9TH ANNUAL
MIT SLOAN
SPORTS ANALYTICS CONFERENCE
FEBRUARY 27 - 28, 2015 BOSTON CONVENTION AND EXHIBITION CENTER

Graphical Model for Baskeball Match Simulation

Min-hwan Oh, Suraj Keshri, Garud Iyengar
Columbia University
New York, NY, USA, 10027
mo2499@columbia.edu, skk2142@columbia.edu
gaurd@ieor.columbia.edu

Abstract

Conventional approaches to simulate matches have ignored that in basketball the dynamics of ball movement is very sensitive to the lineups on the court and unique identities of players on both offense and defense sides. In this paper, we propose the simulation infrastructure that can bridge the gap between player identity and team level network. We model the progression of a basketball match using a probabilistic graphical model. We model every touch and event in a game as a sequence of transitions between discrete states. We treat the progression of a match as a graph, where each node is a network structure of players on the court, their actions, events, etc., and edges denote possible moves in the game flow. Our results show that either changes in the team lineup or changes in the opponent team lineup significantly affects the dynamics of a match progression. Evaluation on the match data for the 2013-14 NBA season suggests that the graphical model approach is appropriate for modeling a basketball match.

1 Introduction

Predicting the outcomes of professional sports events is one of the most popular practices in the sports media, fan communities and, of course, sport betting related industries. Predictions range from human prediction to statistical analysis of historical data. In recent years, basketball, specifically the NBA, has received much attention as a domain of analytics with the advent of player tracking data. Many new metrics have been introduced to evaluate players and teams. However, there have not been studies that fully take advantage of the rich player tracking data to simulate the outcomes of basketball matches. Most of the previous simulation approaches in basketball have focused on win-loss predictions, ignoring the progression of matches. In efforts to obtain detailed "microsimulation" of a basketball game, Shirley [1] and Štrumbelj and Vračar [2] used a possession-based Markov model to model the progression of a basketball match. However, they treat each team as merely a single entity rather than collective union of individual players. These previous studies ignore that in basketball the dynamics of ball movement is very sensitive to the lineups on the court and unique identities of players on both offense and defense sides.

One can ask, "Is Miami Heat the same team without LeBron James? Or, can Oklahoma City Thunder be an elite team without Kevin Durant and Russell Westbrook?" Taking individual players into account in a simulation process is not just about addressing the issues with trades or changes in the roster in the preseason but also changes in lineups of teams during a season, which appear almost on a day-to-day basis, either in a starting lineup or bench lineup — whether it is due to injuries or strategic reasons. These changes do have an impact on final game results.

Fewell et al.[3] used network analysis in which they analyzed ball movement of teams, mapping game progression pass by pass. They assessed differences in team's offensive strategy by their network properties. While their objective was not to simulate matches, they still did not address the unique identities of players, which is prevalent especially in basketball. Questions such as "With Tim Duncan and Tony Parker out tonight, will the Spurs win against the Rockets?" still remain unanswered.

In this paper, we propose the simulation infrastructure that can bridge the gap between player identity and team level network. We model the progression of a basketball match using a probabilistic graphical model. The model shows the ball movement of every play and subsequent game events based on player level pass interaction, shot frequency given teammates and defenders, shot accuracy against the defense, rebound etc. We follow the natural and intuitive flow of a basketball match as shown in Figure 1.

