Duke Basketball Passing Networks

A Thesis
Presented to
Department of Statistical Science
Duke University

Sonia Xu

April 2018

Approved for the
Bachelor of Science in Statistical Science

Dr. Alexander Volfovsky

Dr. Katherine Heller

Dr. Scott Schmidler

Dus X. Name, DUS

Acknowledgements

I want to thank a few people.

Preface

This is an example of a thesis setup to use the reed thesis document class (for LaTeX) and the R bookdown package, in general.

Table of Contents

List of Tables

List of Figures

Abstract

Basketball has traditionally been viewed as a team sport, requiring the collaboration of the entire team to successfully bring a ball to the basket. The scope of this work attempts to model basketball possessions for each game to explain team characteristics that can lead to successful plays. Duke Basketball data from the 2014-2016 seasons was converted into passing networks, with each node representing a unique player and each edge containing information about a pass. These passing networks not only represent the temporal qualities of basketball, but also capture how interactions among players can lead to positive outcomes during a game. Network characteristics, such as measures of centrality, cohesiveness and latent position of the players, were considered as predictors for play outcomes. This approach suggests that successful Duke basketball teams have collaborative team dynamics. These results can be applied to most team settings–uniform collaboration can lead to better results than teams dominated by one leader.

Dedication

You can have a dedication here if you wish.

Chapter 1

Introduction

In basketball, a boxscore provides the statistical summary of the game via defensive, offensive, and overall success metrics. The National Basketball Association's records show that the first boxscore was produced by the Boston Celtics in the 1946-1947 season. Initial records kept track of basic basketball statistics for each player through measures like minutes played (MP), field goals made (FGM), and free throws made (FTM). Seventy-one years later, these metrics are still popular today. While the National Basketball Association has boosted its number of metrics to better summarize the game to include metrics like rebounds per game (RBG), player efficiency rating (PER), free throw attempts (FTA), and 3 field goals made (3FGM), these metrics still cannot capture the entirety of the game because they do not take into account the opposing team's defense/offense, nor previous plays that significantly influenced the flow of the game.

Basketball is not the only sport that has encountered this modeling problem. Soccer, a sport similar to basketball in that it requires a team-oriented approach and it dynamically changes from moment to moment, has also experienced a similar need by academia and major soccer teams to better utilize the data to more fully understand the game. One popular metric that has yet to be uniformly adopted is evaluating a player's passing capabilities and team-value. Although a consensus has yet to be adopted for the best metric, scholars from academia and the National Basketball Association have sought to capture the game of basketball more robustly in a similar fashion–via passing networks.

Chapter 2

Literature Review

Passing forms the backbone of all team contact sports. To advance a ball to the goal successfully, players must work together to dribble/kick/throw the ball to its destination. Each pass to another player can be considered a connection. These connections can be grouped together to form a network of passes. Previous works have captured these passing networks in soccer and basketball both statically and dynamically–this literature review will explore the different methods used to understand the value of a player and team, and the best practices for modeling network data.

"Flow Motifs in Soccer: What can passing behavior tell us?" by Joris Bekkers and Shaunak Dabadghao was released in the 2017 MIT Sloan Sports Analytics Conference, and focused on the static passing networks of the last 4 seasons of 6 big European leagues with 8219 matches, 3532 unique players and 155 unique teams. Passing sequences were denoted as a sequence of all players involved five seconds before an attempted score. This paper created radar graphs that illustrated the most popular passing sequences by player, and compared radar graphs to identify similar players. Passing sequences within teams were also compared between teams by clustering the different passing styles of the different teams. Key players were determined by the frequency that they were included in the passing sequences.

"Exploring Team Passing Networks and Player Movement Dynamics in Youth Association Football (Soccer)" by Bruno Goncalves, Diogo Coutinho, Sara Santos, Carlos Lago-Penas, Sergio Jimenez, and Jamie Sampaio compared the passing sequences of two games played by two groups that differ in age range, which showed that regardless of age, network centrality was distinctive in both groups, and affirmed the long-held belief that more passes lead to better game outcomes. Similar to the first paper, key players were the ones most frequently involved in the passing sequences. This paper created weighted graphs of the passing sequences, which better visualized the passing structure of the team, and made it easier to identify important players.

"Basketball Teams as Strategic Networks" by Jennifer H. Fewell, Dieter Armbruster, John Ingraham, Alexander Petersen, and James S. Waters provided measurements to assess team entropy. First recording the complete 30 seconds of a possession as a passing sequence, they discovered that recording the last three nodes (players) before a shot attempt was a better way to record passing sequences to avoid noisy passing data. Although they were able to recognize various aspects of team dynamics through weighted graphs like the second paper, they did not find a consistent predictor of positive game outcomes. This paper also identified that in general, teams typically range between two playing styles: always passing to the best player or having no distinct patterns in passing. These patterns can be noted by distinct betweenness scores and uniform betweenness scores, respectively. Weighted graphs clearly illustrated the two different playing styles. Also, the paper found that the positions most involved with successful shots were: 1. PG 2. SG 3. SF 4. PF 5. CN.

Joachim Gudmundsson and Michael Horton summarised a variety of methods that utilize object tracking data to analyze team and player performances in "Spatio-Temporal Analysis of Team Sports – A Survey." Their research survey spanned modeling passing networks via graph theory to calculating rebound probability with spatial coordinates. In particular, work conducted by Daniel Cervone, Alex D'Amour, Luke Bornn, and Kirk Goldsberry attempted to capture the game wholelistically via a new measure called Expected Possession Value (EPV) in the paper "A Multiresolution Stochastic Process Model for Predicting Basketball Possession Outcomes." This new metric uses three models–a Microtransition Model, Macrotransition Entrance Model, and a Macrotransition Exit Model–to capture the spatial biases of each player and the in-game effects of pressure, so that it can measure the likelihood of a successful play (made shot) given the previous sequence of events. To compare players against the league-average scores, they also calculated Expected Possession Value -Adjusted as an application for teams.

Peter Hoff explains in "Bilinear Mixed Effects Models for Dyadic Data" the structure of the AMEN package by describing the different components of the model, which reinforces AMEN's suitability to model network data. A Monte Carlo Markov algorithm, the model encompasses modelling linear, bilinear, and dyadic covariates with multivariate normal distributions. A dataset of international relations in Asia was used to demonstrate the robustness of this model in revealing the transitivity and clusterability of the observation.

Bailey Fosdick and Peter Hoff use AddHealth data in "Testing and Modeling Dependencies Between a Network and Nodal Attributes" to introduce a joint model that accounts for network factors and attributes. The AddHealth dataset captures samesex friendship between high school students, where students were asked to rank their top five friends. Applying the model to this dataset via the AMEN package, network features include rank information between students and nodal attributes like exercise frequency of each student. Hoff and Fosdick compare the performance of their joint model against a model that only captures the effect of nodal attributes and show that the joint model has a lower mean squared error in predicting missing values over a 20-fold cross validation. While the paper mainly focuses on demonstrating the robustness of this model, there still exist challenges in determining the level of dimensionality.

Peter Hoff in "Modeling Homophily and Stochastic Equivalence in Symmetric Relational Data" proposes the benefits of modelling data in a latent space. Models that transform datasets that contain network features into latent space can capture two characteristics: homophily and stochastic equivalence. Stochastic equivalence is when nodes can be grouped based on similar characteristics, and homophily is when nodes with similar characteristic nodes are more likely to have a relationship than with different characteristic nodes. Models that measure these relationships through latent eigenvalues perform better than models measured through latent distance or latent class. This result constructs the impetus for the AMEN package to utilize a latent eigenvalue model to capture network and attribute data.

Chapter 3

Dataset

The dataset is from the Duke University Men's Basketball SportsVu tracking data. Features were created by taking snapshots of the game every 1/25th of a second and recording the player's location, action, team, etc. Data was collected for each season from 2013-2016; the dataset totals about 132,000 observations and 98 features. Since the data is owned by the Duke Men's Basketball team, the data is private and cannot be shared.

The dataset was presented in 3 different XML files:

Boxscore Data: This dataset shows the overall player statistics (assists, points, rebounds) for each game.The Boxscore dataset was used as a reference for player performance when modeling the posterior means of the latent eigenvalue model.

Play by Play: This dataset provides a moment summary (dribble, foul, pass) at time t for each game. The Play by Play dataset was used to divide the raw data into possessions, which was then converted into individual passing networks.

Sequence Optical: This dataset provides the locational summary of each player for each game. The Sequence Optical dataset was used to map the passing order for each possession.

Chapter 4

Data Cleaning

Initially provided as XML files, the datasets were converted into csv files and then merged to create a final dataset with 132,000 observations and 98 features. As this project primarily focuses on passing, the data was converted into network data for each game. Each game consists of an array of matrices that represent the passing count between players for each possession.

Below is an example of a 10x10 matrix for a possession. The rows indicate the passer, and the column indicates the receiver.

