Quiz 5

Online (In-class) Quiz –

1) [1 point] What are the assumptions about data we make when applying the BFR algorithm?

Clusters are normally distributed around a centroid in an Euclidean space

2) [3 points] Hierarchical Clustering: Imagine we are clustering the number of items bought on Amazon for a given user each day. We wish to perform a hierarchical clustering of the number of items bought among all users: 1, 3, 8, 16, and 30. Show what happens at each step **until there are two clusters**, and give these two clusters. Assume clusters are represented by their centroid (average), and at each step choose to merge two clusters whose resulting cluster has the **smallest diameter**.

Your answer should be in steps where each step shows the members of the new cluster formed, and its centroid. More specifically, if you are merging a cluster C1 = {x, y, z} of centroid c1 with a cluster C2 = {p, q} of centroid c2, you should report {x, y, z, p, q} in the table, as well as the centroid obtained with these 5 points).

Initial clusters : {1}, {3}, {8}, {16}, {30}
Centroids: 1, 3, 8, 16, 30
First step : {1, 3}, {8}, {16}, {30} [0.5 points]
Centroids: 2, 8, 16, 30 [0.5 points]
Second step : {1, 3, 8}, {16}, {30} [0.5 points]
Centroids: 4, 16, 30 [0.5 points]
Third step : {1, 3, 8}, {16, 30} [0.5 points]
Centroids: 4, 23 [0.5 points]

3) [0.5 point] Modularity Q is positive if the expected number of edges within the group is less than the number of edges within the group.

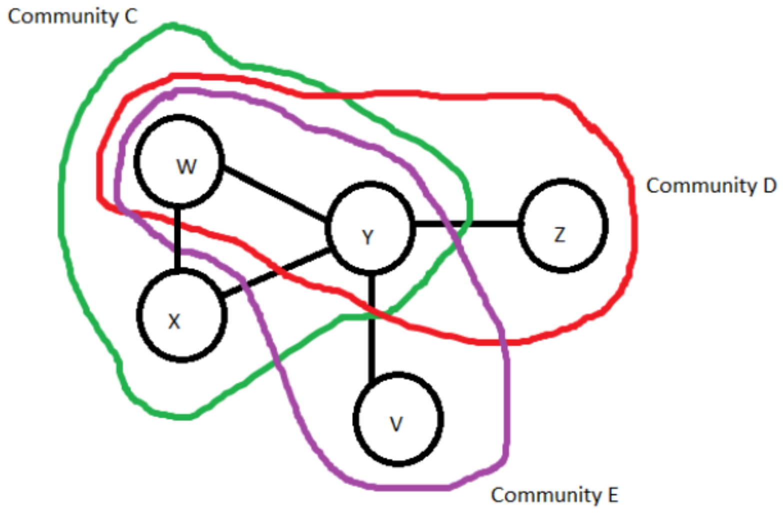
Answer - True

4) [0.5 point] Is the adjacency matrix of an undirected graph always symmetric? (True/False)

Answer - True

5) [3 points] Given the graph and its community below, write the maximum likelihood equation of this graph in terms of Pc, Pd and Pe, where Pc is the probability that an edge belongs to community C. If any 2 nodes A,B do not share any vertices, Assume, Pab = (1 - E)

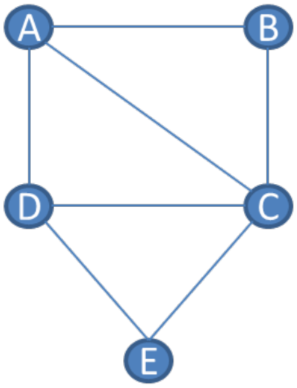
Community C - green color
Community D - red color
Community E - purple color



Answer-
Maximize - Pwx.Pxy.Pvy.Pyz.Pwy.(1-Pwz).(1-Pvw).(1-Pxz).(1-Pxv).(1-Pvz)
I.e., Maximize Pc.Pc.Pe.Pd.(1-((1-Pc).(1-Pd).(1-Pe))).(1-Pd).(1-Pe).(1-E).(1-E).(1-E)