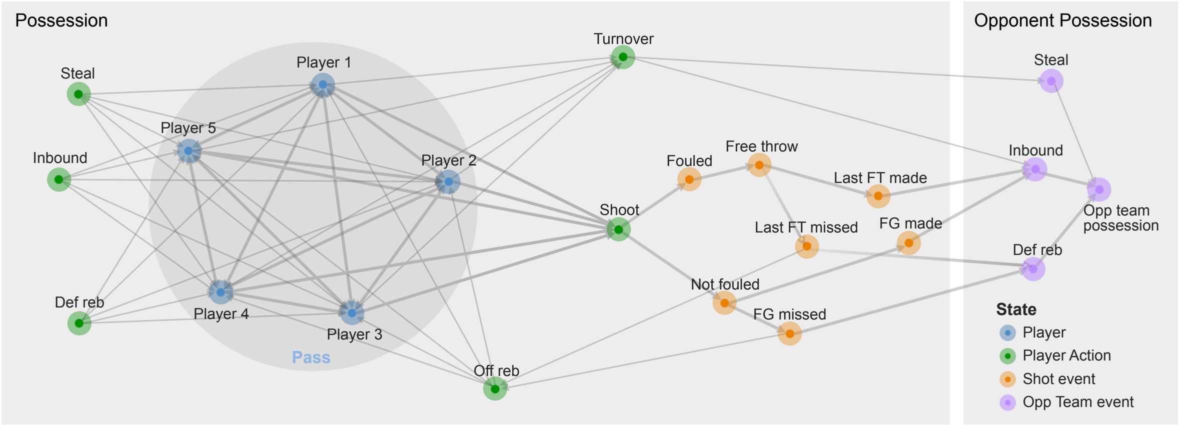


Figure 1: Graphical Model for sequence of events in each possession

2 Method

Our model is calibrated using the player tracking data and play-by-play game log data from the matches for the 2013-2014 season. Both these data sets are available on NBA.com. We also used the lineup data available on Basketball-Reference.com. We model every touch and event in a game as a sequence of transitions between discrete states. We treat the progression of a match as a graph, where each node is a network structure of players on the court, their actions, events, etc., and edges denote possible moves in the game flow. We learn the conditional probability of the edges from the data. We simulate ball movements between players, how likely a player is to take a shot, and how defense and teammates affect the dynamics.

Table 1: Summary of Notation

Notation	Meaning	Section
L_o	Offensive team lineup	2.2, 2.4, 2.6
L_d	Defensive team lineup	2.2, 2.6
γ_i	Player i's propensity to take a shot	2.2
$\widetilde{\gamma}_i$	Player i's propensity to take a shot given the defensive lineup	2.2
β_i	Player i's ability to deter shot attempt	2.2
α_{ij}	Tendency of player i to pass to player j	2.4
θ_{id}	Shooting ability of player i at basis d	2.3
ϕ_{id}	Defensive ability of player i to reduce shot accuracy at basis d	2.3
ψ_{id}	Player i's ability to draw a shooting foul at basis d	2.5
ζ_{id}	Player j's foul proneness at basis d	2.5
ρ_i^d	Defensive rebound grabbing ability of player i	2.6
ρ_i^o	Offensive rebound grabbing ability of player i	2.6
τ_a	Average possession time of team a	2.7

2.1 Start of Possession

We model the start of a possession by a team as a multinomial distribution between players on the court. If the possession starts with an inbound pass, we sample the starting player according to distribution of historical backcourt touch data. On the other hand, if the possession starts with a defensive rebound or a steal, then it is trivial since the player who starts the possession has been already decided. Methods to compute rebound probabilities and to sample a steal event are discussed in later sections.

2.2 Shot Frequency

We model the probability of a field goal attempt for a given touch as a Bernoulli distribution with probability

$$p(S_i=1\mid L_o,L_d,\gamma,\beta)=\sigma\left(\widetilde{\gamma}_i+\left(\widetilde{\gamma}_i-\frac{1}{4}\sum_{k\in L_o,k\neq i}\widetilde{\gamma}_k\right)\right)$$
$$\text{where}\quad \widetilde{\gamma}_i=\gamma_i-\sum_{j\in L_d}w_{ij}\beta_j$$

with $\sigma(x)=\exp(x)/(1+\exp(x))$. S_i is an indicator for whether player i attempts a shot given a touch. L_o and L_d represent the lineups of the offensive team and defensive team respectively. γ_i is a parameter which determines how likely a player is to take a shot and β_j is the defensive ability of player j to reduce shot frequency. The weight w_{ij} determines how much player i is affected by the defense of player j which is proportional to the time that player i is guarded by player j[1]. Franks et al.[4] models shot frequency without taking the teammates in account. However, in our model, we also have an offset term $\left(\widetilde{\gamma}_i-\frac{1}{4}\sum_{k\in L_o,k\neq i}\widetilde{\gamma}_k\right)$ which is negative (or positive) if the propensity of player j to take a shot is less (or more) than average teammates' propensity. The reasoning behind this model is that an event of a player taking a shot depends not only on his propensity to shoot and his defender but also on the propensity of his teammates. We can take Kevin Durant and Russell Westbrook of the Oklahoma City Thunder for an example: when Kevin Durant is not on the court, Russell Westbrook tends to shoot more.

