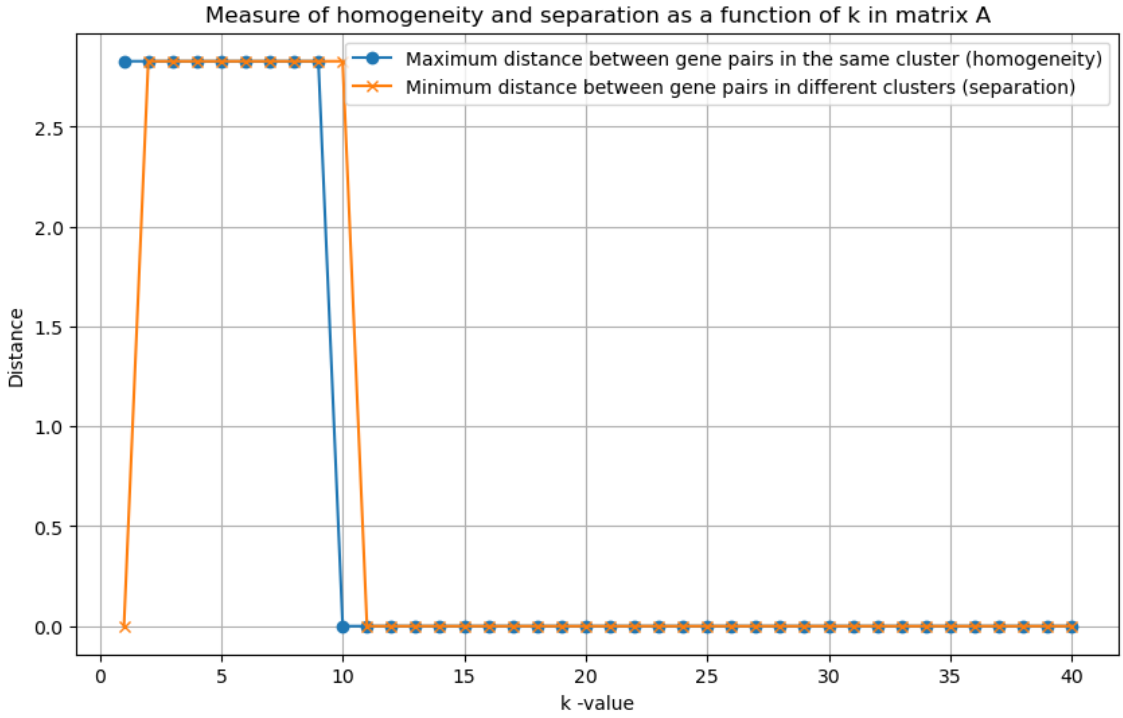
02-725 HW3

Anushka Sinha

1.1 HIERARCHICAL CLUSTERING



1. This plot shows 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 in the matrix A.

2. The best value for k is 10 because it produces the highest separation and smallest homogeneity as shown by the plot.

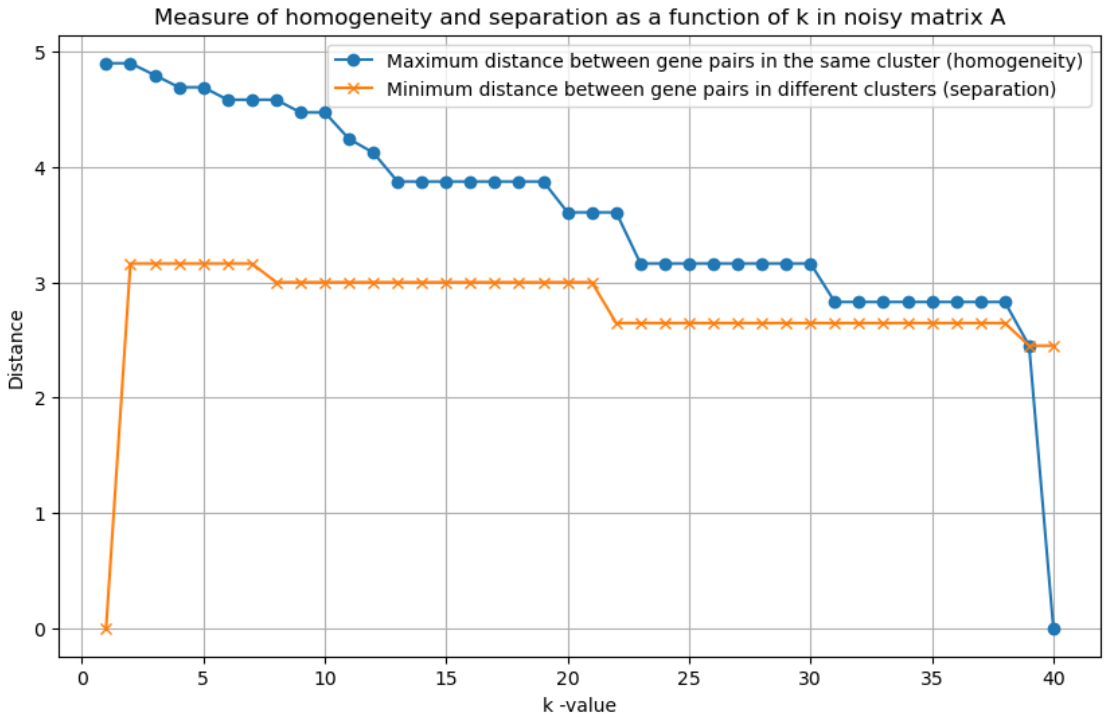
3. The clusters for k=30 are:
$[[5,13],[4,11,17,32],[1,2,3,7],[0,9,33,36],[6],[8],[10],[12],[14],[15],[16],[18],[19],[20],[21],[22]$
$[23],[24],[25],[26],[27],[28],[29],[30],[31],[34],[35],[37],[38],[39]]$

1.2 BICLUSTERING WITH SAMBA

1. The maximum degree of the bipartite graph is 4.

2. Comparing the Hierarchical clustering (k = 10) results on matrix A with Biclustering with SAMBA on matrix A, the jaccard distance obtained is 0 and jaccard similarity is 1. In the calculation for jaccard distance and similarity, intersection is the number of times every gene pair is present in the same cluster for both hierarchical clustering and biclustering. The union is the number of times every gene pair is present in the same cluster for either hierarchical and biclustering algorithms. A jaccard distance of 0 indicates that every pair of genes that is clustered together in hierarchical clustering is also clustered together in SAMBA biclustering, and vice versa.

2. NOISY MATRIX



1. Based on the graph, the best value of k is not clear for noisy matrix A as there is no value of k that maximizes separation and minimizes homogeneity. In this case, k = 30 might be the best number of clusters as around this area on the graph, the separation remains consistent but homogeneity is very low. When compared to the hierarchical clustering on matrix A, it seems like hierarchical clustering on noisy matrix A produces some unnecessary clusters which are not present when the non-noisy data is used.
   The homogeneity measure in the non-noisy data graph is continuously low for all values of k, indicating that either the clusters are extremely tight or the gene pairs within them are very similar to one another. Because the noise is probably affecting the biclustering process, the homogeneity measure in the noisy data graph starts higher and fluctuates as k increases, suggesting that there is more variability within clusters and the the clusters are looser.
   The separation measure is constantly high for all values of k in the non-noisy data, indicating a distinct separation between the clusters. The separation fluctuates more as k increases for the noisy data, indicating that the clusters may not be as well-defined and that there may be less distinction between them.

2. Biclustering using SAMBA on noisy data produces more unique biclusters, as well as biclusters of varying sizes, compared to non-noisy data. For example there are clusters with two and three genes. However, biclusters of non-noisy data show clusters of mostly size 4, if we do not consider the clusters formed by just one gene. The larger biclusters of noisy data are less frequent, and when they do appear, they tend to be smaller than those in the non-noisy data.This indicates that biclustering of noisy data is giving rise to irrelevant groupings that were not formed when biclustering was done on non-noisy data. It also shows that noise is obscuring true relationships, making it harder for the algorithm to find larger and meaningful clusters.