**DATA 620 Week 4 Assignment**

**Centrality Data Analysis**

**Erik Nylander**

**Data:**

The data source that I have identified for this week and for the project one is the full game results of 2010 NCAA Football season. The data set contains the date of the game, the first team in the game, the teams score, the second team in the game, their score, and finally the location of the game if it was at a neutral site. The data is in a text file and was created by James Howell which can be found at <http://homepages.cae.wisc.edu/~dwilson/rsfc/history/howell/cf2010gms.txt>. This data is not in an easy to use format since the spacing is neither coma separated or tab separated. I have done some basic cleaning of the file and will be hosting the cleaned up version of the file on my GitHub page with attribution given to Mr. Howell’s site and work.

Each of the nodes in the data set will be a team with the edges indicating that the teams played each other. I will also be computing the Bowl Eligibility of each of the teams and classifying the teams if they are eligible to play in a bowl game or not. Eligibility is defined as a team having a winning record, a minimum of .500 winning percentage, which is typically six wins in a given season.

**Plan for Analysis:**

To perform the analysis on the centrality of the data I will doing the following process.

Step 1: Download and clean the data. Unfortunately, the data is in an odd text format where it was formatted for ease of human readability and not machine readability. The data has already been downloaded and cleaned up and I will host a copy of the machine readable version of the text file on my own GitHub account. Attribution to the original authors of the data is given in the first two lines of the text file.

Step 2: With the data in a machine readable format, the next step is to load in the data and generate two different object. The first is a dictionary containing all of the teams in the list and their wins and losses. I will trim off any games played after December 11th, the Army-Navy Game, since these are the bowl games and don’t count toward bowl eligibility. Any team that has less than 10 total games played is likely to be a FCS (old division 1AA) team and would not be eligible for a bowl game. I will then compute bowl eligibility for all of the teams in the data set based on winning percentages.

Step 3: Constructing the graph. Now that the data and bowl eligibility have been calculated I will construct the graph. Each team will be a node and the edges will indicate that a game was played between the two teams. For this construction of the graph an edge will be created if the two team played and each node will have the extra categorical information on bowl eligibility.

Step 4: I’m actually unsure what will found when we look at the centrality of the teams that are bowl eligible vs the teams that are not bowl eligible. There should distinct clusters that represent each of the conferences and sub conferences that the teams play in. There should also be teams that are connectors between a conference and other conferences. Given the way that major colleges attempt to schedule “beatable” teams, Louisiana Tech comes to mind, for their out of conference games I could see that the teams with a high degree of eigenvector centrality tend to lose more games and tend not to be eligible to play in bowl games where teams that play mostly in conference and against beatable teams will have a higher degree centrality and may be more likely to be bowl eligible. This of course is all conjecture and may not be a factor at all.