**INFS 770 Advanced Data Mining Applications**

**Assignment 5**

**Due 5/7/2021 with no Grace period**

In this assignment, you will need to conduct some network analysis on an Enron email data set. The text file “Email-Enron.txt” included in the assignment 5 folder contains the edge list of the communications. The description of the data set can be found at <http://snap.stanford.edu/data/email-Enron.html>.

**Things to submit:** Please develop and submit an iPython notebook titled 770\_21\_a5\_*yourlastname*. In the notebook, you need to finish the tasks below, where (C) indicates that you need to write code for the task, (O) indicates that you need to show output, and (A) that you need to type your answers using Markdown text. Please use “run all” to run your code before you submit so that your iPython notebook will show the outputs of your code. You will lose 1 point if you do not “run all”. You probably want to copy and modify the code from the iPython notebook “socialnet\_example” posted on D2L.

In your iPython notebook, at the beginning of each cell, you need to indicate which task the cell is about. If you do not clearly label the cells, you will lose 1-2 points (out of 18 points). Whenever you see “print” in the questions, you need write print statement to print the intended outputs.

To read the data into a networkx graph, use the following codes:

**import networkx as nx**

**# specify the file location**

**file\_name = "<the path to the data file>"**

**# read an edge list into a networkx graph**

**g = nx.read\_edgelist(file\_name, # file location**

**comments="#", # the character used to indicate the start of a comment**

**delimiter="\t", # the character that separates values in a line**

**create\_using=nx.Graph() # create as undirected graph**

**)**

Q1. Please write code to print the number of nodes and the number of edges in the entire network (use the method print nx.info(g)). Please write code to print the density of the network. (C)(O)

Q2. Please create a textbox and tell me: 1) What do the nodes and edges mean in this network? 2) Are the edges weighted? If so, what does the weight mean? 3) What does density of network mean? (you don’t need to write code for these questions) (A)

Q3. Please write code to plot the degree histogram. (C)(O) Please create a textbox in which you try to interpret the histogram (hint: you must talk about the distribution shown in the histogram).(A)

Q4. Please write code to print the number of connected components in the network? Please write code to print the number of nodes and the number of edges in the largest component? (C)(O)

Q5. Write code to name the component having 16 nodes as comp\_A (here you need to write a for loop to traverse the list of components, find the component with len(…) ==16, and set comp\_A =g.subgraph(…) ) and the component having 14 nodes as comp\_B. Please write code to visualize comp\_A and write code to visualize comp\_B. (C)(O)

Q6. Please create a textbox to summarize the difference between comp\_A and compB. You need to at least compare these two subgraphs with respect to *average degree* and *density.* You don’t need to write code for this task. (A)

Q7. Please write code to use the Louvain modularity method to detect communities within Component A with 16 nodes. Please visualize the different communities using different colors as I did in socialnet\_example.ipynb. Please write code to print the number of communities you discovered. (C)(O)

Q8. Please write code to print the number of cliques in comp\_A. Please write code to show the nodes in the largest clique.(C)(O)

Q9. Please create a textbox. Please tell me in the textbox which two nodes (people) in comp\_A are likely to have similar opinions and why, and which two nodes in comp\_A are likely to have different opinions and why. (you don’t need to write code for this) (A)

Q10. Find important nodes in comp\_A using at least three centrality measures. For each measure, please write code to print a datafame with two columns “node” and centrality (e.g., betweenness\_centrality or closeness\_centrality) and the dataframe needs to be sorted by the centrality measure in a descending order. **Pease note you just need to run the centrality computations for comp\_A** rather than the whole network. It takes very long time to obtain the centrality measures for the whole network. (C)(O)

Q11. Please create a textbox to summarize the results you obtained from the different centrality measures (you need to tell me if these different measures give similar results in identifying the most important people in the network; if not, why the results are different (here you cannot just say the results are different because I used different measures. You need to compare different measures and discuss their implications)). (A)