Emory University, Goizueta Business School

Social Network Analysis of Steroid Use in Major League Baseball

Jake Arendsen, Adil A Kumar, Halle Steinberg

Social Network Analytics – ISOM 673

Professor Demetrius Lewis

December 9, 2019

Introduction and Research Question

When Senator George J. Mitchell's 20-month long investigation into the underground steroid market revealed the names of some of baseball's most successful and most popular players, the entire Major League Baseball (MLB) community was forced into action and reform of the league's policies and regulations at the time. Fans questioned the authenticity of the game itself and wondered how the controversy might influence younger players. Allegations that began with just a number of players soon developed into a large web of dozens of connected professional athletes, growing year after year, as players began to introduce teammates and friends to the network of performance enhancing drugs (PEDs).

In general, it is fair to assume that players are looking for and using illegal steroids in order to boost overall performance, and in turn, secure a higher-paying contract. Higher-paying contracts are typically fueled by a player's historical performance, so it makes sense that many players would seek PEDs or growth hormones to help them improve on performance. Although increased performance is typically the goal for PEDs users, is this something that a player can ensure will happen? Our analysis focuses on performance of players in the network, as we explore whether there is an obvious difference in performance among the players in our network. We dive into the interactions between players and dealers in the steroid network and how a player's decision to purchase drugs from a certain dealer influences his career performance.

The type of drug dealers in the steroid network ranges from bat boys to trainers to large pharmaceutical companies across the country and the league. Ease of access as well as cost of drug would be logical factors that a player might consider when deciding where to purchase drugs. More likely, however, where a player got his steroids is the result of where he knows his friends and teammates also got their drugs, without getting caught. In our analysis, we explore the relationships among players and where their drugs come from, examining core-periphery structures in the network, as well as large, dense clusters that represent the most popular suppliers.

As life-long baseball fans, the following report is motivated by our team's love for the game and dives into a scandal that transformed the way the game had been played prior to the release of the Mitchell Report. In the following analysis, we report our findings on the Mitchell Report as a social network, by analyzing three separate networks: a network of all players, connected by their relationships with the same drug dealers; a network of all drug dealers, connected by the players who got their drugs from them; and a network of both players and dealers, representing the spread of drugs and information throughout the Major League baseball community. We will begin by exploring each of these three networks, looking at structure and network statistics for players and dealers within each network. We will then dive into how these different relationships among players and dealers contribute to a player's career performance and whether or not certain dealers are more likely to be connected to better players.

More specifically, we will examine what the MLB steroid network looks like, analyzing differentiating factors between both players in terms of performance and suppliers in terms of quality of drug, quantified by user performance. We will look into whether there is an obvious enhancement in performance for players in our network, examining whether or not the Mitchell Report found that there was a serious impact on player performance.

Background

As mentioned above, our analysis will be performed on the steroid network thoroughly described in the Mitchell Report. In order to best communicate the results of our analysis, we want to give some context to the Mitchell Report, how it came about, how the investigations were performed, and what the results of the investigations were.

In the late 1990s and early 2000s, baseball was unknowingly being rocked by a massive steroids scandal throughout the league. Popular baseball stars like Barry Bonds, Mark McGuire, and Alex Rodriguez were shattering all kinds of records, creating must-watch television and hype around the idea of the “21st-century athlete”. The powerful network that generated such incredible buzz around the league managed to live for quite a few years in a state of secrecy; everybody had something to gain, so nobody shared. People soon started talking, and the secret drug network soon came under fire. First came Jose Canseco’s *Juiced*, an autobiographical count of Canseco’s time in the big leagues, with heavy emphasis placed on the use of steroids. Then came *Game of Shadows*, an investigative piece on the Bay Area Laboratory Co-Operative, that heavily detailed Barry Bonds’ alleged steroid use. After Major League Baseball Commissioner Bud Selig read the account, he quickly understood this was likely occurring around the league and that action must be taken.

Bud Selig appointed George J. Mitchell, a former United States Senator, to undertake a thorough investigation into the use of steroids in Major League Baseball. The League gave Mitchell near free reign to any information necessary to conduct his investigation. The Major League Baseball Players Association, however, was much less helpful. Mitchell and his associates interviewed over 700 different people through the process, many of whom were club officials, coaches, or other ballpark staff. Only 68 of the 500+ former players agreed to be interviewed, and very few current players agreed (the exact number was not included for animosity on the players’ behalf). As a result, the report lacks the full picture from the players’ perspective.

Through the investigation, Mitchell heard rumors that as many as half of all Major League Baseball players have used PEDs in some capacity. However, only 89 players were explicitly named in the report. Mitchell only named players for which he had very strong evidence of steroid use. In most cases, that meant he received an admission from the supplier, with some form(s) of evidence even corroborating the supplier. Due to this, the network Mitchell presents is very focused on only a few suppliers. It lacks the connections players make through the drugs, as well as many of the likely multiple connections to different dealers players probably had. It's very possible that players got drugs through other players or dealers not discovered by Mitchell in his investigation; as a result, they would not be included in the report despite very much so being players in the network.

Regulations and random testing were in place in the early 2000s to try to detect the use of steroids in players around the league. Random testing that began in 2004 seemed to reduce the overall use, but because human growth hormone is not detectable through urine samples, many players began turning to this drug. After the submission of the Mitchell Report, however, new regulations were created to prevent players from using these undetectable drugs. Previously HGH was unable to be detected, but MLB's Joint Drug Prevention and Treatment Program now tests for 74 PEDs, including HGH. Testing has since become more frequent, and punishments result in longer suspensions than those previously given.

