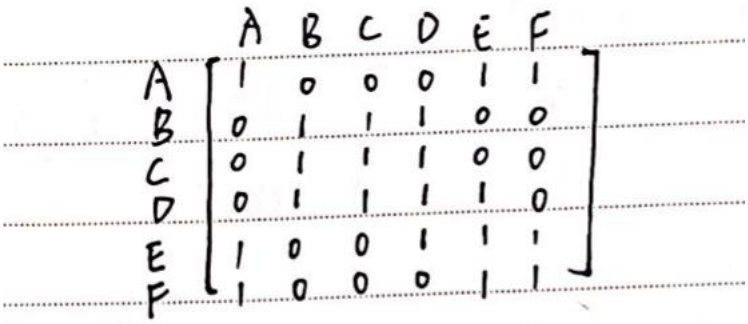
Question 1:

a) Dissimilarity metric: we will use the adjacent relation to be our metric and represent them using binary variables To be more specific, if there is an edge to connect node i and j, then x_{ij} = 1. For the special case i = j, the value is also 1. For other cases, the value should be 0. Following this metric, we can have the following data matrix:

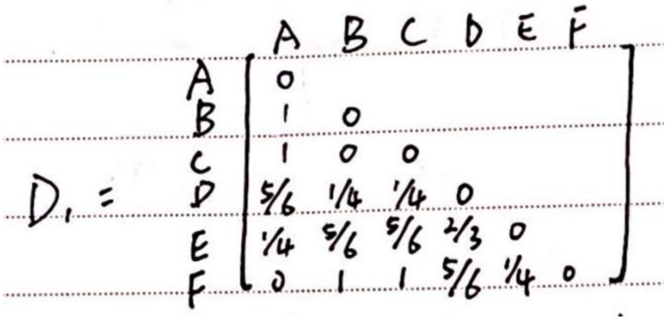


$$\begin{array}{c|cccccc} & A & B & C & D & E & F \\ \hline A & 1 & 0 & 0 & 0 & 1 & 1 \\ B & 0 & 1 & 1 & 1 & 0 & 0 \\ C & 0 & 1 & 1 & 1 & 0 & 0 \\ D & 0 & 1 & 1 & 1 & 1 & 0 \\ E & 1 & 0 & 0 & 1 & 1 & 1 \\ F & 1 & 0 & 0 & 0 & 1 & 1 \end{array}$$

The adjacent relation is an asymmetric variable. Therefore, we will adopt the Jaccard coefficient.

$$d(i,j)= \frac{b+c}{a+b+c}$$

Then, we can get the dissimilarity matrix:



$$D_1 = \begin{array}{c|cccccc} & A & B & C & D & E & F \\ \hline A & 0 & & & & & \\ B & 1 & 0 & & & & \\ C & 1 & 0 & 0 & & & \\ D & 5/6 & 1/4 & 1/4 & 0 & & \\ E & 1/4 & 5/6 & 5/6 & 2/3 & 0 & \\ F & 0 & 1 & 1 & 5/6 & 1/4 & 0 \end{array}$$

b) Following the matrix in a):

$$D_1 = \begin{array}{c|cccccc} & A & B & C & D & E & F \\ \hline A & 0 & & & & & \\ B & 1 & 0 & & & & \\ C & 1 & 0 & 0 & & & \\ D & 5/6 & 1/4 & 1/4 & 0 & & \\ E & 1/4 & 5/6 & 5/6 & 2/3 & 0 & \\ F & 0 & 1 & 1 & 5/6 & 1/4 & 0 \end{array}$$

$\min\{d_{ij}\} = d_{BC} = d_{AF}$, we join B, C.

$$D_2 = \begin{array}{c|ccccc} & A & (BC) & D & E & F \\ \hline A & 0 & & & & \\ (BC) & 1 & 0 & & & \\ D & 5/6 & 1/4 & 0 & & \\ E & 1/4 & 5/6 & 2/3 & 0 & \\ F & 0 & 1 & 5/6 & 1/4 & 0 \end{array}$$

$\min\{d_{ij}\} = d_{AF} = 0$. we join A, F.

$$D_3 = \begin{array}{c|cccc} & (AF) & (BC) & D & E \\ \hline (AF) & 0 & & & \\ (BC) & 1 & 0 & & \\ D & 5/6 & 1/4 & 0 & \\ E & 1/4 & 5/6 & 1/4 & 0 \end{array}$$

