# ISYE 6740 Homework 2

## 1 Spectral clustering [50 points]

1. (20 points) Consider an undirected graph with non-negative edge weights $w_{ij}$ and graph Laplacian $L$. Suppose there are $m$ connected components $A_1, A_2, \ldots, A_m$ in the graph. Show that there are $m$ eigenvectors of $L$ corresponding to eigenvalue zero, and the indicator vectors of these components $I_{A_1}, \ldots, I_{A_m}$ span the zero eigenspace.

2. (30 points) Real data: political blogs dataset. We will study a political blogs dataset first compiled for the paper Lada A. Adamic and Natalie Glance, "The political blogosphere and the 2004 US Election", in Proceedings of the WWW-2005 Workshop on the Weblogging Ecosystem (2005). The dataset nodes.txt contains a graph with $n = 1490$ vertices ("nodes") corresponding to political blogs. Each vertex has a 0-1 label (in the 3rd column) corresponding to the political orientation of that blog. We will consider this as the true label and try to reconstruct the true label from the graph using the spectral clustering on the graph. The dataset edges.txt contains edges between the vertices.

   Here we assume the number of clusters to be estimated is $k = 2$. Using spectral clustering to find the 2 clusters. Compare the clustering results with the true labels. What is the false classification rate (the percentage of nodes that are classified incorrectly).

## 2 PCA: Food consumption in European area [50 points]

The data food-consumption.csv contains 16 countries in the European area and their consumption for 20 food items, such as tea, jam, coffee, yoghurt, and others. There are some missing data entries: you may remove the rows "Sweden", "Finland", and "Spain". The goal is to perform PCA analysis on the data, i.e., find a way to perform linear combinations of features across all 20 food-item consumptions, for each country. If we extract two principal components, that means we use two singular vectors that correspond to the largest singular values of the data matrix, in combining features. You will need to implement PCA by writing your own code.

1. (15 points) Write down the set-up of PCA for this setting. Explain how the data matrix is set-up in this case (e.g., each dimension of the matrix correspond to what.) Explain in words how PCA is performed in this setting.

2. (15 points) Suppose we aim to find top $k$ principal components. Write down the mathematical optimization problem involved for solving this problem. Explain the procedure to find the top $k$ principal components in performing PCA.

3. (10 points) Find the top two principal component vectors for the dataset and plot them (plot a value of the vector as a one-dimensional function). Describe do you see any pattern.

4. (10 points) Now project each data point using the top two principal component vectors (thus now each data point will be represented using a two-dimensional vector). Draw a scatter plot of two-dimensional reduced representation for each country. What pattern can you observe?