	100023	100283	839023	456782	222789	134783	111124	098783	352671	213416
100023	0	1	0	3	0	0	0	0	0	0
100283	0	0	0	0	0	0	0	0	0	0
839023	0	1	0	0	0	0	0	0	0	0
456782	0	0	0	0	0	0	0	0	0	0
222789	0	0	0	0	0	0	0	0	0	0
134783	0	0	0	0	0	0	0	0	0	0
111124	0	0	0	0	0	0	0	0	0	0
098783	0	0	0	0	0	0	0	0	0	0
352671	0	0	0	0	0	0	0	0	0	0
213416	0	0	0	0	0	0	0	0	0	0

4.1 Changes in Shot Clock Time

As college basketball is a consistently changing sport, the NCAA changed the play rules for the 2013-2014 college basketball season. Instead of a 35 second shot clock, the NCAA established a 30 second shot clock. Since this work does not have a temporal component, the rule change does not affect the results of model building drastically. However, the extra five seconds may have allowed players to pass the ball more frequently, which would affect the passing matrices.

Chapter 5

Exploratory Data Analysis

Initial analysis of the data focused on understanding the many features available in the Duke Men's Basketball dataset. This exploratory data analysis explores shot attempt patterns through the years, as well as potential biases with shot location.

5.1 Changes in Shot Attempt Patterns

As one of the best basketball programs in the nation, Duke University Men's Basketball draws in a number of highly desirable and NBA-ready recruits each year. For this, most players stay for only a year before signing and playing for the National Basketball Association. A popular trend for many skilled basketball players, this transition to professional basketball has been coined by players as being "one-and-done." Duke had two players (Rodney Hood, Jabari Parker) drafted in the 2014 draft, three players (Jahlil Okafor, Justise Winslow, and Tyus Jones) drafted in the 2015 draft, and one player (Brandon Ingram) drafted in the 2016 draft. With so many players playing the minimum in college, this paper concentrates on the analysis of players who played more than one season with the Duke Men's Basketball team, and had significant minutes with their time at Duke. With these requirements, it is difficult to find the perfect player for analysis because players like Marshall Plumlee, only had significant playing time his senior year because it took time to fully develop him as a competitive player.

Player 103929, on the other hand, serves as an interesting example because he had consistent minutes for the 2013-2015 seasons. Player 103929's shot attempts were thus divided into each year to understand how his shooting style has changed during his time at Duke.

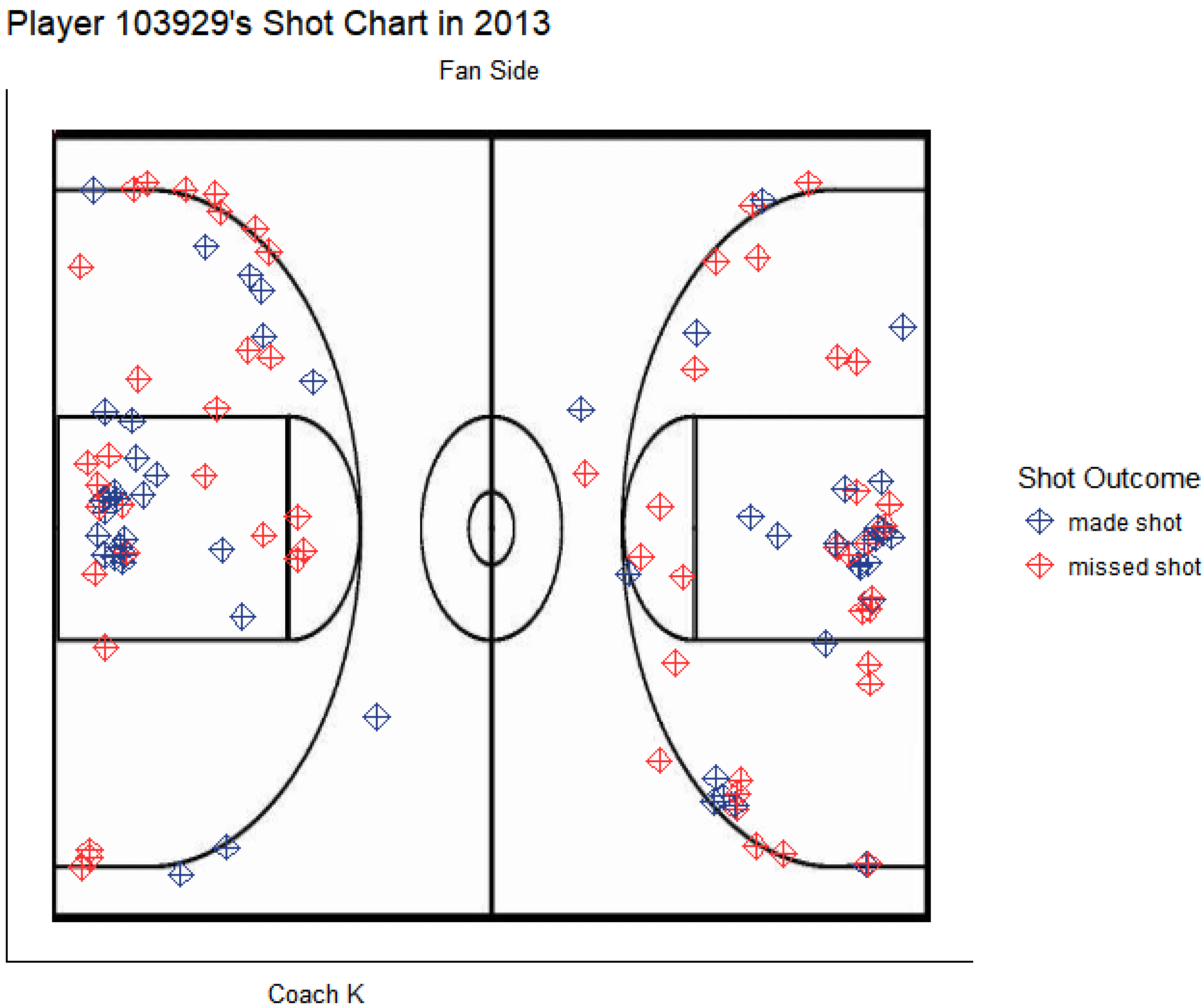


Figure 5.1

Looking at the Player 103929's shot attempts for his junior season, he was fairly even with his shooting, missing most of his 3 point shots, and hitting most of his 2 point shots in the paint. It appears as though he prefers to shoot from the right wing slightly more than he shoots from the left wing.

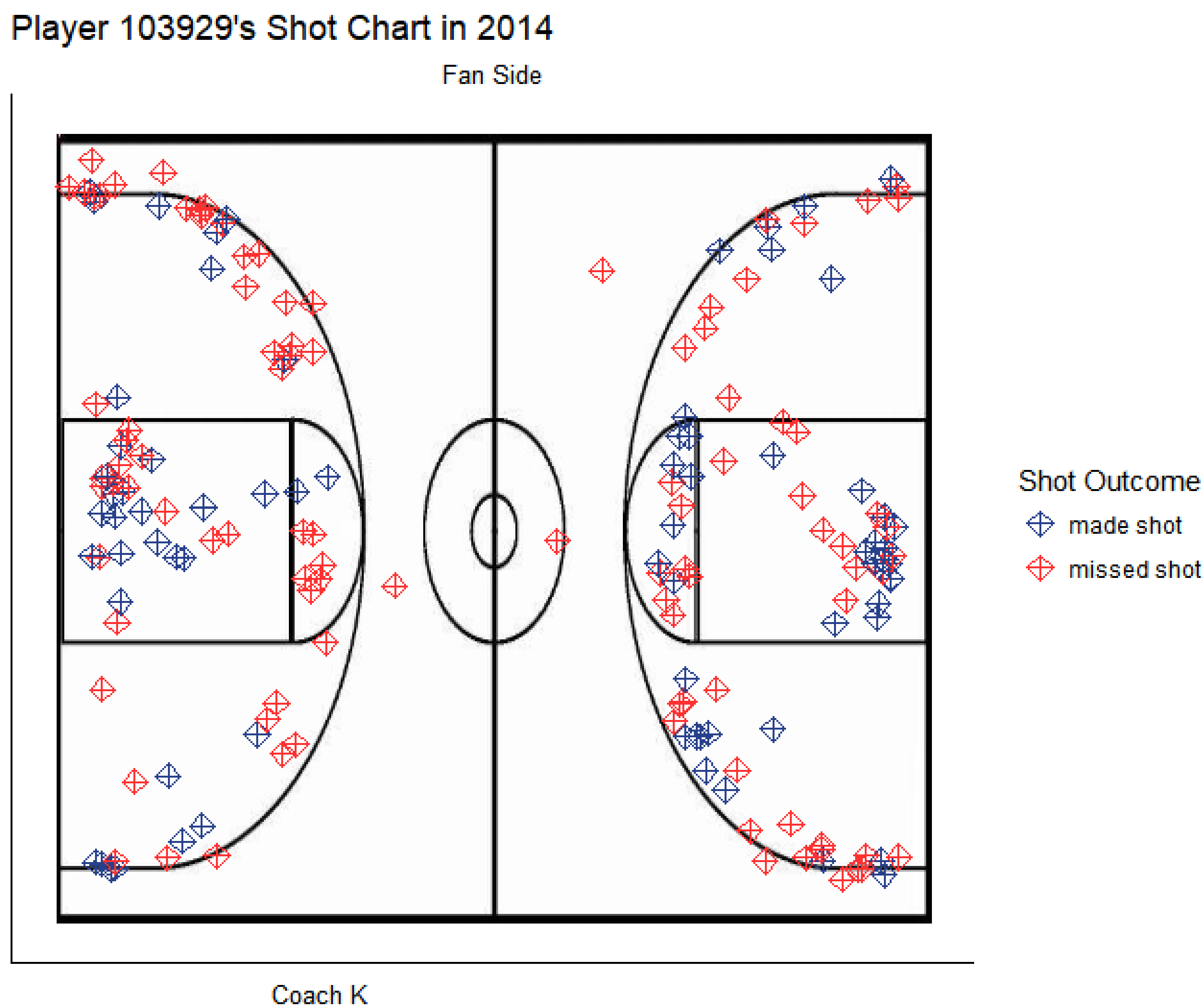


Figure 5.2

In 2014, however, it can be noted that Player 103929 has transitioned to shots that are closer to the basket and minimized the amount of 3 point shot attempts. He brought his shot attempts closer inwards, which aligns with the trend that he is better at shooting when he is closer to the basket. Compared to 2013, he attacks more along the nail, which could be attributed to Player 103929's growing strength as an off the dribble jump shooter.