$\min\{d_{ij}\} = d_{(BC)D} = 1/4$, we join (BC), D.

$$D_4 = \begin{array}{c|ccc} & (AF) & (BCD) & E \\ \hline (AF) & 0 & & \\ (BCD) & 5/6 & 0 & \\ E & 1/4 & 2/3 & 0 \end{array}$$

$\min\{d_{ij}\} = d_{(AF)E} = 1/4$, we join (AF), E

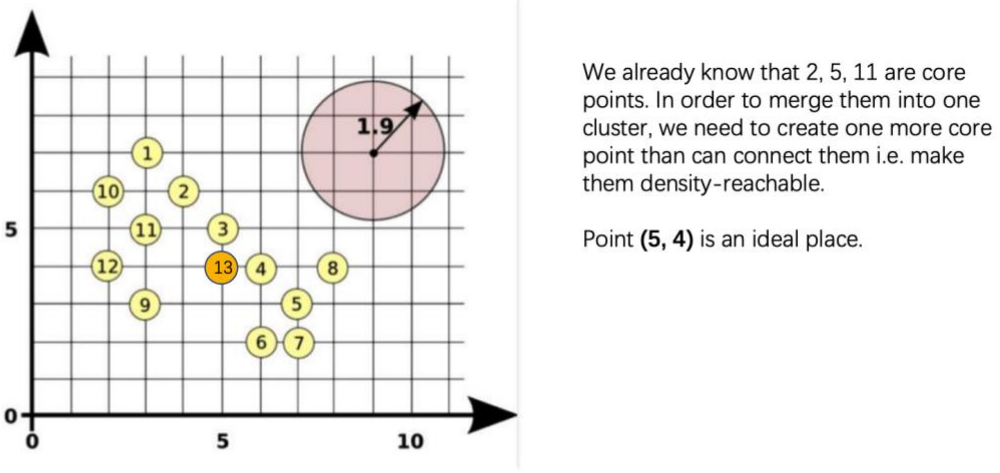
$$D_5 = \begin{array}{c} \\ (AFE) \\ (BCD) \end{array} \begin{array}{c} \begin{array}{cc} (AFE) & (BCD) \end{array} \\ \left[\begin{array}{cc} 0 & \\ 2/3 & 0 \end{array} \right] \end{array}$$

Then join (AFE) and (BCD). End.

Stage	Groups
P_1	A, B, C, D, E, F
P_2	A, BC, D, E, F
P_3	AF, BC, D, E
P_4	AF, BCD, E
P_5	AFE, BCD.
P_6	ABCDEF

Question 2:

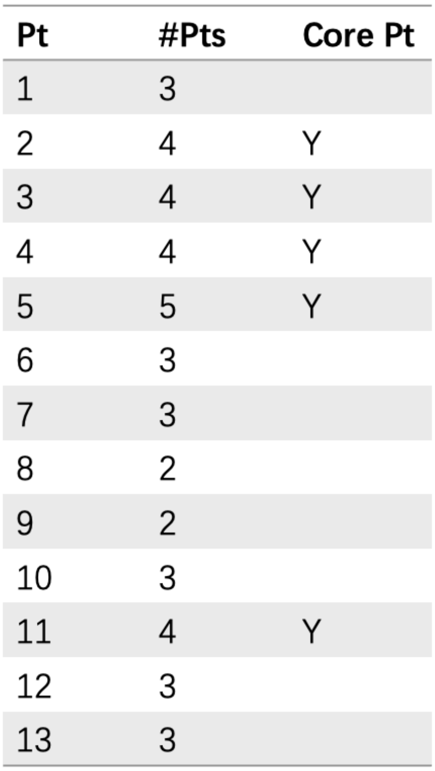
a)



We already know that 2, 5, 11 are core points. In order to merge them into one cluster, we need to create one more core point than can connect them i.e. make them density-reachable.

Point **(5, 4)** is an ideal place.

b)



Pt	#Pts	Core Pt
1	3	
2	4	Y
3	4	Y
4	4	Y
5	5	Y
6	3	
7	3	
8	2	
9	2	
10	3	
11	4	Y
12	3	
13	3	

Core points: 2, 3, 4, 5, 11
Border points: 1, 6, 7, 8, 10, 12, 13
Noise point: 9

c)

Theoretically, if e is large enough, there will no noise point. We are now trying to find a minimum e. The only noise point is 9 and its nearest core point is 11. If e=2, point 9 will be a border point.