2.3 Shot Efficiency

To model shot efficiency of a player, our approach is similar to Franks et al.[4] Given that a player attempts a field goal, we model shot efficiency (the probability that the player makes a shot) as a function of the offensive player's skill, the defender at the time of the shot, and the location of the shot on the court.

$$p(Y=1\mid d,\theta,\phi)=\sigma\left(\theta_{id}-\phi_{jd}\right)$$

Here, Y is an indicator for whether the shot was made, d represents the basis from which the shot was taken, θ_{id} is the shooting ability of player i at basis d, and ϕ_{jd} is the defensive ability of the closest defensive player j to reduce shot accuracy at basis d. Note that our model is slightly different from Franks et al.[4] In particular, we do not take into account the distance between the offensive and defensive player. The justification is that the ability of the defender is also characterized by how closely he defends the player. Thus, the parameter ϕ_{jd} takes into account how closely is player j able to defend basis d. This assumption is important in our simulation model because we are not modeling the distance between the defender and the shooter while the shooter attempts a shot. To each offensive player, we a priori assign the weighted average of defense depending on his position and the position of the defenders on the court. Thus, while simulating the game, once a player decides to take a shot, we sample the shot location using the basis loadings. Then, the success probability of the shot is given by the shot efficiency model.

2.4 Pass Network

We model the passes between players as a network with edge weight parameterized by α_{ij} $(i\neq j)$. The probability that player i passes to player j if player i chooses to pass is given by

$$p(i\to j\mid\alpha,L_o)=\frac{\alpha_{ij}}{\sum_{k\neq i,k\in L_o}\alpha_{ik}}$$

where we only take into account α's for players on the court. Note that the probability that player i passes to player j depends not only on players i and j but also other teammates on the court. To learn the α matrix, we use EM algorithm on the data of total number of passes between each player and the total number of possession each lineup had in every game. Figure 2 shows an example of the α matrix we learned for the San Antonio Spurs, with α_{ij} as the i,j entry of the matrix.

[1] For simulation purposes, we set the weight w_{ij} according to similarity of positions of player i and defensive player j. For example, if player i is a point guard, we give more weight to the point guard and shooting guard on the defensive team than to the opponent center

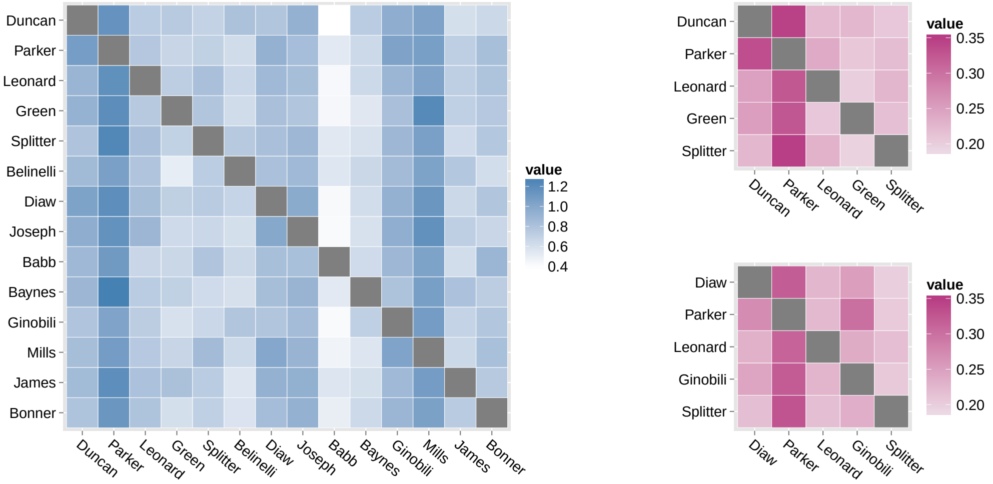


Figure 2: Rows are passers and columns are receivers. Note that the diagonal entries are set to zero. The α matrix for the San Antonio Spurs 2013-14 roster (left) shows that Tony Parker and Patty Mills are more likely to receive passes from most of the other players. This was expected due to their position and role as primary ball handler. We create a pass probability matrix for different lineups (right) by extracting a corresponding entry of the α matrix and normalized by each row. We observe that replacing two players in the lineup results in a different pass probability matrix. This allows us to obtain the passing distribution of any arbitrary lineup in a team.

