Name : Pawan Suresh Bathe Roll No.: CS18M519

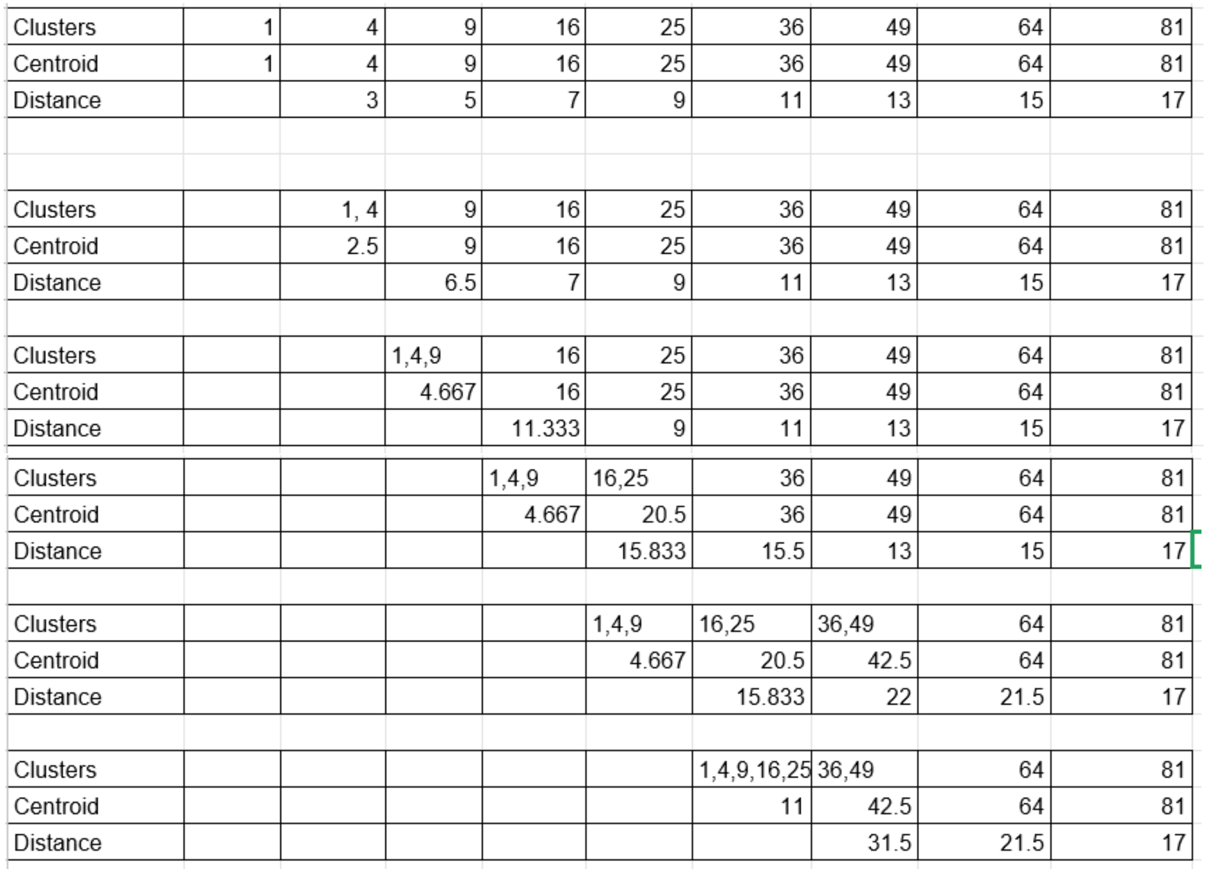
CS6745: Mining Massive DataSets

Tutorial 8

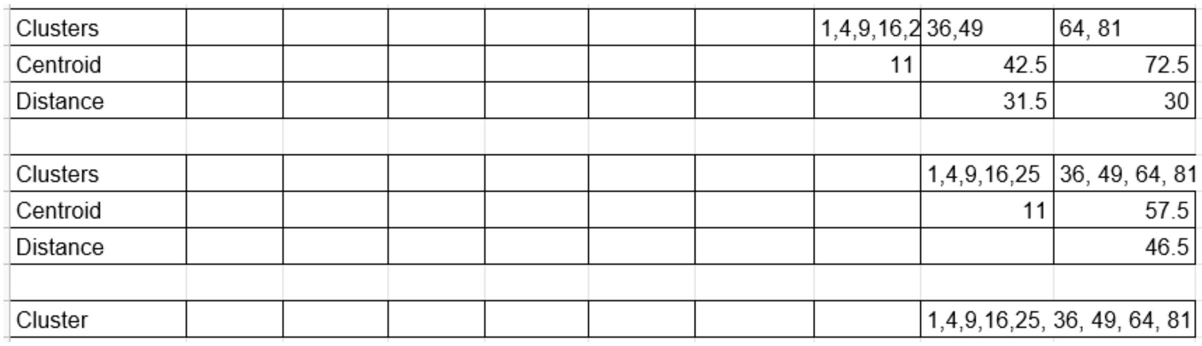
November 5, 2019

- Write your name and roll number in the space provided
- Be neat, and use the space judiciously.
- **Rough sheets won't be evaluated.**

1. (3 marks) Perform a hierarchical clustering of the one-dimensional set of points 1, 4, 9, 16, 25, 36, 49, 64, 81, assuming clusters are represented by their centroid (average), and at each step the clusters with the closest centroids are merged.



Clusters	1	4	9	16	25	36	49	64	81
Centroid	1	4	9	16	25	36	49	64	81
Distance		3	5	7	9	11	13	15	17
Clusters		1, 4	9	16	25	36	49	64	81
Centroid		2.5	9	16	25	36	49	64	81
Distance			6.5	7	9	11	13	15	17
Clusters			1,4,9	16	25	36	49	64	81
Centroid			4.667	16	25	36	49	64	81
Distance				11.333	9	11	13	15	17
Clusters				1,4,9	16,25	36	49	64	81
Centroid				4.667	20.5	36	49	64	81
Distance					15.833	15.5	13	15	17
Clusters					1,4,9	16,25	36,49	64	81
Centroid					4.667	20.5	42.5	64	81
Distance						15.833	22	21.5	17
Clusters						1,4,9,16,25	36,49	64	81
Centroid						11	42.5	64	81
Distance							31.5	21.5	17



Clusters							1,4,9,16,2	36,49	64, 81
Centroid							11	42.5	72.5
Distance								31.5	30
Clusters								1,4,9,16,25	36, 49, 64, 81
Centroid								11	57.5
Distance									46.5
