**DATA 620 Week 4 Assignment**

**Centrality Data Analysis**

**Krishna Rajan**

**Data:**

The data source that I have identified for this week and for the project one is the full game results of 2019 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 which can be found at

https://www.sports-reference.com/cfb/years/2019-schedule.html#site\_menu\_link. This data is not in an easy to use format & will be cleaned up and hosted to the github site

Each of the nodes in the data set will be a team with the edges indicating that the teams played each other. Task is to compute 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. The Data from the website is not readily usable , I will be cleaning it and have it ready in txt format ready for the model

Step 2: Once the Data is formatted , 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. The cutoff for this will be the December 14th 2019 Army Navy game , since these are the bowl games and don’t count toward bowl eligibility.

Step 3: Constructing the graph. After the data is computed and bowl eligibility unearthed , time to generate graphs. 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: It will be curious to see 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. While telling us that There should be teams that are connectors between a conference and other conferences. Given the way that major colleges attempt to schedule “beatable” teams, for their out of conference games It can be revealing to 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.