CS 6965 Advanced Data Visualization
Project 1

Yulong Liang

January 30, 2018

2 The Cat Example (2 pts)

Q1: Some of the small clusters in 20 percent graph will **disappear** in 80 percent graph, while other clusters will grow **larger** than before. Moreover, the edges between clusters are **shorter** in 80 percent graph because the length of edges represent the force between two clusters, i.e., the overlap between two clusters.

Q2: The total number of clusters increases and the average size of each cluster decreases.

3 The Bunny Example (3 pts)

Q3: In the three dimensional point cloud, one of the ear of the bunny is separated from the rest of the body. When clustering, there are **no overlaps** between that ear and the rest of the body. Thus it results in two connected components.

Q4: The lens of the data changes from `sum` to the projection to `Y-axis`.

Q5: Yes. For each interval/cube in the space, the algorithm divides the points into two clusters, which leads to 30 clusters instead of 15.

4 The Digits Example (5 pts)

Q6: Since `t-SNE` algorithm initial the data points in the lower dimensional space stochastically, the location of the neighborhoods (a collection of data points

which are neighbors) is not identical for each experiment. When using `DBSCAN` algorithm to cluster the neighborhoods, the ones that are closed to each other will gather into one cluster. Because of the variance of the locations provided by `t-SNE`, the `DBSCAN` results are different from time to time.

Q7: The result using `Spectral Embedding` produced less and unevenly distributed clusters and within each cluster, the images of different digits are mixed. It did not generate a satifying result.

Q8: Modifying the parameter to `n_components=3`. The projected subspace generated by `Spectral Embedding` was changed from 2-dim to 3-dim, which led to better separated data points. The `mapper` algorithm will then convert a 3D subspace to a 2D visualization.

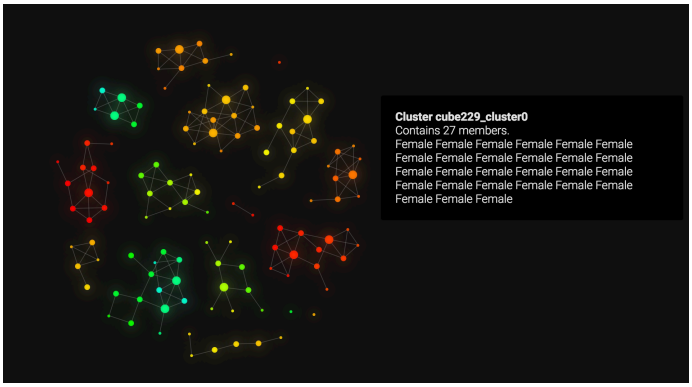
5 Your Own Dataset (5 pts)

On the Kepler Mapper website, the developer disclosed their future examples: Iris, Diabetes 100k, and Customer purchase behaviour. The first two datasets can be found on UCI Machine Learning Repository, so I picked them to analyze.

1. **Iris**

Iris Setosa can well separated from the other two species in terms of the length and width of petal and sepal. For species of *Iris virginica* and *Iris versicolor*, although a large amount of the flowers are clustered purely, there exists some flowers that are not separable. Those flowers have similar attributes but are different species.

2. **Diabetes 100k**



Diabete patients are not separable in terms of **age** and **race**. However, male patients and female patients never get clustered into the same group, which means diabetes vary quite a lot in terms of **gender**.