Major Leage Baseball as a Social Network

John Lalor for CSC 495

June 10, 2014

1 Introduction

Every year, over one thousand professional athletes play in the United States professional baseball league, Major League Baseball. These athletes play for one of thirty teams, and can move from team to team via trades or by being released by one team and picked up by another. Because there are several players that move between teams during a season and between seasons, the relationships between players and teams over a period can be modeled as a network. There also exists a large quantity of statistics compiled over the years that can be used to describe the players and teams in this network. This report explores the relationships found between MLB players and teams, and attempts to identify player and team attributes that are significant in the formation of the network in its current state.

2 Description of Data

The first professional baseball team was founded in 1869. Since then, statistics for players and teams have been kept with varying degrees of regularity. Into the 1960s and 1970s, statistics were kept and compiled by sabermetricians, individuals who believed that baseball statistics were underused and could be used to make baseball more objective, as opposed to more traditional scouts and general managers who relied on intangibles to scout players. However, there is not a comprehensive set of social network data for MLB, so one had to be constructed for this report. For the purposes

of analysis and processing considerations, the network of MLB players and teams is limited to the period 2008-2013.

The source of the data used for this report is the Lahman Baseball Database (LBD). The LBD is a collection of baseball statistics for a number of areas, which has been aggregated and refined over a number of years. The LBD is stored as several comma-delimited files, and has statistics going back to the first professional teams, but the dataset is more complete from the 1970s onward. Several of the comma-delimited files in the LBD had statistics that were used in the construction of the network:

- *Batting.csv*: This file contains batting statistics for all MLB players, using a unique playerID to identify each player. While none of the batting statistics were used, this file was the baseline for building out the network edgelist. Each player is included in this file for each team they played with during a particular season, and each stint that they had with a particular team. A stint is defined as a period of time spent on a particular team by a player, during one season. A player can have numerous stints in a single season, due to trades or being dropped by one team and picked up by another. Stints are tracked and measured in the network as edge weights.

- *Master.csv*: This file includes player names and biographical information. This table is used to lookup player names from their playerIDs, and also to include player statistics, such as height, weight, and country of birth, as player node attributes.

- *Teams.csv*: Team names and statistics by season for each MLB team. This dataset includes annual statistics, such as wins, losses, final divisional rank, and annual attendance, among others. For this report, the average of these annual statistics were taken over the period and used as node attributes.

- *Salaries*: Annual player salary data for each player. This dataset was used to calculate an average annual salary attribute for each player in the network.

With this data, I created an edgelist of player-team connections and created an initial network object in R. the stint attribute was added to each edge as an edgeweight. From this initial network,

I ran a bipartite projection on the graph to separate players and teams, and added attributes to each of the resulting projection graphs.[1]

Players:

- Player Name
- Years Active
- Average Salary
- USA or Foreign Born
- All Star Appearances
- Hall of Fame Inductee (Y/N)
- Player Height
- Player Weight

Teams:

- Team Name
- Average Annual Attendance
- Average Win Total
- Average Loss Total
- Average Home Run Total
- World Series Wins
- Division Titles
- Average Division Rank

[1] Refer to the appendix for R code used to create the edgelist and network, and to perform the necessary projections and attribute additions.

3 Analysis

With the network complete, I ran a variety of calculations on the data to try to identify some interesting statistics about the network. While the original player-team network is quite large, some analysis was possible. The majority of the analysis, however, focused on the player-player and team-team projections of the network.

3.1 Full Network

Due to the size of the network and computing limitations, it was not possible to perform many calculations on the full network. However, one interesting value that I found was in computing the diameter of the network. The diameter, or longest geodesic, of a network is a way to idenitfy how long it would take to traverse from one end of a network to the other. In the case of the 2008-2013 MLB network, the diameter is 6. In this particular network, it is possible to reach any node, either player or team, from any other node in no more than 6 steps. This is in line with the idea of six degrees of separation. Since the source data is relatively small and specialized, it is not surprising that the diameter is so small.

In this network, there are 5 players with weighted degree values greater than 14: Octavio Dotel, Edwin Jackson, Chad Gaudin, Casey Kotchman, and Luis Ayala. Of these "most connected" players, four are pitchers and one is a first baseman. These players have each had several stints will multiple teams, and have been in the MLB for multiple years as "journeyman" players.

3.2 Player-Player Projection

One area worth exploring was whether different player attributes could be predicted based on a linear modeling of the player-player projection. Specifically, does a players centrality in a network affect other player attributes? For this analysis, I calculated several betweenness measures, and calculated the correlation between them. The betweenness measures were all highly correlated, so I could only use one at a time in my modeling.[2] I tried to model average player salary and nubmer of All Star selctions based on several variables, with mixed success. The best model found was one

[2]Refer to the appendix for the correlation charts

that models the number of times a player is selected as an All Star as a function of his average annual salary, years active in the MLB, whether he is USA-born or not, and his degree centrality in the network:

Coefficients	Estimate	Std. Error	t value	Pr(t)
(Intercept)	3.605e-01	4.089e-01	0.881	0.3783
Average Salary	9.707e-08	5.570e-09	17.428	2e-16
Hall of Famer	-8.628e-01	3.904e-01	-2.210	0.0274
USA Born	-1.447e-02	4.013e-02	-0.361	0.7184
Years Active	9.352e-02	1.273e-02	7.349	4.58e-13
Betweenness	-4.389e-04	6.938e-05	-6.326	4.00e-10

Average salary, years active, and degree centrality were all significant values in the model. The first two make sense, as a player that is highly paid is more likely to be an All-Star, hence justifying the high salary. Also, a player that has spent many years in the league can gain experience and become an All-Star type player.

