Homework #2 Due on 10/15/2021

Instructions: While discussion with classmates are allowed and encouraged, please try to work on the homework independently and direct your questions to me.

Instructions: Please interpret your analysis results using concise and clear language and focusing on interesting findings. Remember to include your Python codes and only necessary Python outputs.

1. **Principal Component Analysis (PCA)**: Unsupervised techniques are often used in the analysis of genomic data. In this work, we will illustrate PCA on the NCI60 cancer cell line microarray data. The data contains expression levels on 6830 genes from 64 cancer cell lines. Cancer type is also recorded.

   The file `NCI60_data.csv` is a 64 by 6830 matrix of the expression values while the file `NCI60_labs.csv` is a vector listing the cancer types for the 64 cell lines.

   (a) Given the provided datasets (as CSV files), load them in Python.

   (b) Data preprocessing: Check to see if the data is standardized. If not, standardize the data matrix X so that all variables are given a mean of zero and a standard deviation of one.

   (c) PCA:

      i. Fit the PCA model and transform to get the principal components

      ii. Plot the first few principal component score vectors (i.e. $Z1$ vs $Z2$ and $Z1$ vs $Z3$), in order to visualize the data. Comment on your visualization. **Note**: The observations corresponding to a given cancer type should be plotted in the same color, so that we can see to what extent the observations within a cancer type are similar to each other.

      iii. Plot the Proportion of Variance Explained (PVE) of each principal component (i.e. a scree plot) and the cumulative PVE of each principal component.

2. **(Page Rank)**: Based on the links in Figure 1, obtain the link matrix L and then accordingly compute the PageRank score for each webpage. Provide a barplot of the PageRank score. Which pages come to the top-four list?

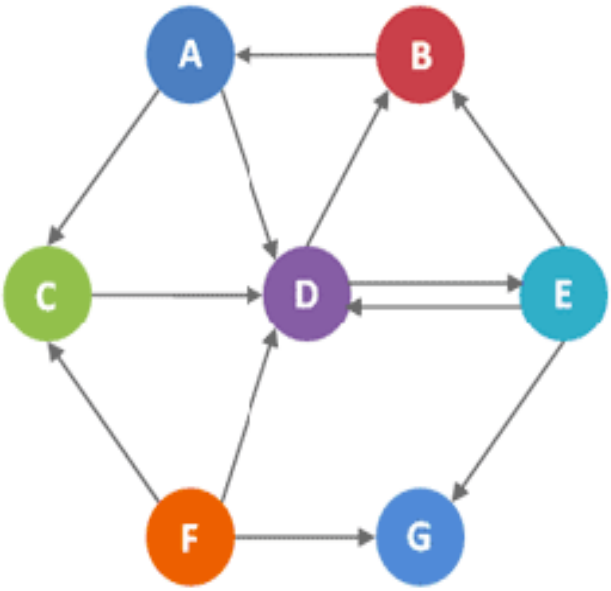

Figure 1: Links among Several Webpages.