PageRank: Who Wins the 2016-2017 NBA MVP Award? (and more!)

Akilesh Bapu Prashanth Ganesh Nitin Manivasagan

March 16, 2017

1 Introduction

This year's NBA MVP candidate race has multiple qualified candidates competing for the prize. While Russell Westbrook is going all out this season, posting up a triple double almost every other game he plays, James Harden surprised the NBA community with his ability to keep the Houston Rockets near the top of the Western Conference. Other candidates such as Kawhi Leonard have been enjoying a "complete" season, contributing equally on offensive and defensive ends of the floor. And as always, LeBron James, considered by many to be the greatest of all time, is a contender for the prize as well.

2 Overview

To evaluate MVP candidates, we first considered applying our PageRank algorithm on all of the All-Star players this year. We had to discard certain players from consideration based on whether their teammates were also playing at the game. For instance, all-star players like Stephen Curry, Klay Thompson, and Draymond Green are not considered in this algorithm because these three, along with Kevin Durant, play for the Golden State Warriors. Of these four players, we analyzed player win-shares (how much a player has contributed to the team) and chose the player with the highest win-share number. We repeated this process for the Cleveland Cavaliers, who had multiple all-stars participate. Another team, the New Orleans Pelicans, also featured two all stars, but we disqualified Demarcus Cousins because he was traded to the team mid-season:

- Cleveland Cavaliers: **LeBron James** (10.5), Kyrie Irving (7.3), Kevin Love (5.4)
- Golden State Warriors: **Kevin Durant** (11.3), Stephen Curry (9.4), Draymond Green (7.3), Klay Thompson (5.2)
- New Orleans Pelicans: **Anthony Davis** (9.4), Demarcus Cousins (traded, so disqualified)

After removing players from the same time, the players we were left to analyze were: Anthony Davis (NOP), Carmelo Anthony (NYK), DeAndre Jordan (LAC), Giannis Antetokounmpo (MIL), Gordon Hayward (UTA), Isaiah Thomas (BOS), James Harden (HOU), Jimmy Butler (CHI), John Wall (WAS), Kawhi Leonard (SAS), Kemba Walker (CHA), Kevin Durant (GSW), LeBron James (CLE), Marc Gasol (MEM), Paul George (IND), Paul Millsap (ATL), Russell Westbrook (OKC)

3 The PageRank Algorithm

PageRank as Google uses it ranks web page quality based on the number of links to the web page and the quality of the links. We wanted to apply a similar algorithm to the top players in the NBA to see rank who is the best/who should get the MVP award. Just as websites are nodes and links are edges in PageRank, in our model, the nodes would be all the all-stars and the links would model how much better a certain player is than another player. Before we say we can solve this problem using PageRank, we have to model this problem just like the webpage ranking problem. Essentially, the webpage ranking problem is saying that when a link exists between two websites, the website getting linked to is only of higher quality if it has more incoming links that are of higher quality. With our example, we have framed it the same way where we have links from every player to every player but the weights define the importance of the link. So, while in PageRank, every webpage does not have a link to every web page, we have approximated that approach with weights between players that compare their head-to-head performances along with others as outlined below. This means that one edge with a weight is intended to represent k different edges, one for each type of weight and when that k^{th} weight is 0, it is as if there are k-1 edges now.

The PageRank algorithm tries to find the stationary distribution and use that to rank the pages in order. We used a similar approach, modeling the Markov chain as a transition matrix and using its eigendecomposition to find the eigenvector associated with the eigenvalue of 1. Then, to make sense of it better, we normalized it, took the absolute value, and sorted to get the ranking.

A $\in R^{nxn}$ is a transition matrix where each column to row mapping indicates how much better a player is based on their overall season performance and head-to-head performance against the other player. Then we found the eigenvector of A that corresponded to the largest eigenvalue (1) using numpy, made the data look nicer, then evaluated.

4 Methods

4.1 Generation of the Markov Chain:

Our Markov Chain comprised of nodes (All-Stars) and edges (a measure of how much more valuable one player is than another). The end result is a Markov Chain that, when the PageRank Algorithm is applied, yields an invariant distribution corresponding to the value of each player.

We experimented with several different methods of generating edge weights for the Markov Chain. We then experimented with various combinations of these methods that yielded the best results.

4.2 Method 1: Head-to-Head Plus/Minus Rankings

Since the PageRank algorithm uses links between different nodes to compute the stationary distribution, our first method used the NBA Plus/Minus statistic as link weights. Plus/Minus is an advanced statistic that captures how important a player is to the performance of a team by measuring how much they outscore (or get outscored by) an opposing team while that player is on the floor.

To capture the value one player directly offers over another player, we analyzed head-to-head matchups between two players (i.e. Kawhi Leonard and the San Antonio Spurs against LeBron James and the Cleveland Cavaliers) and extracted the Plus/Minus for each player against the other. By using this as the edge-weight method, this essentially captures how much more (or less) valuable one player is than another by directly comparing their performance, and resulting value, against each other.