What is interesting is the fact that degree centrality can be used to model All Star selections. There are two possible explanations for this. First, an All-Star can be the keystone of a franchise, and teams can bring in many players to build around the All-Star. This way, the All Star will make connections with many other players over his career, without moving to many other teams. Second, an All-Star can be in high demand from many teams, and can demand a high salary as part of his All-Star status. Teams that are willing to pay the high salary may trade for the player, which would add edges to the network as he moves teams.

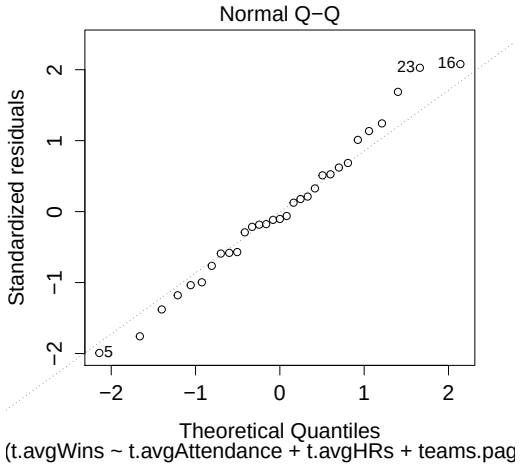
3.3 Team-Team Projection

Similarly to the player projection network, I looked at creating linear models from the team-team projection network. I again ran multiple models to see if there were any that not only were significant, but also included a measure of betweenness as a statistically significant component. Most of

the models noted that wins and division titles can be predicted by annual attendance[3], which makes sense, as teams that are winning can draw more fans to the stadium. Only one model included a statistically significant measure of betweenness, though. This model used average attendance and pagerank to model the number of division titles that a team has won during the period:

Coefficients	Estimate	Std. Error	t value	Pr(t)
(Intercept)	4.935e+01	6.790e+00	7.268	8.1e-08
Average Attendance	5.550e-06	1.476e-06	3.760	0.000832
Average HRs	1.640e-01	3.863e-02	4.246	0.000230
Pagerank	-2.591e+02	1.095e+02	-2.366	0.025400

The QQ plot for this model also shows that it is fairly accurate:

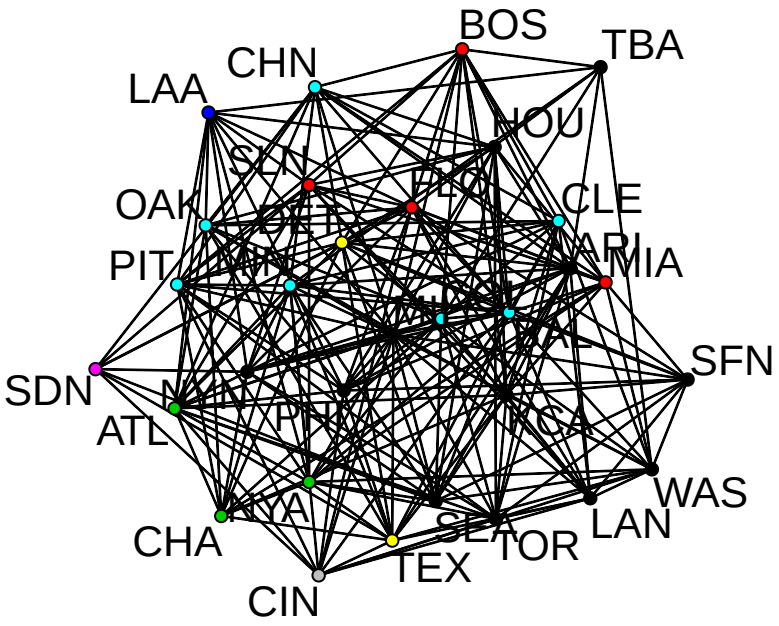


It once again makes sense that teams that regularly win division titles have higher attendance totals. Pagerank, however, is an interesting statistic here. The coefficient for pagerank is negative in the model, so teams with a higher pagerank would have fewer division titles in the model. In this case, more successful teams[4] would be less important in the network, and would have fewer players moving in and out from the team. These successful teams most likely have a core group of players that have been around for several years and have contributed to the success of the team

[3]Refer to the appendix for these models
[4]In terms of division titles

over time.

Finally, I looked at communities within the network of teams, to see if there were community relationships that aligned with divisions or geographical boundaries. While multiple communities were detected, there does not seem to be a correlation between community membership and division or community:



4 Conclusion

Professional baseball in America has a huge amount of statistical data available for use, from MLB organizations to professional statisticians to casual sabermetricians. The project attempted to

create a network using some of that data, and try to identify interesting characteristics about the MLB from a social network perspective. There are some interesting conclusions to be drawn from looking at the MLB as a network, and more work can be done in modeling different areas of the network, or different time periods.

With regards to Major Leage Baseball players, it is possible to construct a linear model to predict a player's All Star selection based on his salary, years active, and betweenness in the network of MLB players. For Major League teams, it is possible to model Division championships as a function of average attendance and pagerank scores.

The construction of the network was a difficult process, and choosing an appropriate time period was crucial, given computing limitations. With more time and processing power, I would like to run additional models and ERGM analyses on both the overal network and the bipartite projections, to determine how likely these networks are to be formed.

Finally, the discovery that the overall network had a diameter of 6 has inspired me to create a program in the vein of the "Six Degrees of Kevin Bacon" game. In reference to the most connected player in my graph, I will call the project "Six Degrees of Octavio Dotel," and hopefully expand the network further into the past, to see how connected Major League Baseball is over the years.

References

[1] Lahman, Sean. "Lahman's Baseball Database." Web. http://www.seanlahman.com/baseball-archive/statistics/