2.5 Shooting Foul & Free Throw

We model shooting fouls as a function of the shooter's ability to draw a foul, defender's foul proneness, and the location (basis) of the shot on the court. The approach is similar to the model for shot efficiency.

$$p(SF(i,j)=1 \mid d,\psi,\zeta)=\sigma\left(\psi_{id}+\zeta_{jd}\right)$$

$SF(i,j)$ is an indicator for whether the player i was fouled by player j while shooting. ψ_{id} is player i's ability to draw a shooting foul at basis d, and ζ_{jd} represents the defender's foul proneness at basis d. Therefore, if a defender is more foul prone, then there is a higher chance of shooting foul. As for free throws, we use free throw percentage for each player to sample a free throw success event. This is a reasonable approach since a free throw does not depend on opponents or teammates.

2.6 Rebound

We model rebound as a competition between the players on court. Since there is a clear difference in effort required to grab a defensive and an offensive rebound, we assume that each player i has a defensive and an offensive rebound ability represented by ρ_i^d and ρ_i^o respectively. Given the current lineup of offensive and defensive team on the court, the probability of player i grabbing an defensive or an offensive rebound is given by

$$p(DR_i=1 \mid L_d,L_o)=\frac{\exp(\rho_i^d)}{\sum_{j\in L_d}\exp(\rho_j^d)+\sum_{k\in L_o}\exp(\rho_k^o)}$$
$$p(OR_i=1 \mid L_d,L_o)=\frac{\exp(\rho_i^o)}{\sum_{j\in L_d}\exp(\rho_j^d)+\sum_{k\in L_o}\exp(\rho_k^o)}$$

DR_i and OR_i are indicators for player i grabbing a defensive rebound and an offensive rebound respectively. In this model , the rebound grabbing ability of a team depends on the players of both the teams on court. This model allows us to estimate rebound grabbing probability for arbitrary lineups.

2.7 Number of Possessions

In our model, we assume that a possession starts with an inbound pass, a defensive rebound, or a steal. To model number of possessions, we assume that team i on an average takes time τ_i to end a possession. Thus, total number of possessions for each team in a game between team a and team b should be close to $\frac{T}{\tau_a+\tau_b}$, where T_i is duration of the game i. We have the data for total number of possessions in a game, η. To learn τ, we minimize the sum of square of error of each game:

$$\min_{\tau} \sum_i \left(\sum_k I_{ki} \tau_k - \frac{T_i}{\eta_i} \right)^2$$

where T_i is duration of game i, η_i is the number of possessions in game i, and $I_{ki} = 1$ if team k plays in game i.

2.8 Turnover

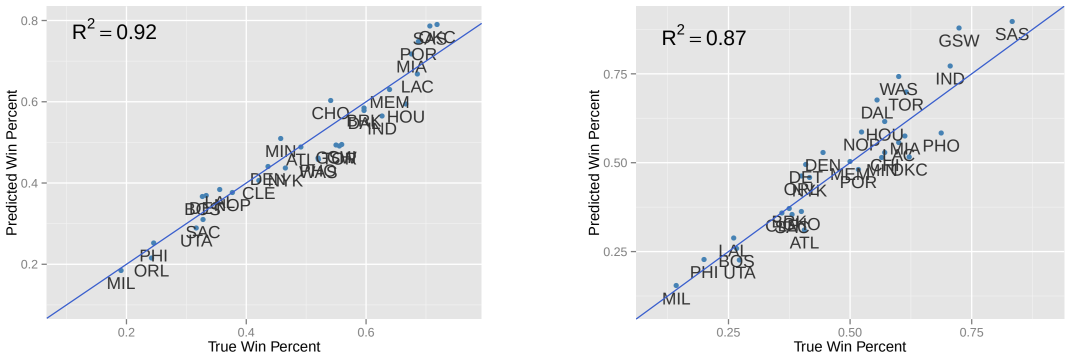
In our model, we assume that there are two types of turnover. One type is stolen balls and the other type includes all the other turnovers that results in an inbound pass (offensive foul, out-of-bounds, etc.). We calculate average probability of turnover per touch for each player from the historical data and use that independent of current lineup. We also sample a stolen ball event from turnover event. Given a stolen ball, we assign a steal to a defensive player with probability proportional to his average steal rate compared to average steal rate of his teammates on court. Given a non-stolen ball turnover, we start from an inbound pass.

