2015 College Football Network

1. Introduction

College football has always had a rich background of unique culture, passion and commitment from fans and players. This same interest can be applied to other areas of the game off the gridiron. As Baseball has evolved in the 21st century with the use of analytics, the same can be applied to college football. This paper will identify the social network of games played during the 2015 college football season. Characteristics of the network will be identified as well as general conclusions.

1. The Data Set

The data set of scores for the 2015 college football season were extracted from Sports Reference, the website that strives to “provide both basic and sabermetric statistics and resources for sports fans everywhere” (Sports Reference). This site was chosen based on personal opinion regarding the ease of use and access to large amounts of useful data.

The data was accessed from the following link: <http://www.sports-reference.com/cfb/years/2015-schedule.html>. This link provided contains all games and scores for each college football game played in the 2015 season. The data is arranged in a clean table making it easy to extract necessary data.

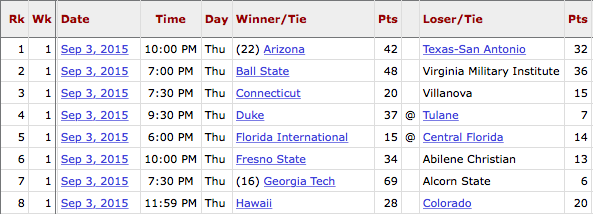


Figure 1. Data Table from Sports Reference

1. Data Extraction

Extraction of data was done by a java program. This program parsed the source code of the sports reference site and extracted necessary attributes of the table. These attributes include, but is not limited to, the winning team, the losing team, the winning score and the losing score.

The program operates from the command line. To run the program, compile the ExtractData.java file followed by the java command to run the main method of the program. Pass a date argument to the command such as: java ExtractData 2015. This example will extract the data from Sports Reference for the 2015 season. Any date between 1869 and 2016 can be successfully entered.

After parsing the data and extracting the necessary information, the program writes a csv file with the required data. For full source code, see the ExtractData.pdf file attached with this document.

1. Network Analysis Tool

Analysis of the network was done with Gephi. Gephi is an open source tool that is ideal for network visualizations and calculations. Many aspects of analysis can be easily done such as calculating centralities and coloring the network as needed.

1. Network Characterization

This section contains characteristics of the network. This information was calculated using Microsoft Excel and Gephi. It is important to note that the calculations on the network were done as a directed graph.

The network of College Football scores for the 2015 season is comprised of 209 nodes, the teams, and 869 edges. An edge exists from A to B if A beat B at some point in the season. Therefor, this is a directed network. The network also consists of 82 strongly connected components and 1 weakly connected component. The diameter of the network is 13.

The five nodes with the highest clustering coefficient are California Davis, Tennessee Martin, Murray State, Idaho State and Texas with a coefficient of 0.5, 0.5, 0.5, 0.5 and 0.28 respectively. Georgia State, Toledo, Georgia Southern, Boise State and Northern Illinois have the highest betweenness centralities. Their scores are 3351, 3141, 2643, 2559 and 2491 respectively.

1. Visualization

This section documents the visualizations of the network. Gephi is used to alter the color and arrangement of nodes to look at different aspects of the network. The size of the nodes represents the number of wins that team had in the 2015 college football season.

Figure 2 represents the generalized network. Labels are excluded from the for this visualization to represent the structure of the network.

noverlap_2.pdf

Figure . Generalized Network.

Figure 3 shows the network using the Fruchterman Reingold layout. This layout assists in visualizing all nodes in the network.Fruchterman%20Reingold.pdf

Figure . Fruchterman Reingold layout.

Figure 4 depicts the network when highlighting over Clemson. All games teams that Clemson played, win or loss, are highlighted along with Clemson. This example is very practical for fans to look at their strength of schedule.

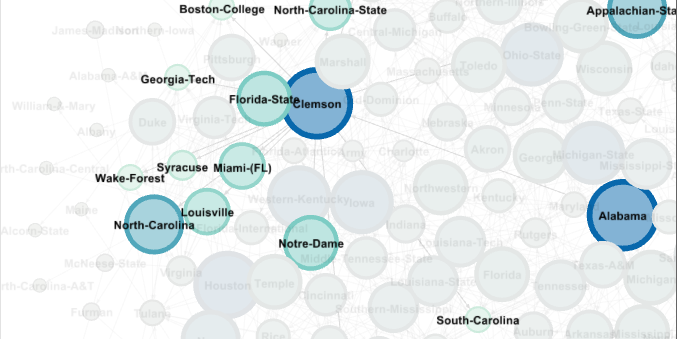


Figure . Clemson Highlighted.

Figure 5 shows only the teams from the power 5 conferences. Red represents the Pac 12, Yellow represents the Big 12, Orange represents the Big 10, Blue represents the ACC and Green represents the SEC.