4.3 Method 2: Individual Player Statistics

Method 1 did not seem to consider a lot of overall statistics and therefore the results we were getting from it werenâĂŹt very accurate. For example, a player could have beaten the team of another player, but that does not really capture the individual performance of a specific player as it does the teams. So while head-to-head plus/minus intuitively seemed effective, we went with another model of comparing 5 different statistics that were not directly between players but rather overall:

- Value over replacement (V) - The proposed value of the player on the team
- Defensive Win Shares (D) - The defensive effect of the player on the team
- Offensive Win Shares (O) - The offensive effect of the player on the team
- Games behind the number 1 seed (G) - The team record of the team the player is playing for relative to the top seed
- Plus/Minus Overall (P) - How many points the player adds/subtracts to the team based on his presence on average

These are ranked in order from most important to least important in determining how good a player is. We came to this conclusion based on the separation of values (i.e. most players had overall plus/minuses in the same range while their value over replacements were very different.

Our weights as decided were from 1 to 4 based starting from 5 to 1 where 2 and 3 had the same overall weight but different individual weights to indicate that one was more important than the other but not by a significant amount. The algorithm is as follows:

$$score = 4 \times V + 0.6 \times 3 \times D + 0.4 \times 3 \times O + 2 \times B + 1 \times P$$

For games behind number 1 seed, we had to assign ranges such that B equals:

- 1 for G $\in$ [0,5]
- 2 for G $\in$ (5,10]
- 3 for G $\in$ (10,15]
- 4 for G $\in$ (15, 20]
- 5 for G $\in$ (15,∞]

4.4 Method 3: Combining H2H Statistics With Individual Statistics

As mentioned above, there were shortcomings with both approaches. Method 1 was biased towards players with strong teams, while method 2 was biased towards players with stronger individual stats. The league MVP is chosen with a combination of both of these considerations, so we decided to modify our edge-weight scheme to have the Head-To-Head Plus/Minus weighted by the individual and team statistics. The advanced metrics mentioned in section 2 are accepted as being fairly representative of these factors.

5 Experiments

5.1 Experiment 1: Head-to-Head Plus/Minus Rankings

One of the main challenges with experimenting with method 1 was how to ensure none of the Head-to-Head statistics were negative. Players like Carmelo Anthony had negative Head-to-Head rankings against other players, so we had add a positive constant to all of the rankings to ensure negative point differentials would not mess up our scheme. Initially, we tried adding 20, but this did not remove the issue of negative edges. So, we added 50 to every ranking, and this solved the issue of negative edge weights.

5.2 Experiment 2: Individual Statistics Rankings

Experimenting with method 2 presented a challenge because we needed to figure out the best way to factor in individual statistics without placing too much/ too little value over certain categories. We decided that value over replacement would be the most important factor in analyzing the value of a player to a team, so we multiplied that value by 4. The next important value was defensive and offensive win shares, because these numbers talk about the influence of a player on both ends of the court. Since offense is more valued than defense (usually), we weighted offense slightly more. The next most important factor was how high a team was seeded: this is important for predicting playoff runs, but we weighted this less because while a player can lead a team, equal pressure falls on teammates to perform up to the necessary level. Finally, overall plus minus does not have as nearly much effect as the other factors, so we weighted this the least. While we had to mess with the offense and defense win rate, this formula gave us fairly accurate answers upon execution.

5.3 Experiment 3: Combining H2H Statistics With Individual Statistics

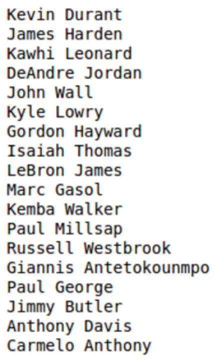
Once we figured out how to create optimal equations for experiments 1 and 2, experiment 3 was easy. We just took these functions we tuned in the first two experiments and multiplied them together, essentially creating weighted averages for each player's individual rankings.

6 Results/Analysis

The results show us that by using each method of evaluating players, what the likelihood is of them winning the award. The first name in the column is most likely to receive the award, and the likelihood of winning decreases as we go down the column.

6.1 Method 1

When we used this method, we calculated the predicted MVP rankings using PageRank and received good results:

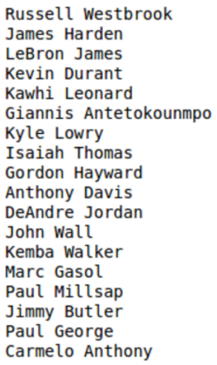


```
Kevin Durant
James Harden
Kawhi Leonard
DeAndre Jordan
John Wall
Kyle Lowry
Gordon Hayward
Isaiah Thomas
LeBron James
Marc Gasol
Kemba Walker
Paul Millsap
Russell Westbrook
Giannis Antetokounmpo
Paul George
Jimmy Butler
Anthony Davis
Carmelo Anthony
```

Figure 1: Ranking Players Based on H2H Match Ups

To the casual NBA fan, these results seem reasonable, since several of the best players in the league top the list. However, upon closer inspection, a more educated NBA fan would recognize that this method gives a higher value to those players who have better teams surrounding them. The general consensus in NBA media is that players such as Kevin Durant and DeAndre Jordan, while being great players, are on such strong teams that their success is in large part due to the strength of the rest of their team. For this reason we investigated another method.