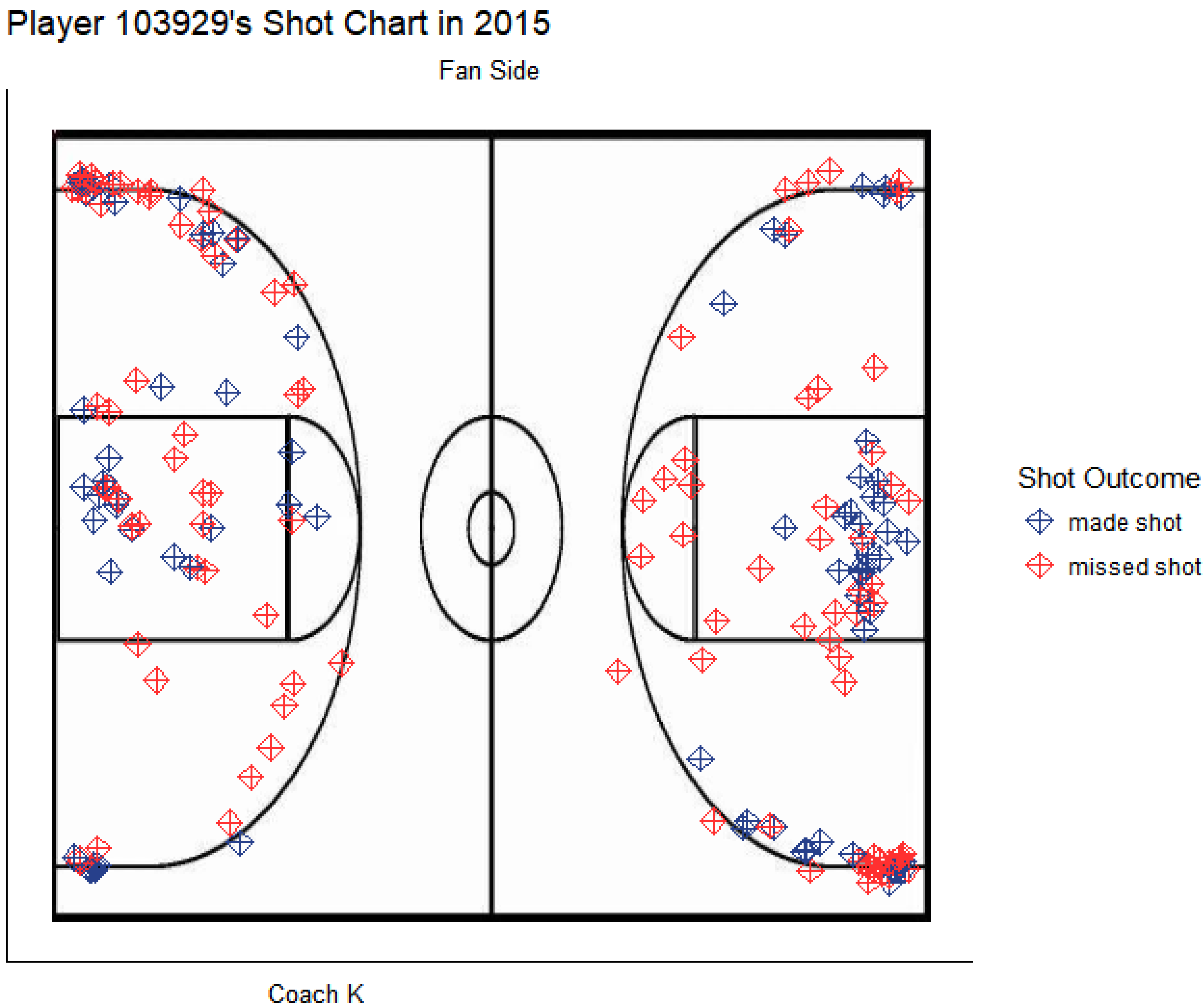


Figure 5.3

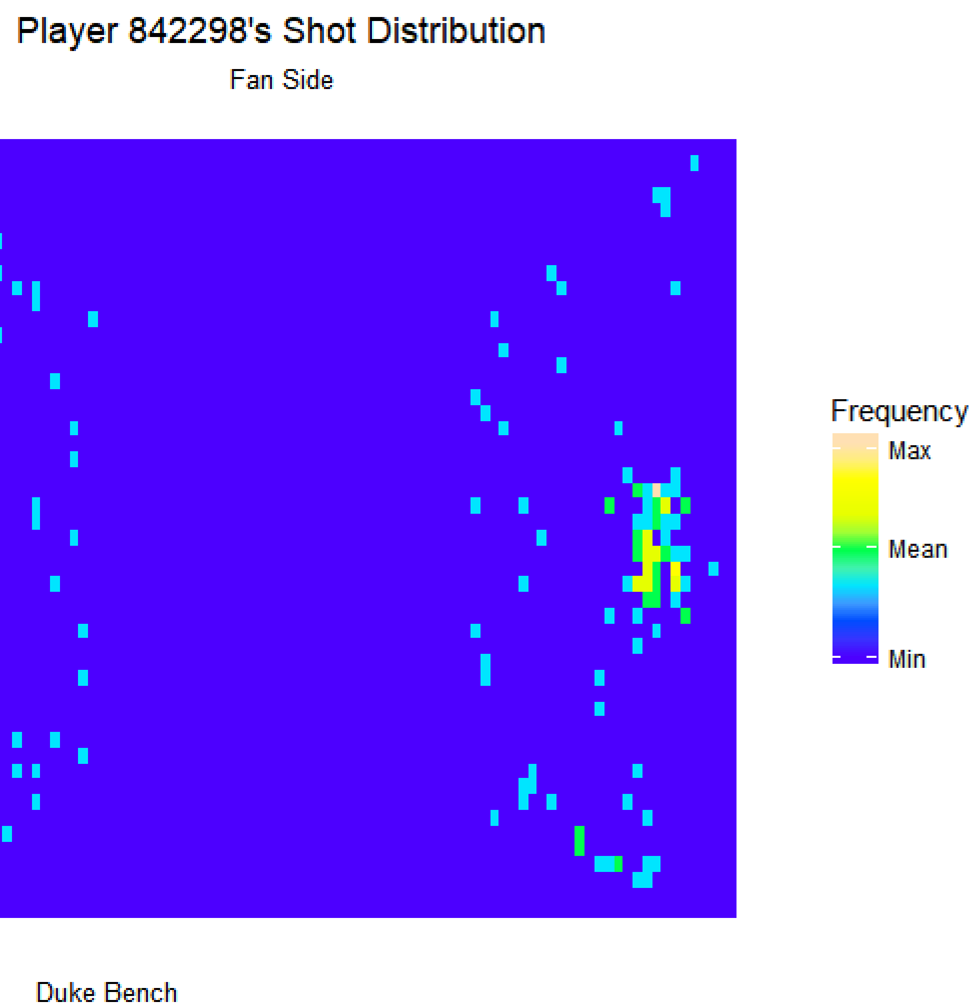
In the 2015 season, Player 103929 moves further out from the basket, attempting more 3s. His preference for shooting in the right wing is more pronounced. A new trend apparent from the graph, however, shows that Player 103929 shoots more corner 3s than the previous two years. While his shot attempts in the paint have slightly changed from 2013, Player 103929 definitely has a unique playing style that has overall been consistent in that he avoids shooting in the extended elbows and short corners.

5.2 Biases in Shot Location

While looking at the shot chart of each player shows their shooting preferences, putting their shot chart in the context of Cameron is another important aspect to note when analyzing a player's shot preferences. Cameron Indoor Stadium's student section, known as the Cameron Crazies, has been ranked as one of the best student sections in the country by Bleacher Report, For The Win, and FOX Sports (to name a few).

Furthermore, during the first half, a team's offense is on the opponent's side and a team's defense is on their home side. Thus, by acknowledging where a player shoots in context to the location of the fans and Coach K may reveal some biases to their shot location. Are players showboating for the Cameron Crazies or are they showboating for Coach K? To assess this trend, multiple Duke players were screened to note any possible trends in shooting habits.