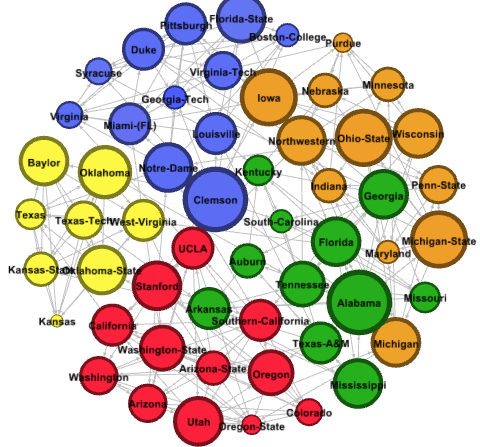


Figure . Power 5 conferences.

1. Analysis

This section will analyze the network from multiple scenarios. Some topics discussed previously will be analyzed and concluded upon.

It can be noticed that the five nodes with the highest clustering coefficient are some of the worst teams in the country. This can be explained by the number of out and in nodes. For the five teams, there is only a small sample of games due to selection of data. The data was limited to division one schools. Therefor, a team such as California Davis would only be in the data set if they played a division one school in the preseason, which would more than likely result in a loss.

The betweenness centrality consists of the five nodes that dominate their respective division one, non power five, conferences. By dominating their conference, other nodes will have to go through these nodes to get to the other nodes on the conference. Also, by winning their conference they better their chance of playing a power five opponent. Regardless of the outcome of this game, the team will be closely connected to a power five school. This in turn would increase the betweenness centrality of the team.

Figure 5 documents only the power five conferences. This graph gives some insight to different issues. First, the PAC 12 missed out on having a team in the college football playoffs. If you look closely at this graph, you will find that the PAC 12 does not have any node that stands out in their conference like the other conferences. In the SEC, we see Alabama as a clear front runner. In the ACC, we see Clemson, and in the BIG 10 we see multiple schools with high degree. This could explain why their were no PAC 12 schools in the playoff.

It could be argued that these schools are quality schools that beat up on each other. This can be argued because the degree of the nodes in the conference appear to be spread out. In other conferences like the ACC, you have some schools with very small degree so schools get could wins against mediocre wins.

Figure 6 7 and 8 show a graph of the degree distribution of the network. Figure 7 shows the in-degree and Figure 8 shows the out-degree.

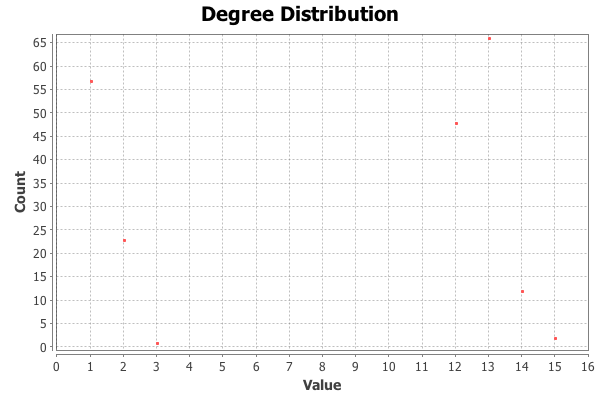


Figure . Degree Distribution.

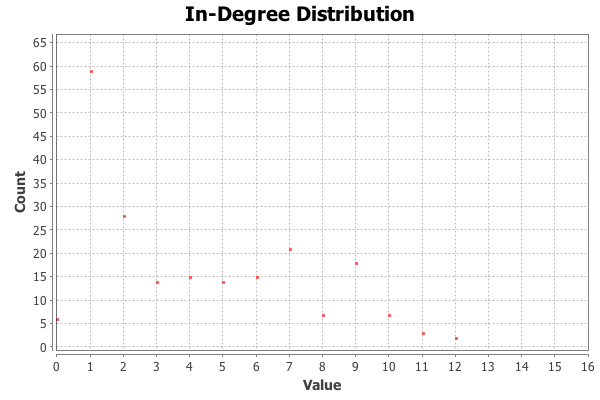


Figure . In-Degree Distribution.

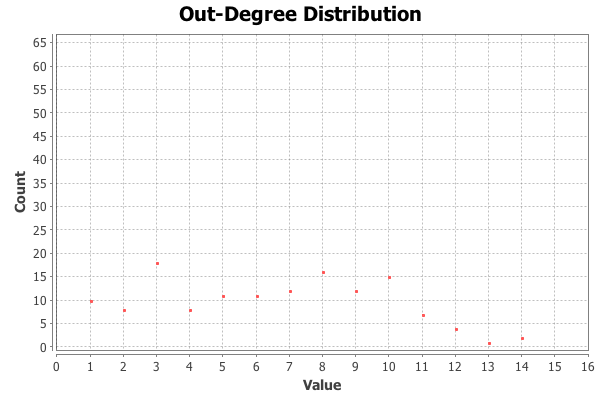


Figure . Out-Degree Distribution.

1. Conclusion

In conclusion, the network of games in the 2015 college football season gives insight about the season that cannot be looked at by just watching the games. The network can help analyze strength of schedule and strength of conference.

This network can be done week by week for the current college football season. This analysis would assist in picking projected winning teams based on past results. This could be used for sports betting or other applications.

1. Work Cited

"Sports Reference | Sports Stats, Fast, Easy, and Up-to-date | Sports-Reference.com." *Sports-Reference.com*. N.p., n.d. Web. 04 Dec. 2016.