Data Overview

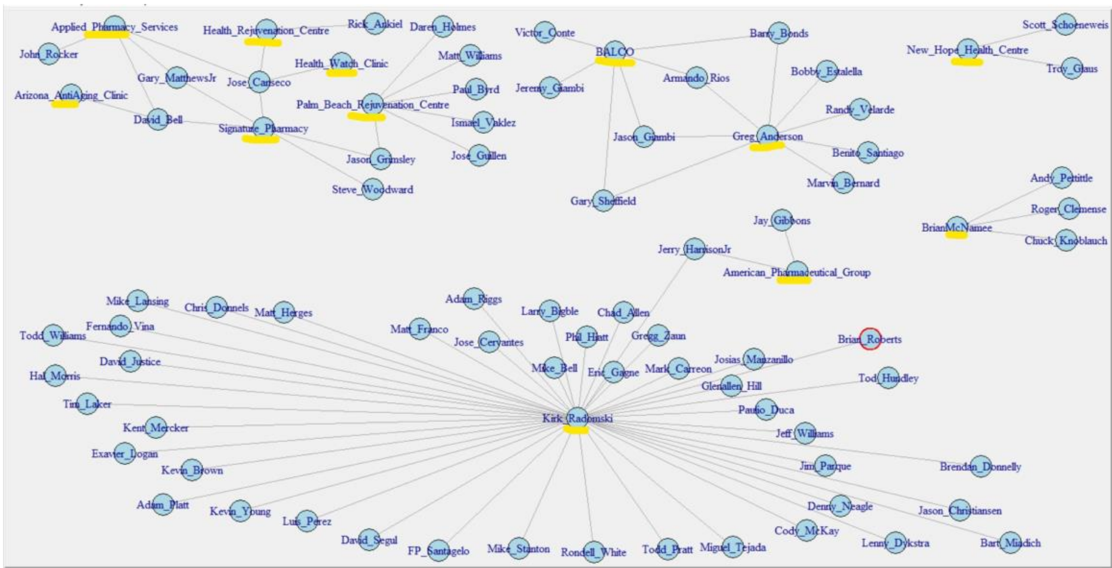
The data we performed our analysis on consisted of two datasets provided to us by Professor Lewis. The first is an affiliation matrix connecting 72 players to the dealer(s) who supplied their drugs to them, as indicated by the Mitchell Report. The other dataset is a player-level adjacency matrix connecting all players who share at least one dealer in common. Both datasets were fairly messy, requiring us to clean it, fix typos, and validate the data. Additionally, we acquired a lot more player data, particularly revolving around age and performance, to allow us to perform analysis related to player performance in the network. Most of the additional player performance data was pulled from Baseball-Reference, a well-known baseball statistics website. Refer to Figure 1 in the appendix for a full list of our data.

Exploratory Data Analysis

Once we had a clean dataset, we decided to create networks from the data and explored the dataset using network centrality measures. First, we created a network out of the complete player and provider network. It is important to note that our ties in this network are undirected, and so it only makes sense to consider degree centrality and not indegree and outdegree. Nodes highlighted in yellow represent the drug dealers and the other nodes represent players.

We can see that the total network consists of many clusters, with lots of players having little degree and little closeness. The largest cluster is around Kirk Radomski, who was a bat boy for the New York Mets from 1985-1995, before transitioning to dealing drugs as a personal trainer. One important group of dealers is the cluster of health clinics located in the top left of the chart below. These dealers were known for giving out human growth hormone(HGH) prescriptions without someone actually needing it or seeing a doctor and were one of the key groups being targeted by Mitchell Report. We also see few players acting as brokers for multiple dealers. Jose Canseco, for example, is connected to four different dealers. Jason Grimsley also acts as a broker because he got drugs from both Palm Beach Rejuvenation Center and Signature

Pharmacy, so he connects both of those groups. He is the sole connection between Palm Beach Rejuvenation Center and the rest of the network.



We then calculated relevant network centrality measures to better understand the network. Below are the summary statistics of network centrality measures:

	deg_out	close	betw	prank	eigen.cent
Min.	1.000	0.0001469	0.000	0.005852	0.0000
1st Qu.	1.000	0.0001784	0.000	0.006533	0.0000
Median	1.000	0.0003123	0.000	0.006533	0.1507
Mean	1.169	0.0002593	3.479	0.007368	0.0935
3rd Qu.	1.000	0.0003123	0.000	0.006783	0.1507
Max.	4.000	0.0003127	88.000	0.020562	0.1543

We can see from the degree measure that, on average, a player receives drugs from only one supplier, while there are a few who receive drugs from multiple suppliers. The range of closeness is pretty small indicating that, in general, players have a similar level of reachability within the network. The difference between the mean and maximum value of betweenness centrality shows that there are few important nodes who can help bridge disconnected groups in an overall network with lower mean betweenness. PageRank centrality shows that on average the flow of information is low. The mean and max of eigenvector centrality shows that all nodes have a similar level of influence in the network.

We then wanted to explore the network level centrality measures by player and see if players who got drugs from certain suppliers had certain trends in network centrality measures. From Figure 2 in our appendix, we see that players who received drugs from multiple suppliers in general were connected to more players in the network than those who received drugs from just one supplier. From Figure 3, it is evident that players who received drugs from multiple suppliers act like brokers in the network and help bridge disconnected groups. The betweenness values of players who received drugs from only one supplier is 0 and indicates that these players were in general disconnected from the rest of the network.

Figure 4 shows that players who received drugs from Kirk Radomski had the highest closeness in the network, indicating that Kirk Radomski helps these players reach different parts of the network and actually improves overall reachability in the network. From Figure 5, we see again that players receiving drugs from Kirk Radomski have high eigenvector centrality, indicating that Kirk Radomski has a high influence in the network. Apart from these players, it is interesting to note that only one player had eigenvector centrality greater than 0 - Jerry Harrison, Jr., who received drugs from Kirk Radomski as well as American Pharmaceutical Group, which further confirms the fact that Kirk Radomski has a high influence in the network.

From Figure 6, we can see that players who received drugs from multiple providers had higher PageRank centrality, indicating that getting drugs from multiple suppliers helps increase the flow of information within the network.