Intuitively, a player's shot chart distributon should be an even reflection of the other half of the court (ie. if half court was inflected onto the other half court, the shot distributions should be similar). From Player 103929's Shot Charts, this intuition is true; it is clear that he prefers shooting from the left side on both sides–indicating that there does not exist an obvious bias in his shot location based on exterior factors. However, when looking at a Player 842298, his shot attempts are more prevalent on Duke's side of the bench, and less present on its complementary side. Perhaps, Player 842298 is showboating for his teammates or Coach K, and plays off of the exterior factors in a game. Further analysis will be conducted in later iterations of this paper to better understand this bias.



Chapter 6

Passing Networks

The main motivation behind this project is to understand the passing structure of Duke players in a game to create a better metric to evaluate players in the game of basketball. For this, each game was decomposed into individual possessions. Players who are in possession of the ball during each of the possessions are identified as vertices, and their passes to other players are edges in a pass network. Each vertex contains attributes about the player (e.g. fouls in the game), and each edge contains attributes about the pass (ie. distance passed).

6.1 A Breakdown of a Passing Network

6.1.1 Game Network

Below is an example of a passing network for an entire game, where each number represents the unique id of a player.

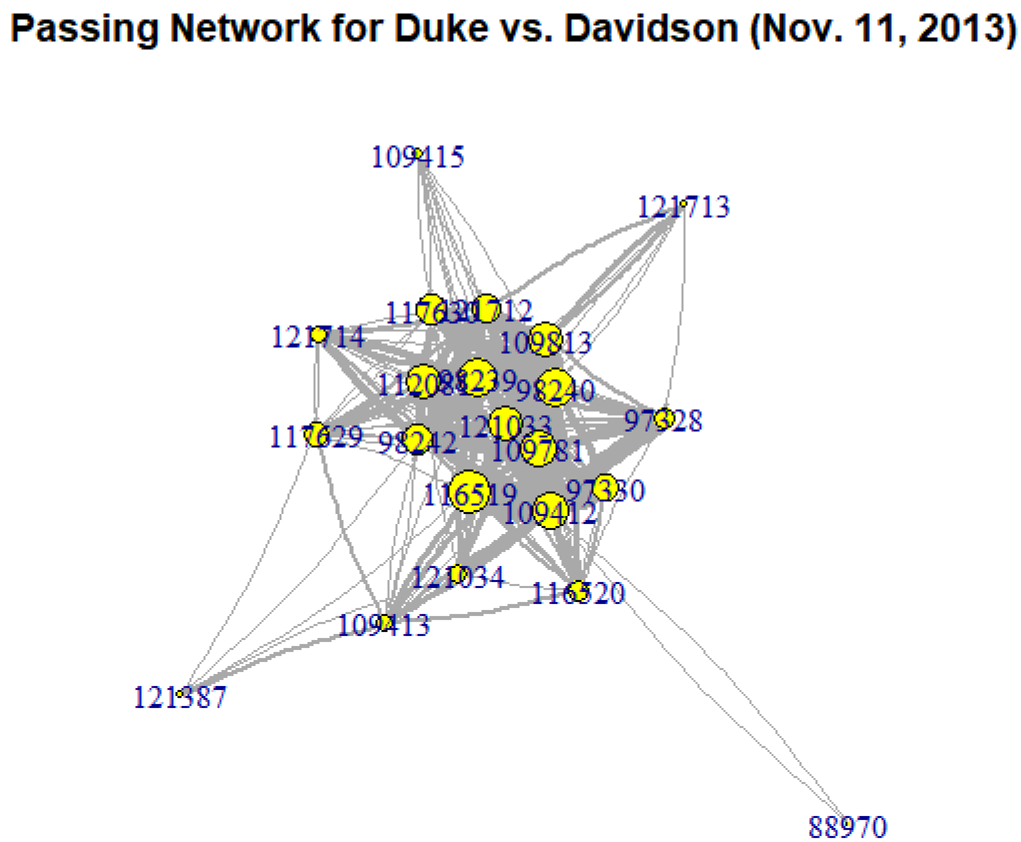


Figure 6.1

6.1.2 Possession Network

Breaking it down into a single game possession, the network becomes reduced to a smaller network. One challenge in identifying a posession was the inconsistency of the dataset's shot clock. For this, a new *possession* for this paper is defined as the moment when a team turns over the ball to the other team. For this, a possession may contain more than five players if players sub in/out within a possession.

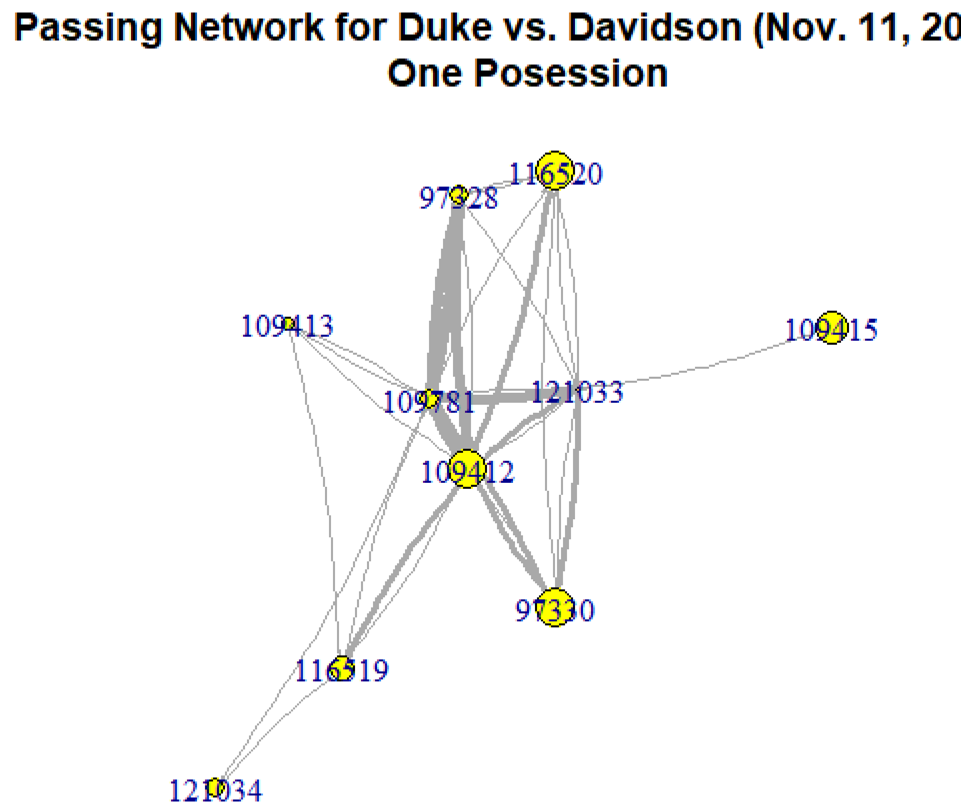


Figure 6.2

6.1.3 A Vertex and an Edge

A single pass between player 109412 and 109413 has a thin line because it only occurred once during this game. The arrow indicates the direction of the pass, and when checking the edge attribute between these two vertices, the distance of the pass between 109412 and 109413 is 22.83 units. Looking at vertex attributes, player 109412's position is a guard.

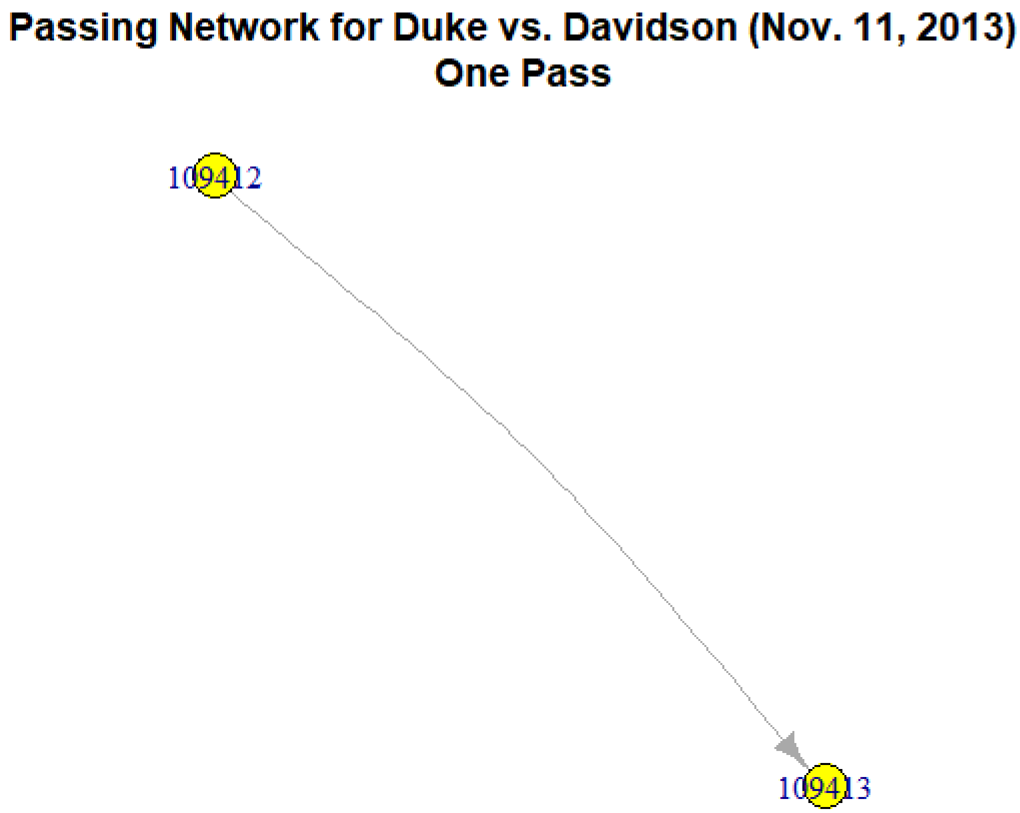


Figure 6.3

6.2 Initial Analysis of Passing Networks

Simply looking at a graph can reveal important characteristics about a player's role within a team. On a possession level, if a player receives many passes (as noted by a thicker edge), then he has a more central role on the team, and his teammates clearly rely on him to make good passs.