6) [2 points] For nodes A,B,C, use the Girvan-Newman algorithm to calculate the betweenness of each edge (do this for C ONLY). Write down the edges and their betweenness values in the format below:

(Edge1, Edge2) = Betweenness Value
Start Node - C



(C, A,C) = 1
(C, B,C) = 1
(C, C,D) = 1
(C, C,E) = 1

Offline (Take-home) Quiz –

1) [2+1 points] Consider we have 2 clusters. Cluster A consists of the points P1(1,2,1), P2(2,1,3) and P3(0,3,2) and Cluster B consists of points Q1(4,5,7), Q2(6,6,6) and Q3(5,7,4). How will both of these clusters be represented if we are to apply the BFR algorithm?
   Now, consider the following points. We have to apply the BFR algorithm to determine what should be the assignment for each of these points. Computations of each point should be done independent of each other. We add a point to a cluster with a threshold of 3 standard deviations from the centroid. For this problem assume standard deviation is root(3) or 1.732

   1. Point X1(3,5,6)
   2. Point X2(3,5,4)

   Cluster A:
   N = 3
   SUM = (3, 6, 6)
   SUMSQ = (5, 14, 14)

   Cluster B:
   N = 3
   SUM = (15, 18, 17)
   SUMSQ = (77, 110, 101)

   Point X1:
   Distance from A: 6.57
   Distance from B: 2.859
   Hence X1 will be assigned to cluster B

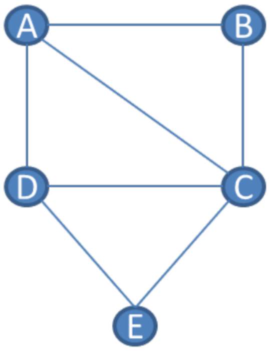
   Point X2:
   Distance from A: 5.0371
   Distance from B: 3.0394
   Hence X2 will be assigned to cluster B
   A is also within range, but B has a smaller distance.

   Some common mistakes from students:
   Not describing the clusters in the way mentioned above.
   Using variance instead of standard deviation in the formula of Mahalonobis Distance. This has led to larger distances and hence wrong answers.

2) [4 points] For nodes A,B,C, use the Girvan-Newman algorithm to calculate the betweenness of each edge. Write down the edges and their betweenness values in the format below:

   (StartNode, Edge1, Edge2) = Betweenness Value

You are expected to give 3 sets of values. For each set, assume only 1 start_node(first A, then B, then C) and write the betweenness values corresponding to this node as the root.



(A, A,B) =1
(A, A,C) = 1.5
(A, A,D) = 1.5
(A, C,E) = 0.5
(A, D,E) = 0.5

(B, A,B) = 1.5
(B, B,C) = 2.5
(B, A,D) = 0.5
(B, C,D) = 0.5
(B, C,E) = 1

(C, A,C) = 1
(C, B,C) = 1
(C, C,D) = 1
(C, C,E) = 1

3) [1 points] What do you mean by modularity when talking about network communities. What is its use?

Modularity is measure of how well a network is partitioned into communities. It takes the value between the range [-1,1]. It is positive when the number of edges within groups exceeds the expected number. It is useful for selecting the number of clusters

4) [0.5 point] In Affiliation Graph Model (AGM), if two nodes u,v share no community then the edge probability P(u,v) is very high.

Answer - False

5) [1.5 point] We are given a set of coin flips : $X = [0,\ 0,\ 1,\ 0,\ 1,\ 0,\ 0,\ 1]$. Figure out the bias of a coin when the Model is $f(\Theta)$ *and it* returns 1 with prob. Θ, else returns 0. We assume the coin flips are independent. What is $P_f(\Theta)$?

Pf(X|θ) = (1-θ)(1-θ)*θ*(1-θ)*θ*(1-θ)*(1-θ)*θ
= θ^3 * (1-θ)^5