We then calculated correlations (Figure 7) between the network centrality measures to see if there were any interesting patterns within the network. The negative correlation between degree and closeness might be because we have a lot of players with high degree and low closeness, indicating the presence of an embedded cluster in the network, which is justified by the cluster around Kirk Radomski. The moderately high correlation between degree and betweenness shows that players with more ties help bridge disconnected groups and these ties are very important to the network. The high correlation between closeness and eigenvector centrality is probably because of the strong influence of Kirk Radomski in the network. Based on the correlations, we also decided what pairs of variables to include in regressions in order to best avoid the problem of multicollinearity.

We also calculated Jaccard similarity to see how players getting drugs from the same suppliers were related in the network. An interesting observation is that 44 of the 70 players have the same Jaccard similarity (0.58) and all these players were tied to Kirk Radomski, the most central dealer in the network.

We also created separate networks for players and dealers, by converting the affiliation matrix into separate incidence matrices. From the dealer network (Figure 8), we can see a distinct set of clusters, one of which is the group of dealers who were giving out HGH. From the player network (Figure 9), we can see 4 clusters, the largest one being those who got drugs from Kirk Radomski. For our analysis, we decided to focus on the total network rather than the individual dealer/player networks to answer all our research questions since we felt that the ties between players and drug suppliers are key to understanding interactions in the network.

Description of Analysis

We set out to answer as many of our research questions as possible. We chose to use regression as it was the primary tool we used in class and homework, as well as the tool we felt most comfortable using. In setting out to use regression, we realized many of our models weren't incredibly predictive, at least in the traditional R-squared view of analysis. This is likely due to

WAR being influenced by many different things, which couldn't simply be covered by drug use and age. However, we cannot add many of the logical things like performance statistics that do impact WAR, as those are impacted by drug use, so we have to accept the low predictive power of our models. That being said, WAR was still our best performance-related variable in terms of getting significant results, so we had to rely on it. Therefore, we focused on coefficients and their significance level in our analysis, rather than the predictive power of the whole model.

We had to make some assumptions about our data from the very beginning of our analysis. Because of the nature of how Mitchell named players in the report, our network is very dealer-focused and definitely incomplete. Due to this, we need to be cautious with any true player-level statistics or results we come up with, but we do feel safe assuming we have the full network for every dealer available to us. The players we are missing are likely connected to dealers not named in the report, to whom some of the players in the report would certainly also be connected. For the very few players who were still active at the time of the report, we also assumed their performance got worse after the report was released, as they would have had to stop using drugs moving forward in their career. As such, we decided to control for 2007 age in some regressions.

Illustration of Results

In order to best answer our questions, we created regressions in order to explore relationships between players and dealers as they relate to various baseball performance statistics. Many interesting insights came about from our analysis, including the possible development of two different networks within the overall steroid network as outlined in Senator Mitchell's report. Below are the results of our analysis.

Are certain players "better" because they got drugs from certain dealers?

```
Call:
lm(formula = WAR ~ BALCO + Greg_Anderson + Palm_Beach_Rejuvenation_Centre +
    New_Hope_Health_Centre + Signature_Pharmacy + American_Pharmaceutical_Group +
    Health_Rejuvenation_Centre + Applied_Pharmacy_Services +
    Arizona_AntiAging_Clinic + Health_Watch_Clinic + BrianMcNamee +
    `2007 age`, data = stats_totalnetwork)

Residuals:
    Min      1Q  Median      3Q     Max
-52.745 -10.218  -2.423   5.989  94.886

Coefficients:
                               Estimate Std. Error t value Pr(>|t|)
(Intercept)                    -54.7592    23.3358  -2.347  0.02245 *
BALCO                           36.1819    13.0280   2.777  0.00741 **
Greg_Anderson                    9.8922    10.8446   0.912  0.36552
Palm_Beach_Rejuvenation_Centre   6.6155    10.0196   0.660  0.51175
New_Hope_Health_Centre          17.3145    16.3807   1.057  0.29497
Signature_Pharmacy              -2.3848    14.7932  -0.161  0.87250
American_Pharmaceutical_Group    9.7272    16.6122   0.586  0.56049
Health_Rejuvenation_Centre      13.8806    23.2648   0.597  0.55311
Applied_Pharmacy_Services        6.0162    17.7266   0.339  0.73557
Arizona_AntiAging_Clinic         4.0797    28.4843   0.143  0.88662
Health_Watch_Clinic              3.0480    38.4137   0.079  0.93703
BrianMcNamee                    65.4979    13.5637   4.829 1.07e-05 ***
`2007 age`                       1.7814     0.6379   2.793  0.00710 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 22.41 on 57 degrees of freedom
  (2 observations deleted due to missingness)
Multiple R-squared:  0.4922,	Adjusted R-squared:  0.3853
F-statistic: 4.605 on 12 and 57 DF,  p-value: 3.647e-05
```

We built this regression to show whether getting drugs from certain dealers results in better performance by players. In this case, we used WAR to indicate player performance. We used all drug dealers except Kirk Radomski as binary dependent variables. We left Radomski out in order to use him as a reference point for all other dealers in the network since he is the most central dealer in our network and most players get their drugs from him, so this makes him a good baseline. We found that most dealers resulted in positive coefficients, indicating that players that get their drugs elsewhere and not from Radomski, had higher WAR on average.

Does being more connected within the steroid network indicate better career performance?

```
Call:
lm(formula = WAR ~ deg_in + betw + close, data = totalnetwork_4)

Residuals:
    Min      1Q  Median      3Q     Max 
-31.423 -11.840  -6.490   6.289 128.677 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)   
(Intercept)  6.361e+01  1.874e+01   3.395  0.00117 **
deg_in       5.700e-01  6.479e-01   0.880  0.38218   
betw        -1.478e-01  5.139e-01  -0.288  0.77449   
close       -1.535e+05  8.884e+04  -1.728  0.08867 . 
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 26.97 on 66 degrees of freedom
  (2 observations deleted due to missingness)
Multiple R-squared:  0.1488,	Adjusted R-squared:  0.1101 
F-statistic: 3.846 on 3 and 66 DF,  p-value: 0.01338
```

We built this model to examine the relationship between various network statistics and a player’s career WAR to see if the players that are more connected within the network are the high-performers or the benchwarmers. Based on our results, we can see that the only positive coefficient in this case is for degree, indicating that players that are connected to many other players and dealers will have a higher WAR on average. Alternatively, players high in betweenness and closeness are more likely to have lower WARs on average. This indicates that the lower performers are more likely to be brokers within the network as well as the ones that others rely on to be connected within the network. This might imply some sort of financial benefit for the worse players, which we dive into in the next model.