2.9 Simulation

For simulation purposes, lineup is an input parameter to our model. One can try different lineups against an opponent team, modifying the number of possessions given to particular lineups. For fitting and testing our model, we use the actual lineups used in each game. After we learn all the required parameters mentioned above, we compute conditional probability for each edge in the graph of possession illustrated in Figure 1. We draw a sample of events using the graphical model for each possession. We repeat this sampling process until we reach the estimated number of possessions. This gives us one sample of single match statistics. We simulate a match multiple times to estimate expected statistics for both players and teams. We then assign a win to a team with more number of wins.

3 Result

We used our model to simulate the 2013-14 season record for each of the 30 NBA teams. We used 70% of 1230 matches in the regular season as the training set and the remaining 30% as the test set. Figure 3 shows that the model provides a good estimate of the teams' actual win percentages with the within-sample R-squared 0.92, and the out-of-sample R-squared 0.87. Our model's performance in predicting average winning percentage

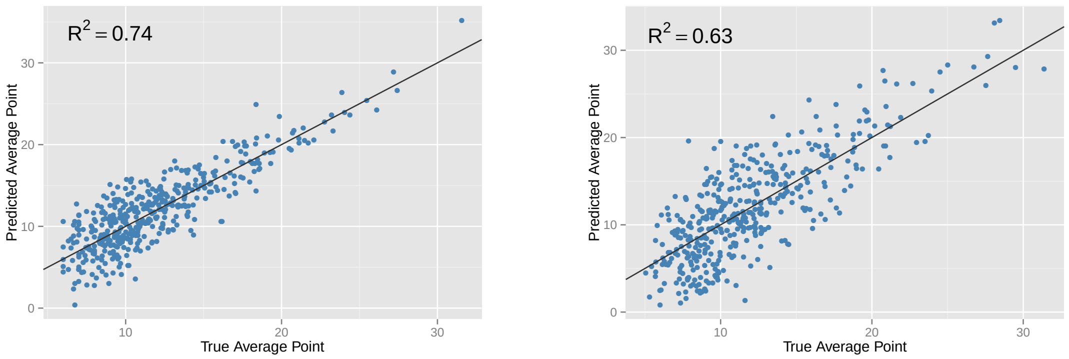


(a) Within-sample result

(b) Out-of-sample results

Figure 3: True vs. predicted win percentages for the 2013-14 season

is comparable to Shirley [1] and Štrumbelj and Vračar [2]. The predicted per-game season average personal statistics such as points per game shows correlation with the actual data (R-square of 0.74 for training data and 0.63 for test data). For player level statistics, we have higher variance and bias in our prediction, especially for players who score fewer points.