Other interesting network calculations are betweeness centrality; this metric can be visualized by the passing network, and noted as the popularity of a player based on how connected/central he is to the play. For this, returning to the Duke vs. Davidson game, we can note that Player 109415 is an important and valuable player for Duke because his betweenness centrality score is the highest score as denoted by the table below:

109412	109781	121033	116519	116520	109413	97330	97328	121034	109415
21.7	21.5	14.5	12.5	7.5	5.5	2	1.8	0.5	0

Furthermore, we can presume that players who are most connected to the ball should be able to best handle the ball. For this, we expect the players with the highest betweeness score to be the starters for Duke's 2013-2014 Men's Basketball team. Checking the starting line-up from Duke Men's Basketball for the 2013-2014 season, the betweenness score correctly matches Coach K's starting line-up.

Chapter 7

Network Modeling

7.1 Posession Analysis

Each possession in basketball typically ends in a made or missed shot, turnover, offensive or defensive rebound, or foul. Each possession has network characteristics unique to the play–number of triangles, passing reciprocity, betweenness centrality, etc. Using possession-level network characteristics to predict the outcome of a play can shed light on the utility value of certain team characteristics. If high centrality is a significant predictor of successful shots, then having a superstar player is the better playing style for basketball.

7.2 Multinomial Model

A multinomial logistic regression was initially fit to determine what features were important to predicting the outcome of a possession. Three categories were created for the outcome of a posession: good outcomes (made shot, offensive rebound), bad outcomes (missed shot, turnover), and neutral outcomes (inbound ball). Predictors included network characteristics like number of triangles, passing reciprocity, and betweenness centrality. Below is the formula:

$$y_{outcome,i} \sim \beta_{tri} x_{tri,i} + \beta_{recip} x_{recip,i} + \beta_{betweenness} x_{betweenness,i} + \epsilon_i$$

for posession i

7.3 Results

Network characteristics, although informative in summarizing possessions, were not significant predictors of the outcome of a basketball possession. When predicting the full dataset with the model, it correctly predicted the true outcome of a possesion 40.9% of the time. Higher reciprocity and triangle dependence, indicators of collaborative teamwork, had a direct relationship with good outcomes. These results from the preliminary model are largely directional, affirming the notion that network characteristics are not informative enough to capture the game of basketball. While previous papers have used exploratory network analysis to explain playing dynamics or to model the success of a player (e.g. "Basketball Teams as Strategic Networks"), this model shows that there exists limited value in solely relying on network characteristics.

7.4 Additive and Multiplicative Effects Network (AMEN) Analysis

The need to capture the game of basketball more robustly led to the final model, an additive and multiplicative effects network (AMEN) model. The AMEN package implements a latent eigenvalue model through a Monte Carlo Markov Chain. Hoff explains his preference for utilizing a latent eigenvalue model in "Modeling Homophily and Stochastic Equivalence in Symmetric Relational Data." This model captures the game of basketball more robustly than a multinomial model because it doubly captures network and nodal attributes. The output of this model provides posterior means of the row, column, multiplicative row, and multiplicative column effects for each player. Overall game performance for each player can be used as a response against the output of the latent eigenvalue model in order to check if a player's network and nodal attributes are significant influencers; points per game for each player was used as a response for a Poisson regression that used the posterior means of the latent eigenvalue model as its features.

$$y_{ij} = \beta_d x_d + r_i + s_j + u_i^T v_j + \epsilon_i$$

where

$$r_i = \beta_i x_i + a_i$$

and

$$s_j = \beta_j x_j + b_j$$

The model captures the network structure of each pass possession by transforming the passing networks into adjacency matrices. Row and column nodal covariates are px1xn vectors indicating the contribution of each player p for each possession n. The dyadic features are pxpxnxi arrays that capture the shared features between players.

Nodal Attributes (Response in the Poisson Regression): Points per game for each player

Nodal Features ($\beta_i x_i, \beta_j x_j$): Was in previous player (0/1), currently in possession (0/1)

Dyadic Features ($\beta_d x_d$): Shared position, shared height, shared weight, shared class

Network Attributes (y_{ij}): Passing network matrix

7.5 Structural Zeros

Currently, the AMEN package takes on a pxp matrix of players to model the passing relationship between players. Currently, the model uses data from the 2014-2015 season, so p = 10. These 10 players represent the five players on the court, and the five players on the bench. For the five players on the court, if there does not exist a pass between two players, then a zero populates the matrix to account for the nonevent. However, if a player is on the bench, he similarly cannot receive a pass, so all bench players will always have a zero populated in the 10x10 matrix. This creates a challenge in modeling the data because the zeros in the matrix represent two different events–players who had the possibility to receive the ball but did not and players who never had the chance to receive the ball.

The model includes a binary column feature that signals a player's status (on court or on bench). While it does not solve the structural zero problem entirely, this feature accounts for the differences between active and nonactive players.

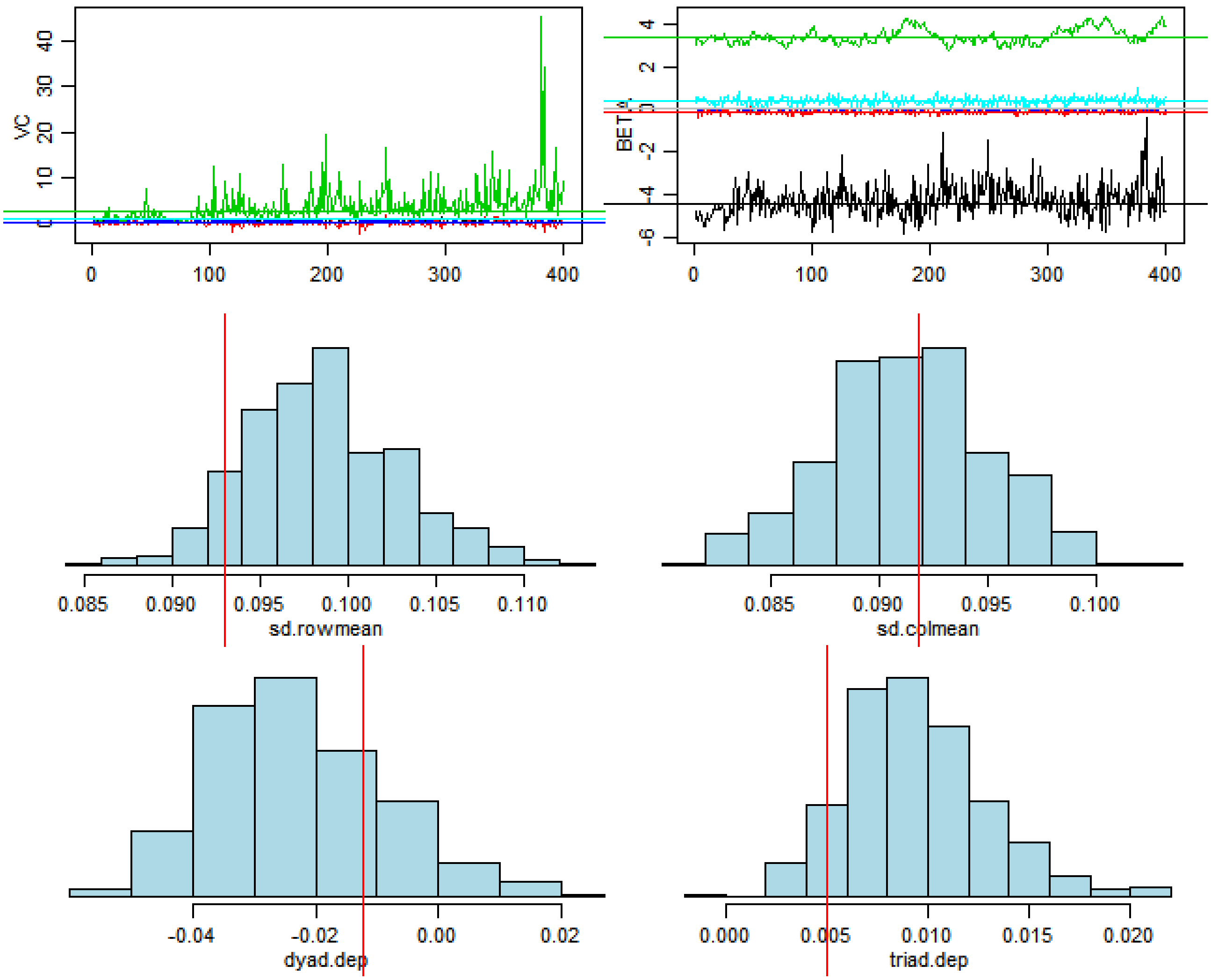
7.6 Model Fit

The model was fit at two stages: latent eigenvalue model and poisson regression.

