CIS 310 01

Project 2

Rodney Fulk

For this project I used a data set available from Stanford University as part of their SNAP project. The data I used was taken from an app that allowed them to track a user’s friends listing and track connections between friends and some related meta data. There ended up being a total of 10 different groups of people that were included in this data set. The data set had all personal identification stripped from the data set to only have an anonymized dataset for the purpose of investigating the connections between people. The full data set contained a series of information about the groups. I made use of only a portion of the data set that contained user information such as birthday, education including area of study or schools attended, hometown, employer and similar types of information.

The data set I used had an adjacency matrix but also included labels for the nodes and used a tab as a separator. I used Excel to strip the excess data out and to standardize the format of the matrix. I then entered the files into a programmer’s editor called Geany and converted the tabs to spaces to allow them to be imported into netzer. Once I got them into netzer I inspected the number of lines they had and looked at the graphs. I chose to use the two smallest graphs as they only had about 50-60 nodes in them. The others had hundreds of nodes in them and were difficult to readily see details when displayed.

With the smaller graphs I created 8 communities and then searched for the top 5 influential nodes. I noticed that Netzer does a poor job of saving out attributes for graphs when you save the file but I was still able to reproduce the graphs and the related information.

I also made a summary of the results in an Excel document to inspect for tendencies across the different files. Some of the things that I noticed with the data is the average path length seemed to be random on these files and was both negative and positive. Density was directly related to the ratio of edges to nodes. For these data sets the average degree was either close to 10 or close to 7 regardless of the number of nodes and edges.

I used the two smaller data sets as they allowed reasonable displays. I made a total of 8 communities in each graph and looked for the top 5 influencers. In the one graph it appeared that there were several strong influencing nodes so just choosing 5 was probably not good enough but to get a more accurate look at the data probably would have required more sophisticated software able to process more points. The other graph had far more sparse connections and most of the major influencers were in the two most dense communities which makes sense.

I ended up spending several hours on this project trying to make sense of the data. In the end If I had spent more time processing the data up front and either using a different utility or taken the data and converted it to a more usable format it may have been more telling about connections between users and how they were connected. Some of the data I was not able to use in its current format without major reconstruction.

This project showed me that to get successful results it requires a good amount of work up front to insure you have usable data. This project could be extended further by possibly separating the data into a different configuration of data sets and including more users. There was an all-inclusive dataset available that included all data points but as it was it was difficult to separate out the data points. This could have been expanded out to use more users and bigger data points but would have likely required specialized software to compress the data to a more practical usable format.