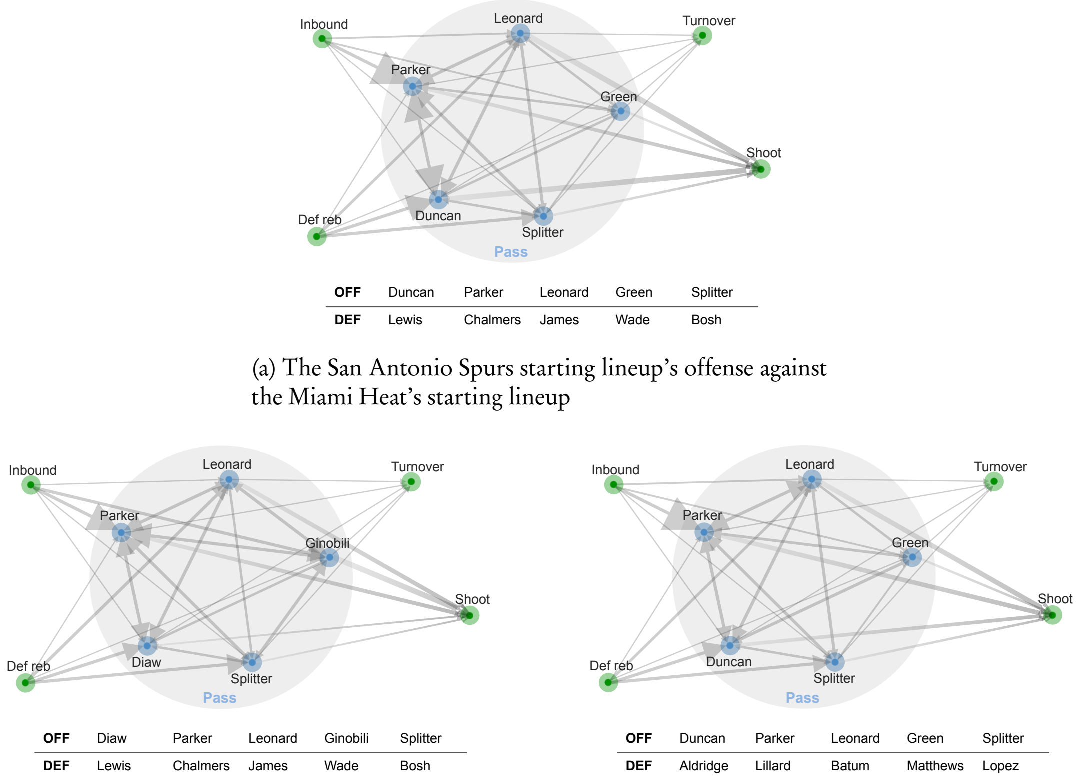


(a) Within-sample result

(b) Out-of-sample result

Figure 4: True vs. predicted average point for players for the 2013-14 season

We used our model on the 2014 NBA Finals matchup between the San Antonio Spurs and the Miami Heat. We computed the conditional probabilities for the pass network and player actions of the Spurs' starting lineup against the defense of the Heat's starting lineup, and the graph for the matchup is shown in Figure 5a. While fixing the defense, player substitutions — Boris Diaw for Tim Duncan and Manu Ginobili for Danny Green — result in a graph with different conditional probabilities of all events (Figure 5b). We observe that Ginobili at-



OFF	Duncan	Parker	Leonard	Green	Splitter
DEF	Lewis	Chalmers	James	Wade	Bosh

(a) The San Antonio Spurs starting lineup's offense against the Miami Heat's starting lineup

OFF	Diaw	Parker	Leonard	Ginobili	Splitter
DEF	Lewis	Chalmers	James	Wade	Bosh

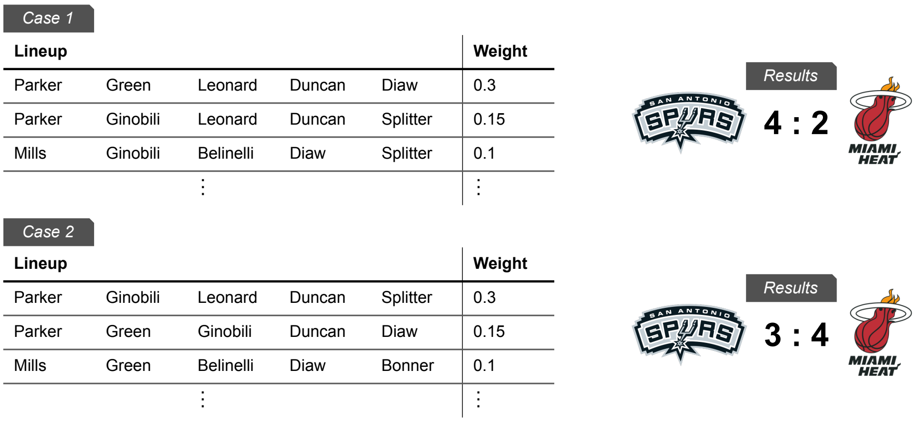
OFF	Duncan	Parker	Leonard	Green	Splitter
DEF	Aldridge	Lillard	Batum	Matthews	Lopez