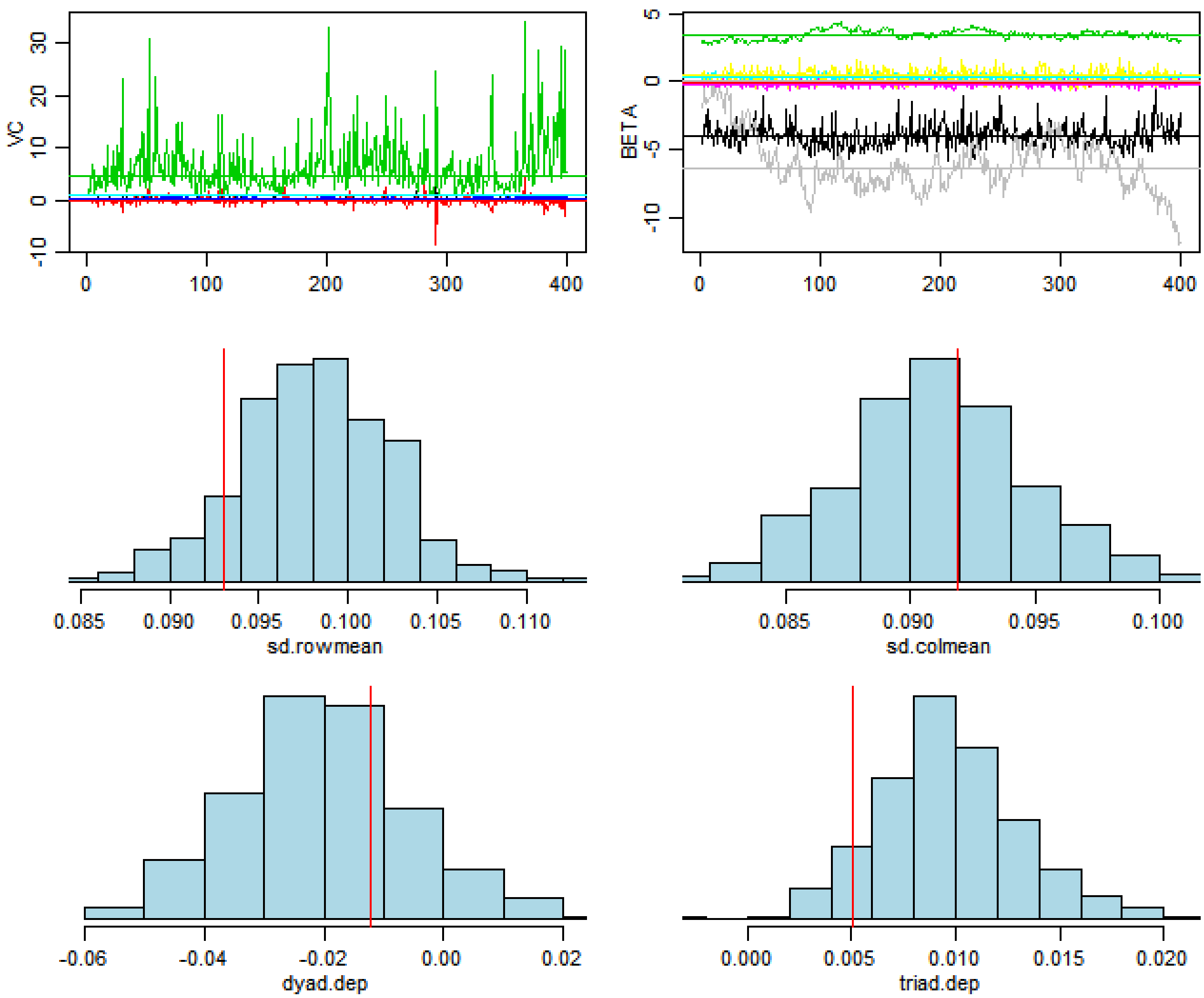
7.6.1 AMEN Fit

Checking the model fit of the AMEN output, the posterior predictive checks perform well for the column effects. A challenge with this model, as noted at the end of Fosdick's and Hoff's "Testing and Modeling Dependencies Between a Network and Nodal Attributes," was determining the appropriate dimensionality for transforming the row and column effects onto a latent space. Initially, the data was fit with a dimensionality of 2. However, as there are always five players on the court, higher dimensionality would be more appropriate for this model. The data was fit again with a dimensionality of four, and there was not a significant improvement in the fit of the posterior predictive checks. This suggests that a dimensionality of 2 is enough to capture the data.

The output below shows the model fit of a game with a dimensionality of 2.



The output below shows the model fit of a game with a dimensionality of 4.



7.6.2 Poisson Regression Fit

The output of the AMEN model was plugged into a Poisson regression, with the response as a player's points per that game. Assessing the fit of the Poisson model, it assumes hetereoscedascity, and fits the checks well. However, the p-value for the deviance goodness of fit test was $9.152 * 10^{-25}$, which indicates with strong evidence that the model fits the data poorly. Assessing the fit of the model with regards to predicting the true points per game, the poisson model had a mean squared error of 73.08. The mean difference between the predicted points per game and true points per game was 5.52–approximately two to three possessions in the context of basketball. Regardless, this model provides a directional indication that row and column effects (passing and receiving) are significant features in predicting the productivity of a player, as denoted by the summary of the model.

Below is the summary output of the Poisson regression's fit and coefficient estimates for each β.

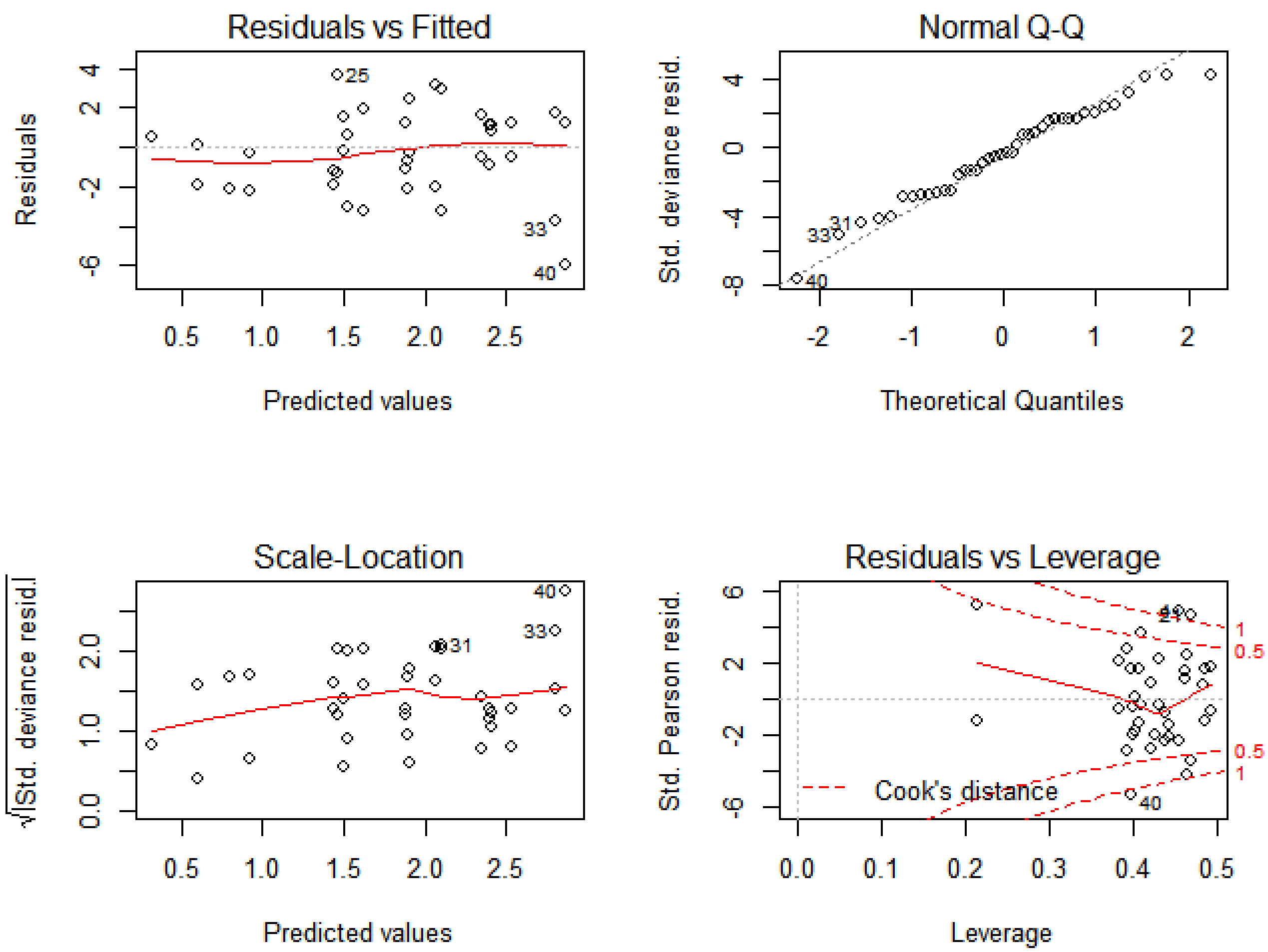


Figure 7.1

```
Call:
glm(formula = points ~ ., family = "poisson", data = data.frame(pred.dat[,
    -c(1, 3)]))

Deviance Residuals:
    Min       1Q    Median       3Q      Max
-5.9340  -1.9315  -0.3091   1.1841   3.6648

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)   2.0536     0.2068   9.929  < 2e-16 ***
apm          -2.5614     1.3698  -1.870 0.061490 .
bpm           0.3579     0.1108   3.229 0.001242 **
```