Does better performance indicate higher connectedness within the steroid network?

```
Call:
lm(formula = eigen.cent ~ WAR, data = totalnetwork_4)

Residuals:
    Min      1Q  Median      3Q     Max 
-0.7075 -0.5466  0.2779  0.3350  0.6745 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept)  0.722178   0.066027  10.938  < 2e-16 ***
WAR         -0.005859   0.001952  -3.002  0.00375 ** 
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.4634 on 68 degrees of freedom
  (2 observations deleted due to missingness)
Multiple R-squared:  0.117,	Adjusted R-squared:  0.104 
F-statistic: 9.012 on 1 and 68 DF,  p-value: 0.003749
```

This regression models the relationship between a player's career performance, represented by WAR, and a player's eigenvector centrality. We can see from the results of this simple model that a player who has a higher eigenvector centrality has a lower career WAR on average. Eigenvector centrality is a measure of a player's influence within the network, so those players with higher measures are the ones connecting other players to drug dealers. The fact that those with higher eigenvector centralities are the ones with lower WARs might indicate that there is some financial gain for these players if they refer others to their dealers. Since these players are not as high-performing, they likely don't make as much money, and are looking to make their money in this way. Similarly, better players might not want to take on the risk of spreading the word about their dealers, so they don't provide as much influence or power to the network as those lower-performing players.

How does similarity among players influence their WAR?

```
Call:
lm(formula = WAR ~ average_Jaccard, data = totalnetwork_4)

Residuals:
    Min      1Q  Median      3Q     Max
-29.467 -12.056  -6.273   3.419 133.385

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)       32.894      5.812   5.660 3.31e-07 ***
average_Jaccard  -37.952     12.654  -2.999  0.00378 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 27.06 on 68 degrees of freedom
  (2 observations deleted due to missingness)
Multiple R-squared:  0.1168,	Adjusted R-squared:  0.1038
F-statistic: 8.995 on 1 and 68 DF,  p-value: 0.003781
```

This regression tells us that an increase in average Jaccard similarity actually results in a decreased total WAR, all else equal. This means that players that are more similar to who they get their drugs from actually have a lower WAR, on average. One implication for why this might be is that most players are buying cheaper drugs, so they're more connected and, because of the worse drugs, have a lower WAR. In this context, it is similar to a lemons problem - none of the players really know the difference in drug quality, so they just buy from the suppliers of cheaper drugs, leaving much less business for suppliers that are actually providing high-quality drugs.

Is it the case that the providers are making the players better through their drugs or that the more skillful players are more likely to find the highest quality drugs?

```
Call:
glm(formula = Kirk_Radomski ~ WAR, data = stats_totalnetwork)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-0.7075  -0.5466   0.2785   0.3356   0.6750

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.722121   0.066132  10.919  < 2e-16 ***
WAR         -0.005857   0.001955  -2.996  0.00381 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for gaussian family taken to be 0.2154643)

    Null deviance: 16.586  on 69  degrees of freedom
Residual deviance: 14.652  on 68  degrees of freedom
  (2 observations deleted due to missingness)
AIC: 95.175

Number of Fisher Scoring iterations: 2
```

```
Call:
lm(formula = WAR ~ Kirk_Radomski, data = stats_totalnetwork)

Residuals:
    Min      1Q  Median      3Q     Max
-28.144 -12.058  -6.283   3.417 132.156

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)     30.644      5.209   5.883 1.36e-07 ***
Kirk_Radomski  -19.912      6.646  -2.996  0.00381 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 27.07 on 68 degrees of freedom
  (2 observations deleted due to missingness)
Multiple R-squared:  0.1166,    Adjusted R-squared:  0.1036
F-statistic: 8.977 on 1 and 68 DF,  p-value: 0.003814
```

For our next research question, we wanted to consider the possible direction of causality between a player's skills and his seeking out certain providers. We wanted to see whether better players were seeking out certain dealers, and on the other hand, whether players that went to certain dealers ended up with worse career WARs as a result. We ran a pair of regressions for every single dealer, so 12 pairs in total. The image on the left above represents an example of our first type of regression, using WAR to predict the probability that a player would get his drugs from that dealer. For our second type, we used the binary variable for each dealer that represents whether a player got drugs there or not to predict the change in career WAR for a player that gets his drugs from that dealer.

Our regressions resulted in many conclusions from our network. Most interestingly, many of the dealers that were convicted of dealing human growth hormones to players without requiring a prescription had negative coefficients for both regressions. This means that good players were less likely to go to these dealers to get their drugs, and on the other hand, those that did get their drugs from these dealers had lower career WARs. We also found that good players had the highest percentage of getting their drugs from Brian McNamee among all dealers. Also, those that got their drugs from him had a higher WAR by over 60 points than those that did not get their drugs from him. These results imply that players that were already good knew where to get their drugs from, and as a result, had even higher WARs afterwards because they got their drugs from those dealers.

As we've seen in much of our analysis, the most connected dealer is not always the best one. In this case, Kirk Radomski dealt out drugs to the most players, but those players were the lower-performing ones among the network. We see this clearly in Radomski's pair of regressions, as players with higher WAR have a lower chance of getting drugs from him, as well as the fact that those that did get drugs from Radomski had a career WAR that was about 19 points lower on average than those that did not.

