***CAP 6315: Soc Networks/Big Data Analytics Programming Assignment #1***

Assignment learning objectives:

* Students will be able to use NetworkX library to create and visualize social networks, as well as extract meaningful metrics from it.
* Students will learn to interpret the results of social network analysis by using the network properties such as network diameter, edge density, and clustering coefficient.
* Students will learn to detect influential nodes by using graph centrality.

***Instructions***

* Download the dataset: <https://networkrepository.com/rt-twitter-copen.php>
* The dataset is in .*mtx* format. You can read the data by using this line of code ‘*matrix = mmread('rt-twitter-copen.mtx')*’, while also importing ‘*from scipy.io import mmread’*. Finally, you can create a graph with ‘*G = nx.from\_scipy\_sparse\_array(matrix)*’.
* ***Your submission should include the complete Jupyter Notebook following the steps outlined below***. Be sure to include everything! It’s helpful if you add the steps as comments in your code, followed by the code block or markdown cells that represent the answers.
* ***Leave the code outputs in the notebook***. Codes without visible outputs will lose 0.25 points for each missing output.
* Your code (submission) should be named ***YourLastName\_Assignment1.ipynb*** and it should be submitted to Canvas before the deadline.

***Assignment***

* Using NetworkX library and Pandas/Matplotlib, follow the steps below:
  1. Load the data and build a graph using network library. (0.5 points)
  2. Visualize the graph. (0.5 points)
  3. Calculate the number of nodes in the graph. (0.5 points)
  4. Calculate the number of edges in the graph. (0.5 points)
  5. Calculate the network diameter. (0.5 points)
  6. Add a markdown cell. In your own words, interpret the meaning of obtained value for network diameter. (0.5 points)
  7. Plot a histogram representing degree distribution. (0.5 points)
  8. Add a markdown cell and discuss what you observe by looking at the generated degree distribution. (0.5 points)
  9. Calculate the average degree of a node. (0.5 points)
  10. Add a markdown cell. In your own words, explain what the obtained average degree of a node mean in the context of this dataset? (0.5 points)
  11. Calculate the average shortest path length between any two nodes. (0.5 points)
  12. Add a markdown cell. In your own words, explain what the obtained value of the average shortest path length means in the context of this dataset. (0.5 points)
  13. Plot the distribution of shortest path length. (0.5 points)
  14. Add a markdown cell and explain what you observe in this histogram. (0.5 points)
  15. Calculate the edge density of this graph. (0.5 points)
  16. Add a markdown cell and answer the following question. What can we conclude about this graph considering its edge density? (0.5 points)
  17. Calculate the global clustering coefficient of this graph. (0.5 points) – *make sure to use the correct function (there are multiple ones. Look for the correct definition)!*
  18. Add a markdown cell and answer the following question. What can we conclude about this graph considering its global clustering coefficient? (0.5 points)
  19. Calculate the number of connected components in the graph. (0.5 points)
  20. Report 10 nodes with the highest degree centrality (1 point)
  21. Report 10 nodes with the highest betweenness centrality (1 point)
  22. Report 10 nodes with the highest closeness centrality (1 point)
  23. Report 10 nodes with the highest Eigenvector centrality (1 point)
  24. Add a markdown cell and answer the following question. What do you observe by looking at the top 10 nodes for each centrality? Explain in your own words (e.g., which nodes appear multiple times, which nodes only showed up in top 10 in one of the centralities, etc.) (0.5 points)
  25. For each centrality, plot a histogram of its distribution (a total of 4 histograms for this question). (4x 0.5 points = 2 points)
  26. Under each histogram, add a markdown cell and in your own words, explain what your observation is. (4x 0.5 points = 2 points)
  27. For each centrality, visualize a network where the size of the nodes corresponds to the centrality values (4 plots). (4x 0.5 points = 2 points)