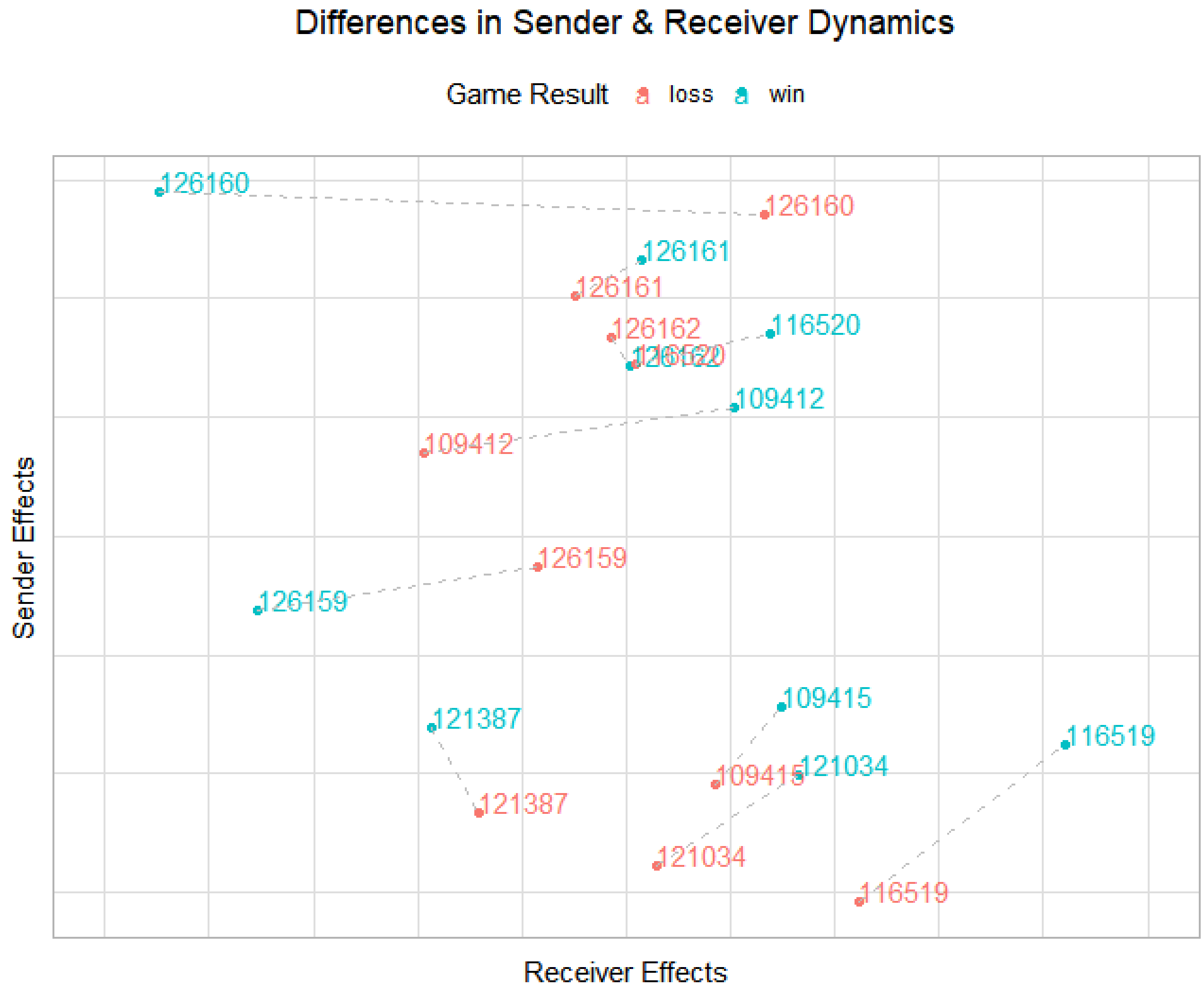
```
u1           -5.5429     6.0272  -0.920 0.357761
u2            5.1948     1.3746   3.779 0.000157 ***
v1            0.8560     0.8369   1.023 0.306395
v2           -0.2401     0.5166  -0.465 0.642043
uvpm1         0.3843     3.7931   0.101 0.919292
uvpm2         6.9119     7.2313   0.956 0.339157
uvpm3         3.3323     2.8435   1.172 0.241244
uvpm4        -8.4270    11.7979  -0.714 0.475055
uvpm5         7.5922     7.5001   1.012 0.311403
uvpm6        11.1989     5.8842   1.903 0.057012 .
uvpm7        -1.5104     3.2222  -0.469 0.639261
uvpm8        11.6098     6.2778   1.849 0.064409 .
uvpm9        -3.6840    10.0160  -0.368 0.713017
uvpm10        5.0079     4.3871   1.141 0.253662
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 282.5  on 39  degrees of freedom
Residual deviance: 171.9  on 23  degrees of freedom
AIC: 327.64

Number of Fisher Scoring iterations: 7
```

According to the coefficient estimates, a higher posterior mean for the additive row effect (apm-passing) leads to lower points per game. On the other hand, a higher posterior mean for the additive column effect (bpm-receiving) leads to more points per game. The second dimension of the posterior mean for multiplicative row effects (u2) leads to more points per game. This model summary reaffirms the significance of passing to the success of a team.

7.7 Results

The latent eigenvalue model with less dyadic features (only shared position) performed comparably to the full model based on the posterior predictive checks. The output of the latent eigenvalue model was used to predict a player's points per game via a Poisson Regression. While both the row and column posterior means were significant influencers of a player's points per game, row effects had a larger coefficient estimate. These results confirm the importance of passing and teamwork for successful plays.

Exploratory analysis of the model output reveals that there are game-level differences in play style for certain players. Comparing the performance of two games with two different outcomes for the Duke 2014-2015 team, there a noticeable differences in the sender and receiver effects. Player 126160, for instance, had higher receiver effects in a loss compared to a win. This result could imply that player 126160 was not fulfilling his role on the team if his performance recorded by a successful game was his baseline. On the other hand, some players did not drastically change betweeen a a win and a loss. Players like Player 109415 who typically had low playing time and thus lower sender and receiver effects overall, intuitively had even lower sender and receiver effects during a loss. The loss of a team, although influenced by many characteristics, can be partially attributed to these differences in individual player performance. Regardless, almost every player had a higher sender (passing) effect in a win compared to a loss, indicating the importance of passing. However, this model does not take into account sufficient statistics for the varying defense that play against Duke. A future indicator to control for the quality of Duke's opposing teams would be one way to solve this challenge.