Cluster								1,4,9,16,25, 36, 49, 64, 81	

Hierarchical clustering

2. (3 marks) As discussed in class, CURE is an efficient data clustering algorithm for large datasets. Is it robust to outliers in the dataset? Explain.

Yes, CURE is robust to outliers in the dataset. In CURE, fixed number of scattered points of a cluster are chosen and they are moved towards the centroid of the cluster. These new points which are moved towards centroid are used as representatives of the cluster. At each step, the clusters with the closest pair of representatives are merged. This enables CURE to correctly identify the clusters and makes it less sensitive to outliers. Because of having more than one point as representative of clusters , CURE can adjust to different boundary shapes and hence is robust towards the outliers.

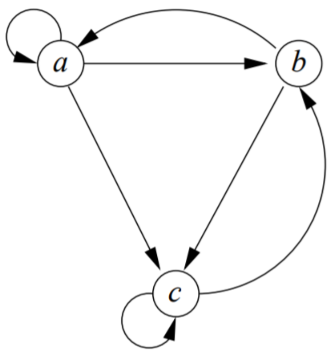


Figure 1: Example Web Graph

3. (a) (2 marks) Compute the PageRank of each page in Figure 1, assuming no taxation.

Assume ranks of a, b, and c are x, y , z respectively.

$$\text{Transition matrix M} = \begin{bmatrix} 1/3 & 1/2 & 0 \\ 1/3 & 0 & 1/2 \\ 1/3 & 1/2 & 1/2 \end{bmatrix}$$

Solving this matrix under the constraint of x + y + z = 1.
We get
rank of a = x = 4/10 = 2/5
rank of b = y = 3/10
rank of c = z = 3/10

(b) (2 marks) Compute the PageRank of each page in Figure 1, assuming $\beta = 0.8$