- This homework is due at 11:59pm on April 23, 2021. Please submit by email to `natalies@cs.unc.edu+comp790`.
- There are a few files provided:
  - Microbiome Network 1: with edgelist given in `supragingival_plaque.edgelist` and node names given in `supragingical_plaque_nodenames.csv`
  - Microbiome Network 2: with edgelist given in `subgingival_plaque_nodenames.csv` and node names given in `subgingival_plaque_nodenames.csv`
- You are welcome to consult with other colleagues, but please write up your own independent solution.
- You are welcome to use Python, Julia, or R here.
- You are welcome to write up your assignment using the `HW2_790-166.tex` template, or write up the solutions in the method of your choice.
- This homework is worth 50 points total.
- Please submit your final writeup as a PDF.

Problem 1

(50 Points Total) **Microbiome Network Alignment**

SparCC `https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1002687` is a method to infer correlation networks between microbial species. In these networks, each node represents a particular microbial species and an edge exists between a pair of nodes if those species have prominent co-occurrence patterns across human subjects. Using data provided with the SparCC paper, we have downloaded microbial interaction networks constructed from data collected in the human microbiome project `https://hmpdacc.org/`. Briefly, this effort focused on characterizing microbial co-occurence patterns in different body sites.

Here, you have been given two networks of microbial interactions observed in 1) subgingival plaque and 2) supragingival plaque. (**Proceed with caution if you google these**). We will use our recent knowledge in graph alignment to compare these two networks and to understand the inferred mapping between their nodes. For each network, we have provided an edgelist and `.csv` file providing the biological name for each node:

- Supragingival Plaque Network (1): with edgelist given in `supragingival_plaque.edgelist` and node names given in `supragingical_plaque_nodenames.csv` $\rightarrow$ referred to downstream as 'Network 1'
- Subgingival Plaque Network (2): with edgelist given in `subgingival_plaque_nodenames.csv` and node names given in `subgingival_plaque_nodenames.csv` $\rightarrow$ referred to downstream as 'Network 2'

Recall REGAL alignment `https://arxiv.org/pdf/1802.06257.pdf`. The following homework sub-propblems will **walk us through implementing the REGAL graph alignment approach**.

1) **Constructing Node Features (5 points):** The first part of REGAL is to create a feature vector for each node that helps to summarize something about its context. We will use a simple k-hop method to construct a feature vector for each node. Recall that for a node, i, its 'k-hop subgraph' can be obtained by considering nodes that are within k hops from i. (Hint: you may find the following useful `https://networkx.org/`

documentation/stable//reference/generated/networkx.generators.ego.ego_graph.html).

We will consider k-hop networks for $k = 1, 2, 3, 4$. **Write a function, where for a particular k, you collect the set of neighboring nodes within k hops of each node and summarize the degree distribution of these collective 'k-hop neighbors' with 4 statistics : {min degree, median degree, mean degree, max degree}.** After doing this for each value of k, you should ultimately be able to represent each node with 16 features (4 considered hops $\times$ 4 summary statistics per hop). As an example, assuming Graph 1 has N_1 nodes, define its node feature matrix, $\mathbf{X}_1 \in \mathbb{R}^{N_1 \times 16}$ matrix.

2) **Intuition Building (5 points):** Use your new function to build the described feature vectors for Supragingival Plaque Network (Network 1). Assuming this network has N_1 nodes, **project these N_1 nodes into two dimensions using your dimensionality reduction method of choice**, based on the 16 computed features ($\mathbf{X}_1 \in \mathbb{R}^{N_1 \times 16}$).

3) **Choosing Landmarks (5 points):** Recall that REGAL constructs an embedding for each node by specifying landmark nodes that have been collected across both of the graphs being aligned. Choose a set of d landmark nodes **collectively** across Graphs 1 and 2. You can play with d later, but considering the total number of nodes is < 200 between graphs 1 and 2, perhaps $d = 30$ is a good place to start. You can choose the set of d landmarks at random, or use a more sophisticated approach. **Explain your choice of landmarks and write a function to return these landmark nodes**.

4) **Computing Similarities to Landmarks (5 points):** In part 1), you wrote a function to compute feature vectors for each node. Assuming Graph 1 has N_1 nodes and Graph 2 has N_2 nodes, **write a function that computes a similarity measure in this 16-dimensional space between each of the nodes in Graph 1 and Graph 2 to each of the d landmarks.** So, you should en up with a matrix, $\mathbf{C} \in \mathbb{R}^{(N_1+N_2)\times d}$.

5) **Extract Landmark $\times$ Landmark Matrix (5 points):** As you know, the $\mathbf{C}$ that you constructed contains the d landmark nodes! Write a function to construct $\mathbf{W} \in \mathbb{R}^{d \times d}$ submatrix of $\mathbf{C}$ where the similarities between the landmarks were stored.

6) **Embedding via Landmarks (5 points)**: Given Theorem 3.1 in https://arxiv.org/pdf/1802.06257.pdf, we can compute the collective node embedding matrix (across Network 1 and Network 2), $\tilde{\mathbf{Y}} \in \mathbb{R}^{(N_1+N_2)\times d}$, as

$$\tilde{\mathbf{Y}} = \mathbf{C}\mathbf{U}\mathbf{\Sigma}^{1/2}$$

.
Recall that here, $\mathbf{U}$ and $\mathbf{\Sigma}$ are obtained through an SVD on the pseudo inverse ($\mathbf{W}^{\text{pinv}}$) of the (landmark $\times$ landmark) similarity matrix, $\mathbf{W} \in \mathbb{R}^{d \times d}$ extracted from $\mathbf{C}$.

$$\mathbf{W}^{\text{pinv}} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^T$$

Hints: These are useful for pseudoinverse (https://numpy.org/doc/stable/reference/generated/numpy.linalg.pinv.html) and SVD (https://numpy.org/doc/stable/reference/generated/numpy.linalg.svd.html).

Given this information, write a function to compute $\tilde{\mathbf{Y}}$.

7) **Putting it All Together Visualization 1 (5 points):** You have now defined an embedding for all nodes in Networks 1 and 2 in some d-dimensional space through $\tilde{\mathbf{Y}}$. **Use your favorite dimensionality**

reduction method of choice to project the collective set of nodes in Networks 1 and 2 into two dimensions. Color the nodes by which network they are from. Comment on any observations.

8) **Alignment Between Graphs (5 points):** Given $\tilde{\mathbf{Y}}$, **calculate a similarity score (your choice) between each node in Network 1 and every node in Network 2.** Remember that you have the names of these nodes in `supragingical_plaque_nodenames.csv` and `subgingival_plaque_nodenames.csv`. **Comment on your mapping in regards to whether or not nodes with the same name are being mapped to each other between Network 1 and Network 2.**

9) **Creativity (5 points):** Now that you have the entire pipeline in place, play around with it a bit. For example, considering changing how you define the features for nodes in part 1), changing the value of d, changing how you choose landmarks, or anything else that is interesting to you! **Re-run steps 1-7 with your modification and comment on how it changes the interpretation of alignment between Network 1 and Network 2 given in $\tilde{\mathbf{Y}}$.**

10) **Creativity Part 2 (5 points):** Imagine a collaborator dropped these two networks on your desk. They are paying you from their grant, so you need to produce something to give them. **Create a visualization of your choice that reflects something about the similarity between Network 1 and Network 2** (in terms of node alignment, clustering structure, etc).

Congratulations! You implemented REGAL from scratch!