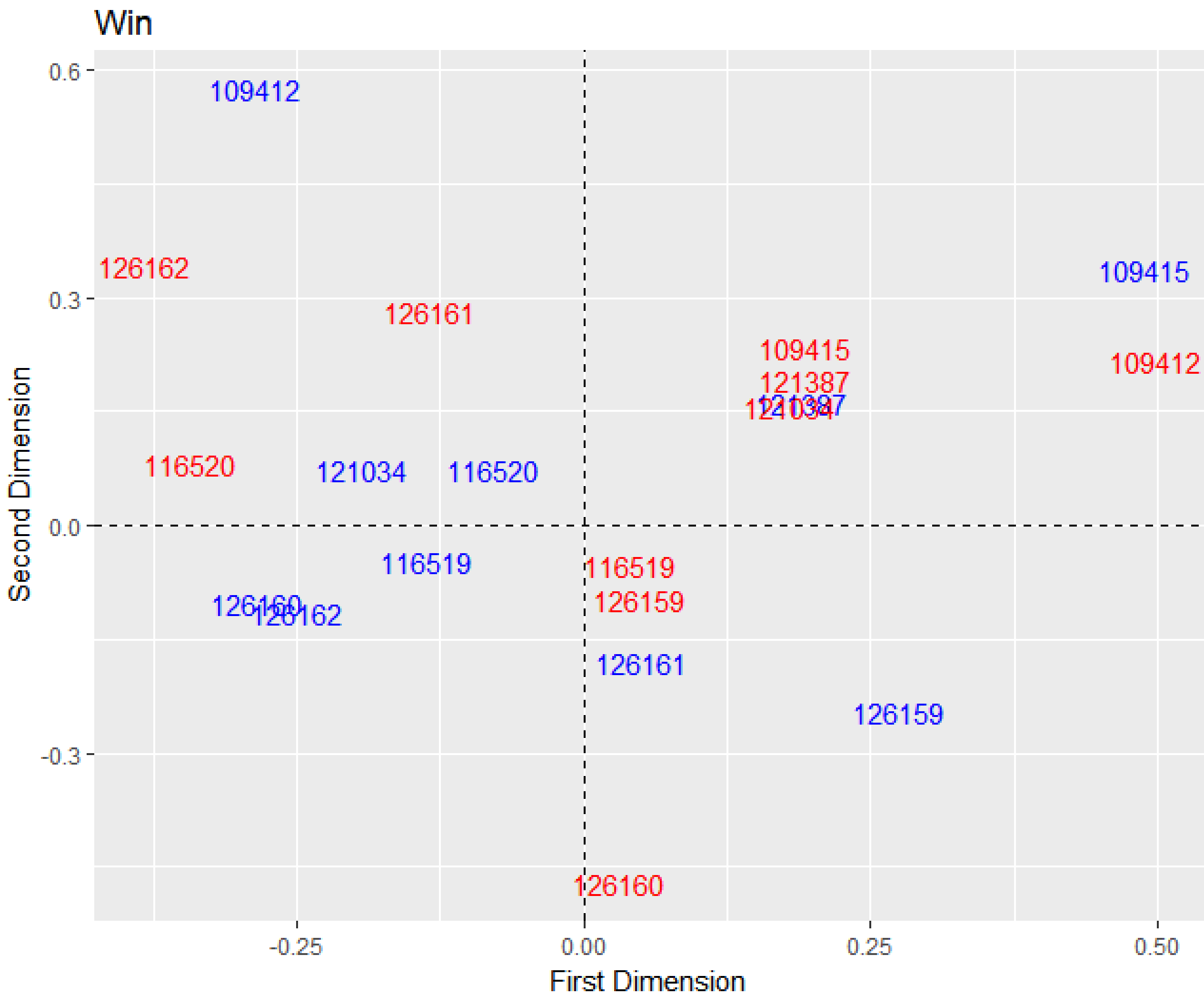


Looking at the posterior means for the multiplicative row (blue) and column (red) effects for a loss, players who are in the same quadrant and differing colors are more likely to interact. Players that lie on the origin are expected to be the point guard because of their neutrality in passing and receiving the ball from other players. However, the two players who lie on the origin were not the point guards for Duke.

Player 126162 (in the red) is a post player, and Player 126159 is a guard. More players lie in the lower right quadrant for a loss compared to a win, especially for the receivers (red).

Figure 7.2

Looking at the plot of the posterior means of the multiplicative row (blue) and column (red) effects for a win, the players are more evenly distributed. Again, the players closest to the origin were not the point guards for Duke. The players with the most playing time were significantly closer to the origin than the players on the edge of the plot.


Win
Second Dimension
First Dimension
0.6
0.3
0.0
-0.3
-0.25
0.00
0.25
0.50
109412
126162
126161
109415
109415
121387
121034
121387
109412
116520
121034
116520
116519
116519
126159
126160
126162
126161
126159
126160

Chapter 8

Conclusion

Duke Men's Basketball has a vast and rich dataset that has much to be explored. Of particular interest is how a player interacts against his teammates and defenders. This paper focuses on modeling player interactions via passing networks–network centrality and betweenness scores identify key players within a team. By evaluating passing networks, not only can a player's value within a team be deduced, but also how a player's value within a team has changed over time. Modeling each posession in a game with network characteristics as features can be directionally useful. A more robust approach utilizes Peter Hoff's AMEN package, which models both nodal and network characteristics. The results through this approach similarly show the significance of passing and receiving the ball. Teamwork and high collaboration leads to successful plays.

8.1 Future Steps

The scope of this work captures possessions of a game on an individual level. However, using the output of the latent eigenvalue model to predict nodal attributes is not sufficient to capture the game fully. Currently, an implementation and adaptation of a model influenced by Luke Bornn's "A Multiresolution Stochastic Process Model for Predicting Basketball" aims to capture the game of basketball more robustly. Future work will use this advanced model to create metrics for assessing a player's production value on a team level. A summary of the model replication can be found in Appendix A.

Chapter 9

Appendix A

9.1 Model Replication

The initial approach to understand how to best capture passing networks sought to replicate Daniel Cervone, Alex, D'Amour, Luke Bornn, and Kirk Goldsberry's paper,"A Multiresolution Stochastic Process Model for Predicting Basketball Possession Outcomes." They attempt to capture the game wholelistically via a new measure called Expected Possession Value (EPV). This new metric uses three models–a Microtransition Model, Macrotransition Entrance Model, and a Macrotransition Exit Model–to capture the spatial biases of each player and the in-game effects of pressure, so that it can measure the likelihood of a successful play (made shot) given the previous sequence of events. To compare players against the league-average scores, they also calculated Expected Possession Value -Adjusted as an application for teams. Below is a brief overview of each model.

This paper is particularly interesting because EPV utilizes the spatio-temporal elements of the game, so it models the NBA game dynamically. Given Duke Basketball data, the motivation is to replicate "A Multiresolution Stochastic Process Model for Predicting Basketball Possession Outcomes," to better understand the Duke Men's team, as well as to compare professional basketball to collegiate basketball individual and team playing styles. Below is a brief overview of each model used in the paper to calculate EPV.

9.1.1 Microtransition Model

$x^l(t+\epsilon) = x^l(t) + \alpha_x^l[x^l(t) - x^l(t-\epsilon)] + \eta_x^l(t)$ where $\eta_x^l(t) \sim N(\mu_x^l(z^l(t)), (\sigma_x^l)^2)$

The microtransition model models the defensive conditions of the game based on the (x, y) coordinates of a player and their acceleration effects $(\alpha_x^l(t))$. It is also assumed that a player's spatial location is normally distributed. Since players play differently, each microtransition model is specifically fitted to the player.

9.1.2 Macrotransition Entrance Model

$P(M(t)|F_t^{(Z)}$ The macrotransition entrance model predicts whether the next move will be a pass (4 options), shot attempt, or turnover. The model is disjoint.

9.1.3 Macrotransition Exit Model

$P(C_{\delta_t}|M(t), F_t^{(Z)})$ Given the Macrotransition Entrance Model predicts a shot attempt, it indexes to a logistic regression model to calculate player l's successful shot probability. Given the Macrotransition Entrance Model predicts a pass,it indexes to a model that predicts where the pass will take place. Otherwise, a turnover is assumed.

9.1.4 Fall Backs on the Implementation of this Model

Currently, the implementation of the model has yet to be completed due to setbacks of incompatible R code. The implementation of this paper is currently still in progress.

9.1.5 Proposal

Regardless, we hypothesize that since both metrics are calculated via a semi-Markov process, EPV fails to capture the full nature of the possession because it only uses the last posession as a prior. The model would be more robust if it captured the entirety of the posession in its prior–however, the computational time of such an ordeal would prevent any real-time analyses. Thus, this paper proposes that a simpler model may perform more quickly and potentially just as robustly to allow for game-time analyses.

References

Bekkers, J., & Dabadghao, S. (2017). "Flow Motifs in Soccer: What can passing behavior tell us? Sloan Sports Analytics Conference. Retrieved from http://www.sloansportsconference.com/wp-content/uploads/2017/02/1563.pdf

Cervone, D., D'Amour, A., Bornn, L., Goldsberry, K. (2016). "A Multiresolution Stochastic Process Model for Predicting Basketball Possession outcomes." Retrieved from https://arxiv.org/pdf/1408.0777.pdf

Fewell, J., Ambruster, D., Ingraham, J., Petersen, A., & Waters, J. (2012). "Basketball Teams as Strategic Networks." PLOS. Retrieved from http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0047445

Fosdick, B.K., Hoff, P.D. (2013). "Testing and Modeling Dependencies Between a Network and Nodal Attributes." Retrieved from https://arxiv.org/abs/1306.4708

Goncalves, B., Coutinho, D., Santos, S., Lago-Penas, C., Jimenez, S., & Sampaio, J. (2017). "Exploring Team Passing Networks and Player Movement Dynamics in Youth Association Football." PLOS. Retrieved from https://doi.org/10.1371/journal.pone.0171156.

Gudmundsson, J., & Horton, M. (2016). "Spatio-Temporal Analysis of Team Sports – A Survey." Retrieved from https://arxiv.org/abs/1602.06994

Hoff, P.D. (2007). "Modeling Homophily and Stochastic Equivalence in Symmetric Relational Data." Retrieved from https://arxiv.org/abs/0711.1146.

Hoff, P.D. (2003). "Bilinear Mixed Effects Models for Dyadic Data." Retrieved from https://www.stat.washington.edu/~pdhoff/Preprints/dyadic.pdf.