PREDICT 420. Individual Assignment 5: Preparing Data for Analysis

Memo from January 2002. We appreciate you early investigations into selected activities at Enron. Now the Federal Energy Regulatory Commission (FERC) and the Securities and Exchange Commission (SEC) request further services as we try to understand who was involved with these activities prior to Enron's filing for Chapter 11 bankruptcy protection on December 2, 2001. Having identified an activity of interest, your next job is to learn about the people involved. What would serve our purposes well would be a picture of the network of Enron executives (perhaps with links to persons outside of Enron) involved with the selected activity. We understand that relevant data for visualizing this network may be obtained from the Enron e-mail archive. For your work, please use the Enron e-mail archive on the PostgreSQL database server. You will need to use your NetID and password to connect up to the SSCC dornick host via a VPN connection. Use the **`psql`** database shell to PostgreSQL to find your way to Enron e-mail messages of interest.

Enron e-mail messages are in PostgreSQL tables **messages** and **messages_table** under the **enron_work** database. Names of Enron executives are provided within **messages_table** columns **`fromaddress`**, **`toaddress`**, **`ccaddress`**, and **`bccaddress`**. Duplications can be avoided by using **`SELECT DISTINCT`**. To cut down on the file size, use **`fromaddress`** and **`toaddress`** only. In gathering e-mail addresses for the network visualization, you are encouraged to stay within the **`psql`** shell. Create a PostgreSQL view, copy this view to a comma-delimited text file, and download that file to your personal computer for further processing in Python—direct **`sftp`** is easy on Mac OSX, and FileZilla works on both Mac OSX and Windows.

After the file of e-mail addresses is on your personal computer, additional processing will be required to organize these text data for input to programs for network visualization. In particular, you will need to create a file of network edges (from-node and to-node e-mail addresses). The file **`sample_plot_input.txt`** shows the required format for input to the Python NetworkX package, and the Python program **`sample_make_plot.py`** shows how to analyze these data by creating a plot of the social network. (Note that you must install **`networkx`** before running this program.) The resulting visualization is **`sample_plot.pdf`**.

Deliverable: Submit on Canvas a SINGLE PDF file that has YOUR OWN QUERY sample_plot.pdf file.