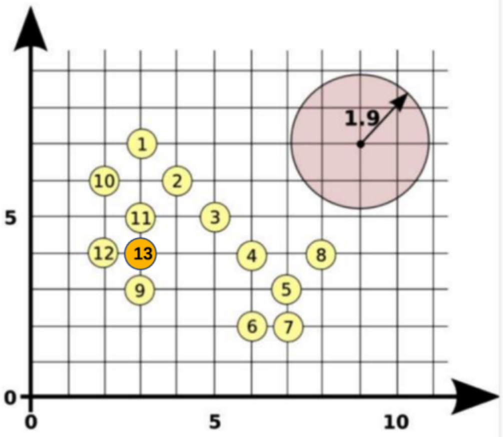
d)

If **<u>MinPts = 3</u>**, there will be only one cluster.

Pt	#Pts	Core Pt
1	3	Y
2	4	Y
3	3	Y
4	3	Y
5	5	Y
6	3	Y
7	3	Y
8	2	
9	2	
10	3	Y
11	4	Y
12	3	Y

Core points: 1, 2, 3, 4, 5, 6, 7, 11, 12
Border points: 8, 9

e)



Point **(3, 4)** is an ideal place.

Pt	#Pts	Core Pt
1	3	
2	4	Y
3	3	
4	3	
5	5	Y
6	3	
7	3	
8	2	
9	3	
10	3	
11	5	Y
12	4	Y
13	4	Y

Core points: 2, 5, 11, 12, 13

Border points: 1, 3, 4, 6, 7, 8, 9, 10

Question 3:

a) To combine the two clustering method, we can first do the hierarchical clustering and stop it at a certain condition. Then, we apply the k-means to continue cluster the nodes. The detailed procedure is as followed:

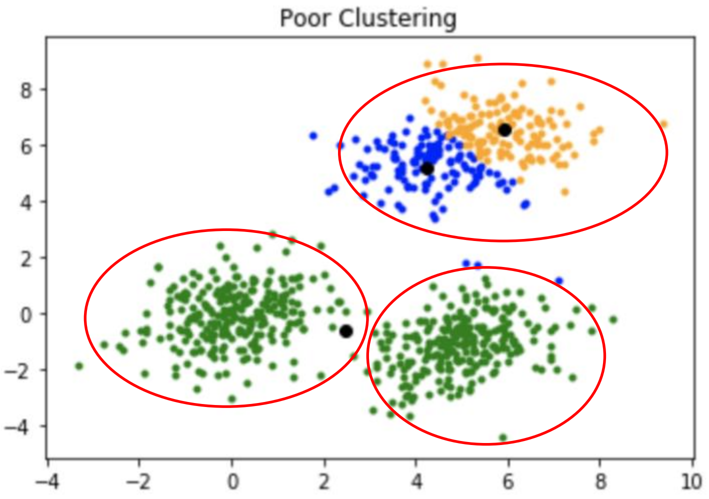
   a. Apply hierarchical clustering to divide the nodes into k-clusters
   b. Use the centers of the k clusters as the initial center
   c. Apply k-means using the centers found in step b.

   Following this way, the complexity is between O(tkn) and O(n^2). First, it is faster than O(n^2) because the second half of the clustering is completed by a more time-efficient k-means clustering method. Because of the same reason, it is slower than O(tkn) because the first half of clustering procedure is completed by hierarchical clustering.

b) Effectiveness: The most crucial weakness of k-means clustering is that it will randomly choose the initial central nodes. Different initialization will lead to different results. With some poor initialization, the clustering effectiveness may be weakened. On the other hand, hierarchical clustering's weak points include both time complexity and termination condition. As the clustering process goes, it may loss some original features. However, the combination of the two methods can **improve** the effectiveness by first leverage the hierarchical clustering to find the good initialization for k-means clustering, and then enhance the time-efficiency by applying k-means clustering.

   Speed: As discussed in the above section, the speed will be in between of the k-means clustering and hierarchical clustering.

c) When the initial centroids of the k-means clustering are not well positioned, my proposed algorithm can perform better. For example, in the following figure, it is obvious that there are three clusters (circled in red). However, because the initial centroids (black points) are not well positioned, the clustering result is poor. With my proposed method, the hierarchical clustering will be first used to find the approximately good centroids, and then start the k-means clustering.



https://www.geeksforgeeks.org/ml-k-means-algorithm/