The dataset that our group has selected details all the connections between a large selection of Twitter users and the social circles that are generated from these connections. This dataset was taken from the Stanford Large Data Network Collection. Specifically, the data was compiled by J. McAuley and J. Leskovec for the purposes of their publication *Learning to Discover Social Circles in Ego Networks*. Each node in the dataset represents one user with each edge representing a follow from one user to another. In traversing this dataset, we plan to use a Breadth First Search. Not only will it be helpful to find all the vertices/nodes connected to each other vertex/node, but it may also assist in our Betweenness Centrality algorithm, one that measures the "centrality" of each node based on criteria such as number and length of edges. We hope to display that output of said algorithm graphically, first by running all the nodes and outputted data into a Graph Coloring algorithm to easily identity the centrality of nodes visually and then through a Force Directed Graph algorithm that should graphically display the centrality and density of different node clusters (i.e. social circles). Our end goals that we hope to achieve for this project are to learn how to graphically display complex data, learn to implement a betweenness centrality algorithm, and to implement graphical analysis to real-world applications.

EDIT: As of the third week of development, our group has decided to revise our original goals to those that would be more reasonable in scope for this project. We have decided to switch datasets from the aforementioned Twitter social circles to Wikipedia user votes, with each node representing a user and an edge representing one user voting for another's edit to a Wikipedia page. The main reason we are changing

datasets is due to the runtime of both Floyd-Warshall and Betweenness Centrality. With our original dataset of 800,000 nodes the runtime would've been astronomical; this new dataset only contains around 7000 nodes. Due to the delays that changing datasets would cause in our development cycle, we've also decided to forego implementation of a Force-Directed Graph and a Graph Coloring algorithm. Generating a graphical output in the form of a Force-Directed Graph would've required too much valuable effort better put towards our aforementioned Floyd-Warshall and Betweenness Centrality. Likewise, if we do not plan to produce a graphical output, then Graph Coloring is not nearly as useful nore needed anymore. Our project goals, then, are revised to learn how to implement a Floyd-Warshall shortest path algorithm, a Betweenness Centrality algorithm, and draw conclusions to real world phenomena from our resulting data output.