(b) Change in the offensive lineup

(c) Change in the opponent team

Figure 5: Graphs of offense with changes in the team lineup or in the opponent lineup/team

tracts the ball from other players more than Green does. Also with Duncan out, we expect more ball movement between backcourt players, and team shooting also shifts significantly towards guards and small forward. Note that Ginobili and Diaw have the same positions as Green and Duncan respectively. However, the graphs are quite different for the two lineups within the same team. This suggests that defining the possession network of a team only in terms of player positions is not sufficient. For another comparison, we also computed the network of the Spurs' starting lineup against the Portland Trail Blazers' starting lineup, i.e. fixing the team lineup but changing the opponent lineup or team (Figure 5c). We observe a clear drop in Duncan's shot attempt probability, compared to the base case against the Heat in Figure 5a. This is due to the effect of his defender, Lamarcus Aldridge, who has a higher defensive ability to reduce shot frequency. Subsequently, we observe an increase in expected shot frequency for Parker. These comparisons suggest that either changes in the team lineup or in the opponent team lineup significantly affects the dynamics of a match progression.



Case 1

Lineup					Weight
Parker	Green	Leonard	Duncan	Diaw	0.3
Parker	Ginobili	Leonard	Duncan	Splitter	0.15
Mills	Ginobili	Belinelli	Diaw	Splitter	0.1
		⋮			⋮

Results

4 : 2

Case 2

Lineup					Weight
Parker	Ginobili	Leonard	Duncan	Splitter	0.3
Parker	Green	Ginobili	Duncan	Diaw	0.15
Mills	Green	Belinelli	Diaw	Bonner	0.1
		⋮			⋮

Results

3 : 4

SAN ANTONIO SPURS
MIAMI HEAT

Figure 6: Simulation results on the 2014 NBA Finals

Having established that the ball dynamics is significantly influenced by different sets of players on the court, our model can be used to evaluate the effect of different lineups on game results. For demonstration, we simulated the 2014 NBA Finals with two distinct sets of lineups of the Spurs while keeping the lineups of the Heat fixed (we used the Heat's lineups used in Game 5 of the Finals for both case 1 and case 2). We assigned different weights to each lineup as shown in Figure 6 in order to allocate different number of possessions given to lineups in a given game. 101 simulations were performed to determine a single win as described in section 2.9. We applied conventional best-of-seven playoff format to determine the winner of the Finals. The results show that changes in lineups and weights affect the outcome of the series.

4 Conclusion

The simulation model we propose helps answer not only the team level questions, e.g. which team will win a match, or which teams will advance to the playoffs, but also player level questions, such as how well a specific player will perform in a given match or the entire season. The model offers the infrastructure for match simulation that will allow the front office or the coaching staff of the team to evaluate the performance of hypothetical lineups against specific opponents. It also gives insight on minute allocation between players. One of the limitations of our current simulation model is not being able to estimate pass network for hypothetical lineup of players from different teams since we do not have the pass data for players from different teams (also because there are distinct pass structures for different teams as argued by Fewell et al.[3]). Also, we do not take assist and block into account in the simulation. Our future endeavor would be to modify our model to overcome these limitations. This will help the teams evaluate the value of future acquisitions in the context of the existing roster. Moreover, the simulation model could also be used to predict performance of Fantasy Basketball teams.

5 Acknowledgements

The authors would like to thank Prof. Martin Haugh for the discussions and constructive comments on the paper.

6 References

[1] Shirley, K. (2007). A Markov model for basketball. Poster presentation at New England Symposium for Statistics in Sports, Boston, MA, September 2007.

[2] Štrumbelj, E. & Vračar, P (2012). Simulating a basketball match with a homogeneous Markov model and forecasting the outcome. International Journal of Forecasting 28 (2012) 532-542

[3] Fewell, J., Armbruster, D., Ingraham, J., Petersen, A. & Waters, J. (2012). Basketball Teams as Strategic Networks. PLoS ONE 7(11): e47445. doi:10.1371/journal.pone.0047445

[4] Franks, A., Miller, A., Bornn, L., & Goldsberry, K. (2014). Characterizing the Spatial Structure of Defensive Skill in Professional Basketball. arXiv:1405.0231