6.2 Method 2



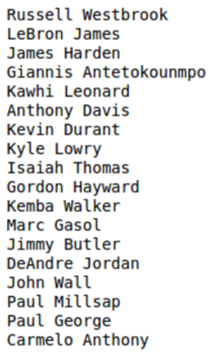
```
Russell Westbrook
James Harden
LeBron James
Kevin Durant
Kawhi Leonard
Giannis Antetokounmpo
Kyle Lowry
Isaiah Thomas
Gordon Hayward
Anthony Davis
DeAndre Jordan
John Wall
Kemba Walker
Marc Gasol
Paul Millsap
Jimmy Butler
Paul George
Carmelo Anthony
```

Figure 2: Ranking Players Based on Individual Statistics

These rankings fit more to the current NBA MVP Race but we still felt that leaving out the head-to-head defeated the purpose of PageRank and Markov Chains in general. There was no need to model this as Markov Chains since calculating these stats is not much more than just ranking players based on what their total score was. This inspired method 3.

6.3 Method 3

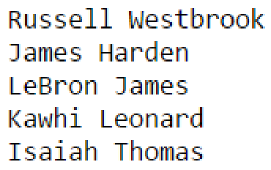
Once we generated edge-weights in this way, our MVP ranking followed:



```
Russell Westbrook
LeBron James
James Harden
Giannis Antetokounmpo
Kawhi Leonard
Anthony Davis
Kevin Durant
Kyle Lowry
Isaiah Thomas
Gordon Hayward
Kemba Walker
Marc Gasol
Jimmy Butler
DeAndre Jordan
John Wall
Paul Millsap
Paul George
Carmelo Anthony
```

Figure 3: Ranking Players Based on Individual Statistics and H2H Statistics

This ranking falls much more closely in line with the current NBA expert predictions. The current expert consensus is (1) James Harden, (2) Russell Westbrook, (3) LeBron James, (4) Kawhi Leonard, (5) Isaiah. In fact, when we run our model against just thse top 5 candidates we get something very similar:



```
Russell Westbrook
James Harden
LeBron James
Kawhi Leonard
Isaiah Thomas
```

Figure 4: Ranking Players Based on Individual Statistics and H2H Statistics

Our predicted ranking is quite strong as evidenced here. This is likely due to the fact that NBA experts take into account individual statistics as well as considerations as to the term valuable while determining the MVP frontrunners. However, the term valuable has long been quite subjective. By using Head-To-Head Plus/Minus to determine value in Method 1, we have taken a step towards objectively combining both.

7 Discussions/Limitations

While creating this ranking system, we ran into many challenges in trying to access the NBA data. Extracting stats from the NBA website proved to be very difficult until we found a python package called nba_py. Through this package, we could make simple API calls in order to extract player information.

Another major limitation is that it is very hard to model factors that are harder to quantify. For instance, a factor that plays a role in selection of NBA MVPs is whether the player has a compelling story. Russell Westbrook, who lost an important teammate in Kevin Durant, has a more exciting story than other MVP candidates such as Isaiah Thomas. Even though Westbrook's team is a lower-ranked team in the western conference, he still has a very good chance of winning MVP this year mainly because of his scenario. Other subjective factors that play into the decision making of MVP (i.e. Player X does not play well at the 3^{rd} quarter) are very difficult to quantify using our model.

8 Further Analysis of Markov Chains: Simulating a Seven Game NBA Series

Note: This was just an activity we decided to do for fun and to explore Markov Chains more in depth. It has nothing to do with our PageRank project.

To further analyze the role of Markov Chains in sports, (specifically basketball), we created a simple simulation at the end of our IPython file. We take 5 all stars and pit them against another group of 5 all stars. (Feel free to play with our Ipython file change the team lists to see how your favorite players would fare against each other). Once we have chosen the teams, we simulate a seven-game series between these two teams to see who would edge out (i.e. which team would win four games first). We replicate this under three different scenarios:

- Scenario 1: The team that wins game i has home court advantage on game i+1
- Scenario 2: The higher seeded team (which we modeled as the team that had a higher average overall ranking using method 2) has home court advantage on games 1,2,5, and 7. The other team has home court advantage on games 3,4, and 6 (this is how the NBA traditionally conducts 7 game series)
- Scenario 3: The team that wins game i loses home court advantage on game i+1

Under each of these scenarios, we ran simulations of 100, 500, and 1000 games. For scenarios 1 and 3, we ran separate simulations assuming the higher ranked team started with home court advantage, and then with the lower ranked team starting with home court advantage. To model home court advantage, we found the probability of victory for each team. Let A be the average rank of players in team 1, and B be the average rank of players in team 2. Then: (P(team 1 wins) $= \frac{A}{A+B}$). Then, from these probabilities, we added a certain handicap probability (0, 0.05, or 0.10 for each trial) if the team had home court advantage, and subtracted the handicap probability from the away team. We then simulated the seven game series as a Markov Process, trying to see if home court advantage really had a huge impact in the team's ability to win the series. Needless to say, we were quite surprised by the results! We represented our findings through graphs in the notebook.