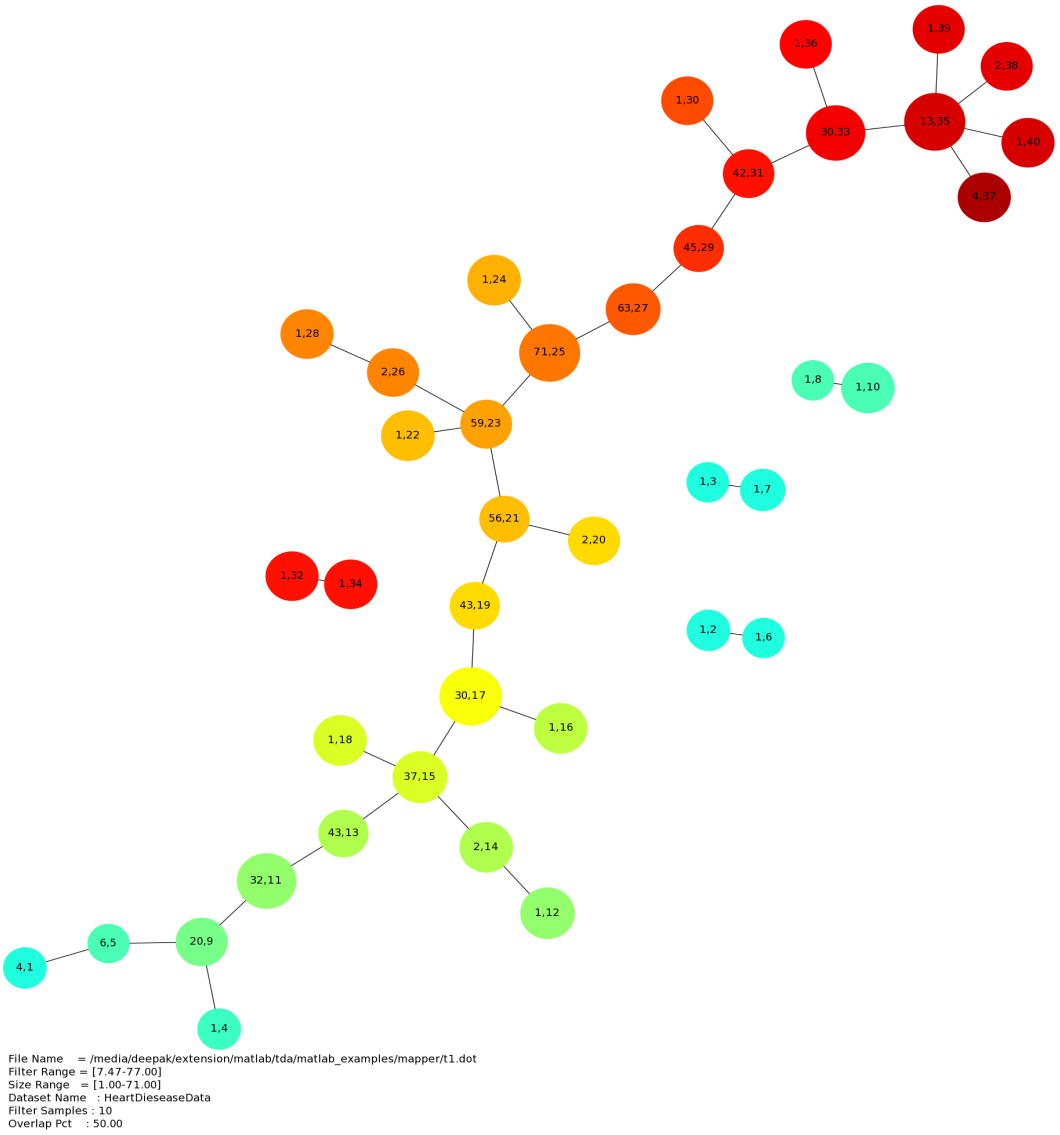
[1] This has its usage in specific kind of data with braches which is explained by Carlsson in [2]



File Name = /media/deepak/extension/matlab/tda/matlab_examples/mapper/t1.dot
Filter Range = [7.24-77.00]
Size Range = [1.00-175.00]
Dataset Name : HeartDieseaseData
Filter Samples : 3
Overlap Pct : 50.00

Each node represents a cluster of points and two nodes having atleast one point in common have a connection.
The first number on the each node represents the number of data points it is representing and the second nunber is just an index for each node.


16,20
56,18
106,16
1,17
1,19
1,5
1,7
122,14
2,15
1,13
1,4
96,12
81,11
1,1
70,8
42,6
2,10
1,9
1,3
7,2
File Name = /media/deepak/extension/matlab/tda/matlab_examples/mapper/t1.dot
Filter Range = [7.28-77.00]
Size Range = [1.00-122.00]
Dataset Name : HeartDieseaseData
Filter Samples : 5
Overlap Pct : 50.00



4 Future Works

There is a lot to explore in this field. It seems that questions are many but we were able to work only on a few. Further works are suggested below:

- We can use more variants for distance metric like modified L_1 norm and cosine distance.
- Existing possibilies for even more filter functions. for eg 1) a filter function which is a linear combination of two of more important attributes rather than just one or even a filter function oof top 5-6 eigenvectors from PCA. 2) a density estimator filter function
- Better insights into theoritical understanding of Betti numbers(or topology) will lead to better conclusions from barcode analysis

- Associate and display statistical data of each nodes on the mapper graph. In case of images instead of statistical data of node rather should represent the image in small size. This would increase the speed of analysis and really deepen the understanding of the mapper algorithm. Something similar on this has been shown in Carlsson's paper [2] page 30.
- After applying mapper and getting clusters, again applying mapper on each node. This would be a very nice recusive approach to understand large data sets.
- We can compare this heart disease data set to another heart disease data set from a different source.

5 Conclusions

5.1 $Betti_0$

After studying betti 0 on $num - of - div = 80$ we see that the modulus of slope of graph is small initially which increases and the again decreases. This leads us to the conclusion that most point have distances that lie in between 1 to 3 as can be seen from $betti_0$ figure.

6 Softwares Used

1. JavaPlex
2. Mapper (Maintained by Comptop.stanford.edu)
3. Matlab
4. Graphviz

References

[1] Gurjeet Singh, Facundo Mmoli, and Gunnar Carlsson. Topological methods for the analysis of high dimensional data sets and 3d object recognition. In Eu- rographics Symposium on Point-Based Graphics, volume 22. The Eurographics Association, 2007.

[2] Gunnar Carlsson. Topology and data. Bulletin of The American Mathematical Society, January 29 2009.

[3] Lum, P.Y., Singh, G., Lehman, A., Ishkanov, T., Vejdemo-Johansson, M., Alagappan, M., Carlsson, J., Carlsson, G.: Extracting insights from the shape of complex data using topology. Sci. Rep. 3 (2013).

[4] Heart Disease Data Set, Hungarian Institute of Cardiology. Budapest: Andras Janosi, M.D.