02-425/725 HW3

February 28, 2024

1 Clustering

Consider the 40x40 gene expression `matrix A` from `matrix_A.txt` file.

1.1 Hierarchical Clustering

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.

1. 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.
2. What is the best value for k in terms of having both separation and homogeneity?
3. Report the clusters in case of $k = 30$.

1.2 Biclustering with SAMBA

Perform biclustering on `matrix A` using the `SAMBA` method and parameters $p_c = 0.9$ and $p_{u,v} = 0.1$.

1. What is the maximum degree of the bipartite graph? Use the maximum degree as the bound on degree.
2. Compare your results to the clustering from 1.1 using Jaccard distance.

2 Noisy Clustering

Repeat part 1 using data from `noisy_matrix_A.txt`. This is a version of `matrix A` with noise added.

1. Compare the clustering of `noisy matrix A` to the clustering of `matrix A`.
2. Compare the bi-clustering of `noisy matrix A` to the bi-clustering of `matrix A`.