HW3

Describe your method and write a script for each of the following problems.

1. Consider the 40x40 gene expression matrix A from “matrix_A.txt” file.
   (A) Write a script to cluster the genes (rows) of matrix A based on hierarchical clustering. Use the average Euclidean distances between the genes in two clusters as the distance metric between two clusters. Plot the “maximum distance between gene pairs in the same cluster” (a measure of homogeneity) and “minimum distance between gene pairs in different clusters” (a measure of separation) as a function of *k* , the number of clusters. What is the best value for k in terms of having both separation and homogeneity ? . Report the clusters in case of k=30.
   (B) Perform biclustering on matrix A using the SAMBA method and parameters $p_c=0.9$ and $p_{u,v}=0.1$. What is the maximum degree of the bipartite graph? Use the maximum degree as the bound on degree. Compare your results to the clustering from part (A) using distance between clusters/biclusters defined below*.

2. Repeat part (A) using data from “noisy_matrix_A.txt”. This is a version of matrix A with noise added. Compare the clustering of noisy matrix to the clustering of matrix A. Compare the bi-clustering of noisy matrix A to the bi-clustering of A.

   Which method is more noise-robust, clustering or bi-clustering ? Which method did you expect to be more robust ? Why ?

***Distance between clusters/biclusters.** Measure the difference between clustering / biclustering using the following metric. For Clusterings / Biclusterings C_1 and C_2, the distance between C_1 and C_2 is defined as the number of gene pairs that are present in C_1 and absent in C_2, plus the number of gene pairs that are present in C_2 and absent in C_1.