Does being more central or similar in the steroid network imply having higher performance during the steroid era?

```
Call:
glm(formula = greater_in_steroid_era ~ eigen.cent + average_Jaccard,
    data = totalnetwork_4)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-0.8479  -0.2306   0.2500   0.2500   0.4807

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)      -0.1751     0.3831  -0.457    0.651
eigen.cent       -5.5462     2.8644  -1.936    0.063 .
average_Jaccard  11.0762     5.5073   2.011    0.054 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for gaussian family taken to be 0.2021313)

    Null deviance: 6.7742  on 30  degrees of freedom
Residual deviance: 5.6597  on 28  degrees of freedom
  (41 observations deleted due to missingness)
AIC: 43.255

Number of Fisher Scoring iterations: 2
```

```
Call:
glm(formula = less_after_steroid_era ~ eigen.cent + average_Jaccard,
    data = totalnetwork_4)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-0.8221   0.1386   0.2095   0.2612   0.2612

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)       0.8740     0.2228   3.923 0.000366 ***
eigen.cent        0.3972     1.8743   0.212 0.833316
average_Jaccard  -0.9115     3.5394  -0.258 0.798202
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for gaussian family taken to be 0.1863171)

    Null deviance: 6.9750  on 39  degrees of freedom
Residual deviance: 6.8937  on 37  degrees of freedom
  (32 observations deleted due to missingness)
AIC: 51.184

Number of Fisher Scoring iterations: 2
```

We ran regressions predicting both whether or not a player's WAR increases during the steroid era, which we define as the period between 1995 and 2003, and whether or not it decreases after the steroid era. We created a binary variable for whether a player's average seasonal WAR was greater during the steroid era than prior or less after the steroid era than during it. For our prediction during the steroid era, we see average Jaccard similarity has a large positive effect on steroid era WAR, while eigenvector centrality has a sizable negative effect on steroid era WAR. This would imply that more similar players saw better steroid era WARs, while more central players saw lower WARs, when compared to their WARs during the pre-steroid era (pre-1995). This falls similar to trends we have mentioned earlier; more central players seem to be motivated by something other than performance to use drugs. Meanwhile, the positive similarity effect is likely caused by those players sticking with the best dealers to maximize their performance. We also checked this regression in reverse to see if there was any evidence of an increase in performance causing an increased or decreased centrality, but found no good evidence (Figure 10).

When examining performance in the post-steroid era (post-2003), the intercept implies most players were worse performers after the steroid era, which falls in line with what one would assume should happen. More central players seemed more likely to do worse after the steroid era, while more similar players seemed more likely to do better. This leads us to believe to some extent that the more similar players were naturally better than the more central players, at least to start, and again, that performance might not be the primary motivation for the most central players.

Conclusion and Implications

Many of our regressions resulted in the conclusion that worse players are the ones that are more highly connected within the network, acting as brokers for certain popular drug dealers and increasing reachability for other players to gain access to those drugs. These players have much lower career performance statistics, indicating that they might be participating in the network with a goal other than to better their own playing performance. These players might be receiving some sort of financial incentive to bring other players to those dealers, as indicated by the high degree for many of these dealers that are connected to low-performing players.

From our analysis, we see baseball's steroid network is not as simple as one would guess upon first glance. There seems to be clear better and worse dealers, and good players seem to have no issue connecting themselves to the good dealers. Bad players seem comfortable enjoying their better connected positions between worse dealers than seeking out better dealers, while good players seem very comfortable enjoying their isolated connections to good dealers. There are plenty of reasons this could occur; one interesting hypothesis we came up with is that this was actually two separate steroid networks. One network consists only of already accomplished players who take drugs solely to get even better, while a second network consists of bad players more intent on increasing the flow of lower quality drugs, potentially due to addiction, a feeling of "fitting in", or a financial gain from the drug flow.

Moving forward, it would be interesting to explore these two possible steroid networks and their current implications in today's baseball world. Since the Mitchell Report is truly the only investigation into steroid use at such a large scale, it is hard to pinpoint those who might be using drugs today and even make assumptions for those that do versus those that don't. However, if the Major League understands that these two types of players, and therefore two types of networks, might exist in today's world, it might be easier for them to narrow down certain starting points for finding steroid users, starting with the extremes of highly-talented stars and below-average benchwarmers and further diving into those players' networks.

Appendix

Figure 1:

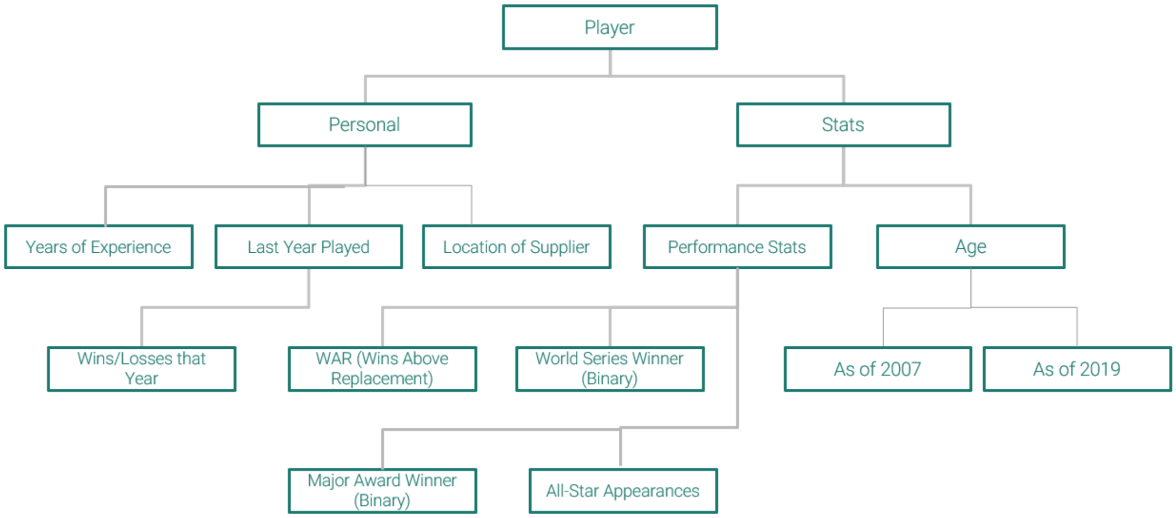


Figure 2:

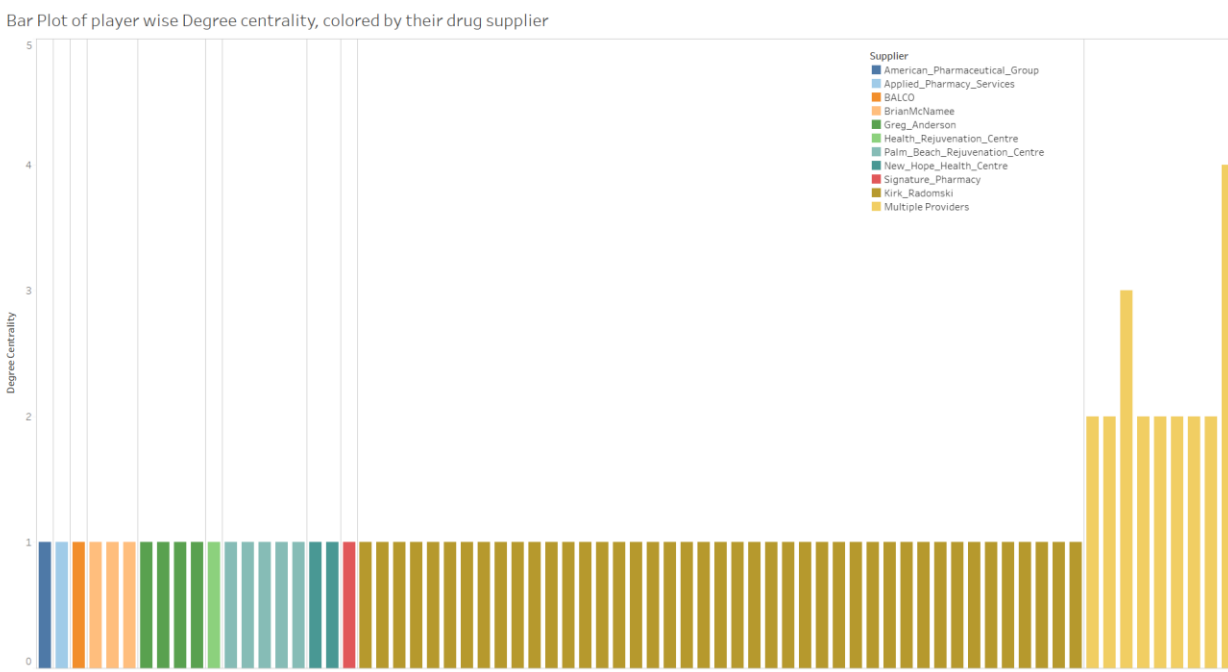


Figure 3:

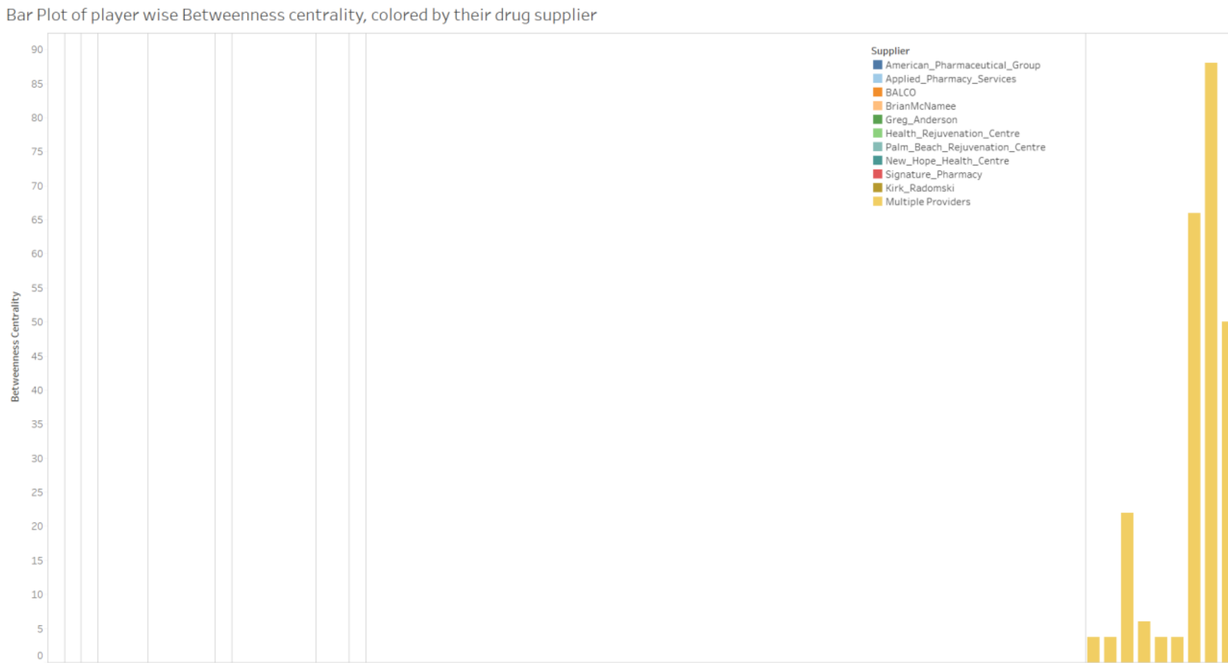


Figure 4:

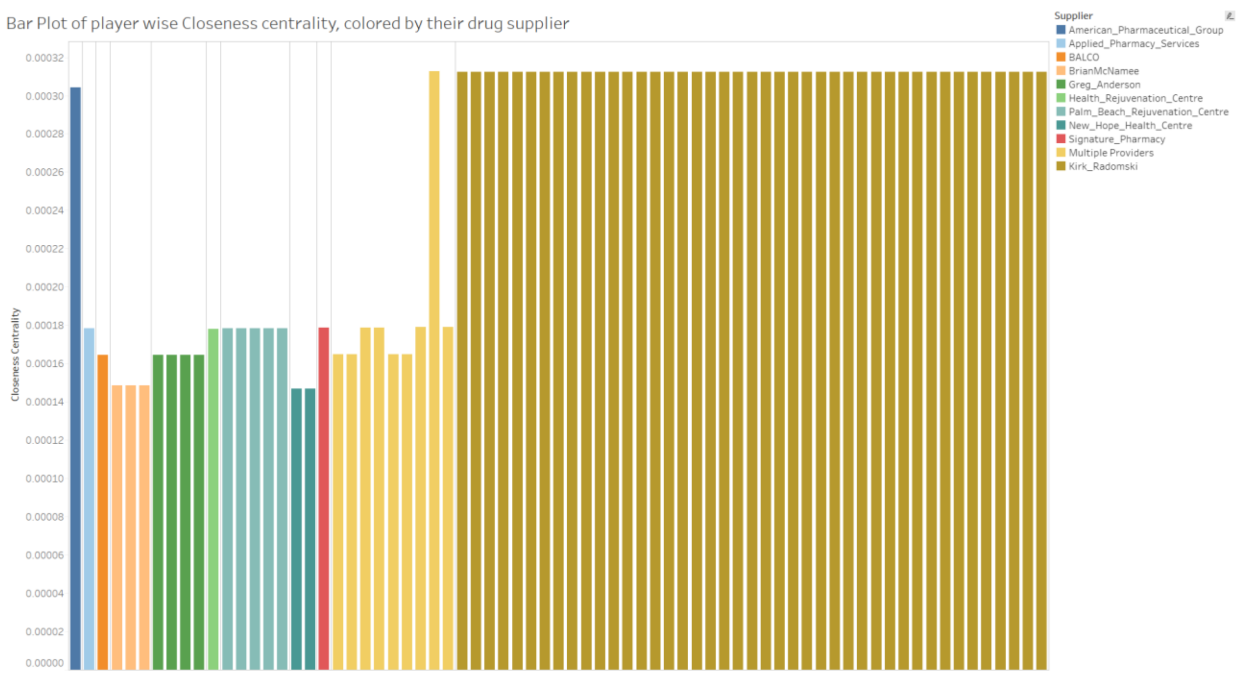


Figure 5:

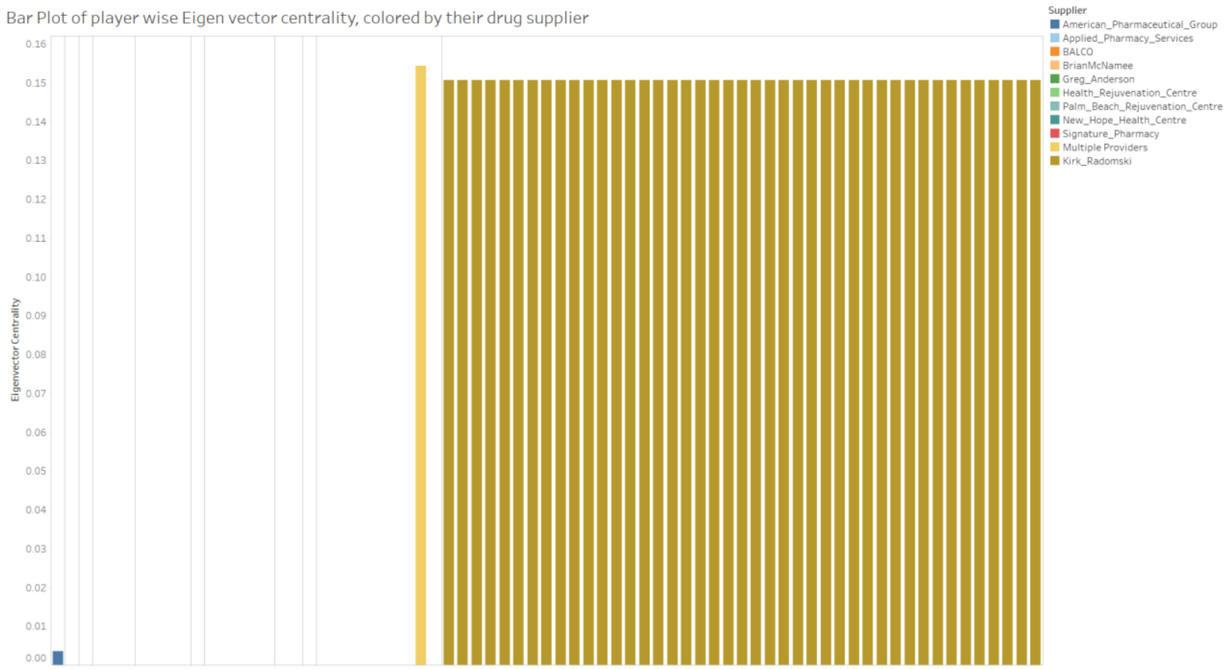


Figure 6:

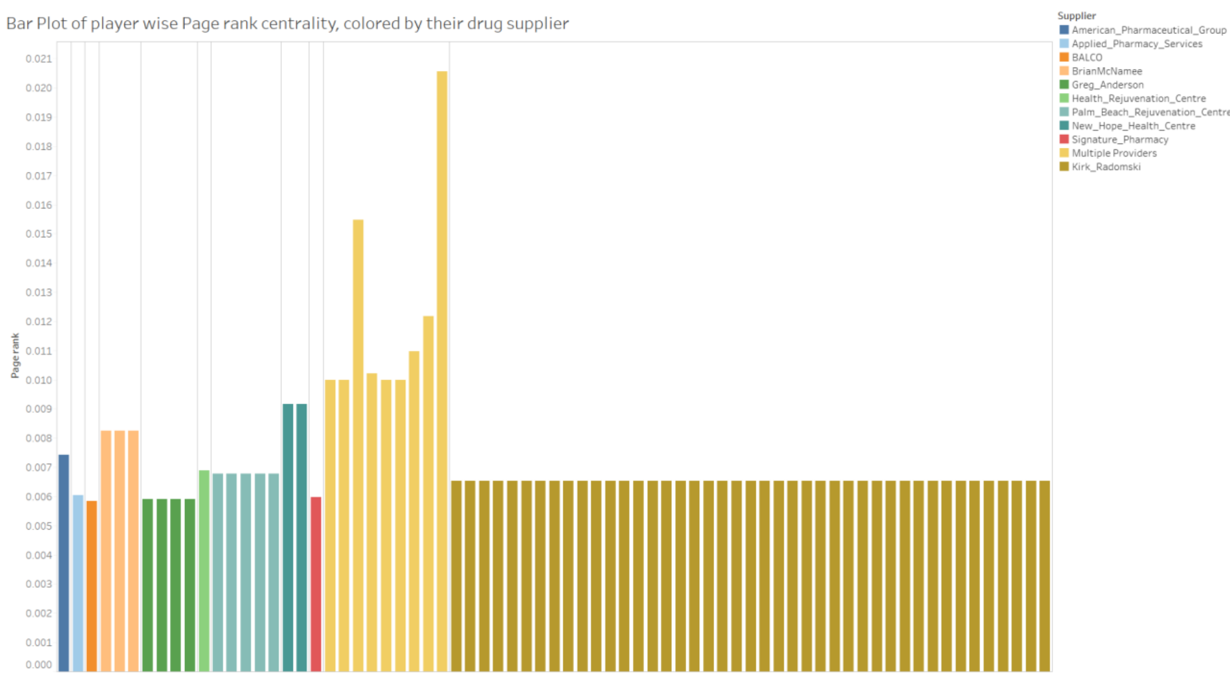


Figure 7: Correlation between network centrality measures

	Degree	Closeness	Betweenness	Eigenvector	Page Rank
Degree	1	-0.36	0.65	-0.37	0.95
Closeness		1	-0.12	0.97	-0.39
Betweenness			1	-0.13	0.68
Eigenvector				1	-0.39
Page Rank					1

Figure 8: Dealer Network

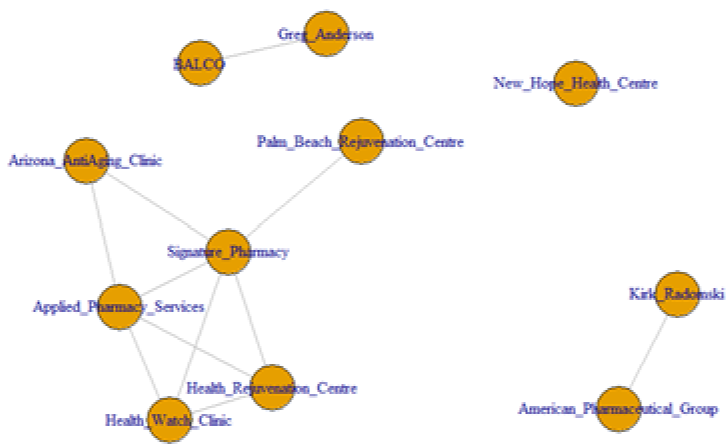


Figure 9: Player Network

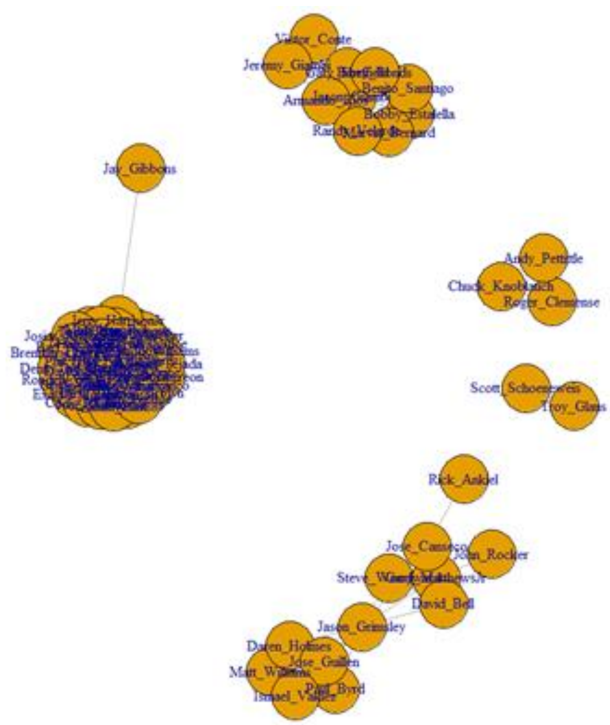


Figure 10: eigen centrality ~ increased steroid era performance regression

```
Call:
lm(formula = eigen.cent ~ greater_in_steroid_era, data = totalnetwork_4)

Residuals:
    Min      1Q  Median      3Q     Max 
-0.7139 -0.4997  0.2856  0.2856  0.4997 

Coefficients:
                       Estimate Std. Error t value Pr(>|t|)   
(Intercept)              0.4997     0.1529   3.269  0.00278 **
greater_in_steroid_era   0.2142     0.1858   1.153  0.25833   
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.4835 on 29 degrees of freedom
  (41 observations deleted due to missingness)
Multiple R-squared:  0.04383,	Adjusted R-squared:  0.01086 
F-statistic: 1.329 on 1 and 29 DF,  p-value: 0.2583
```

Sources

http://files.mlb.com/summary.pdf

https://www.nydailynews.com/sports/baseball/mitchell-report-impact-mlb-10-years-article-1.3694905

https://www.baseball-almanac.com/legendary/Mitchell_Report.shtml

https://www.nytimes.com/2006/05/07/books/chapters/0507-1st-fain.html

https://www.baseball-references.com