Analyzing NBA Player Positions and Interactions with Density-Functional Fluctuation Theory

Basketball
ID: 193938

Introduction

The growing availability of player tracking data in sports, notably in basketball, has the potential to improve how we quantify player abilities and how we understand the interplay between individual talents and overall team strategies. The main challenge lies in effectively processing this tracking data.

In the limit of an infinitely large data set the problem is statistical: *given the positions of the players and the ball, what % of the time is the result a 3-pointer, a 2-pointer, a miss?* In this limit it would be possible to consider how the %outcome changes as a particular player is moved to different locations on the court, as one player is exchanged for another, or generally any aspect of team composition. For finite data sets it is necessary to perform data reductions, to indicate what constitutes as *similar positions* to be able to obtain workable statistics, and traditional metrics precisely employ such data-reduction. For example, the three-point goal percentage (3P%) provides a description of a player skill that reduces data from all situations (different clock times, player positioning, defense composition, *etc.*) to simply whether a shot attempt by the considered player is taken from beyond the 3-point line. In this manner statistical measures such as 3P% are obtained by using similar positions—in this case when the shooter is beyond the three-point line—without ever requiring the presence of truly identical positions.

To obtain detailed measures from tracking data what should constitute as 'similar positions' is non-trivial. However, problems of this form—understanding fundamental relationships from positional data—arise frequently in the physical and social sciences, and approaches developed in these fields have proven to be highly influential. In particular, density-functional theory (DFT) is a Nobel-prize-winning technique that reduces multi-body molecular forces into a tractable form, now representing the cutting-edge of molecular modeling. Density-functional fluctuation theory (DFFT) extends DFT to situations where the underlying interactions are unknown by typically partitioning relationships into social interactions (those between interacting agents) and location-based preferences. DFFT has been successfully applied to complex systems ranging from groups of insects to racial segregation in major cities [1-3].

1.1. Player Densities

In DFFT, interacting agents—in our case basketball players—are replaced by consideration of their densities. When the number of interacting agents is large, these densities can be constructed by counting how many agents are localized at a particular location. However, as the number of interacting agents in basketball is small (there are only 10 players), we generate a 'player density' which represents the influence of each player with a tapering spatial extent. In this work, the density generated by a player will be represented by a 2d Gaussian with a standard deviation of 5 feet, $\sigma = 5$, centered at the player's actual location. We will further at this point treat players of the same team as being indistinguishable, meaning that if two offense players on the team are swapped the same density distribution is generated. An example of offense and defense densities are shown in Figure 1.

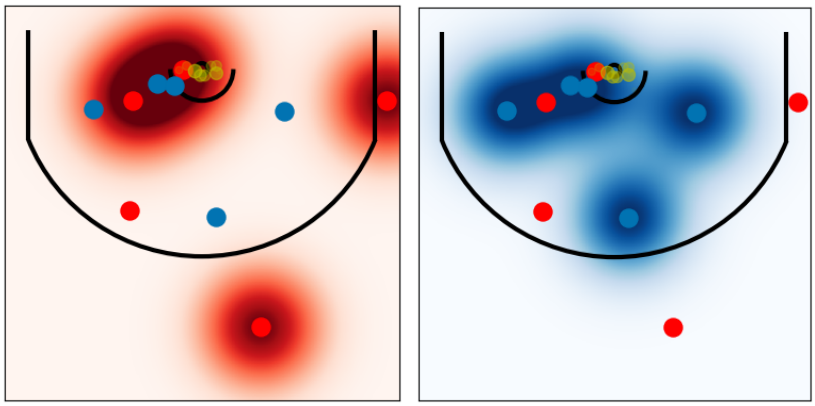


Figure 1: Densities at the time of a shot. Offense (Red, left) and defense (Blue, right). The ball's location (small yellow circle) over the previous few milliseconds is shown.

Now we can consider the densities that occur at different locations on the court for any snapshot of a play. In this work, we will consider a partition of the half-court into a 10x10 grid of blocks ($N_{blocks} = 100$) consisting of 5x5 feet squares and evaluate the cumulative offense and defense density that occurs in each block. Finally, instead of treating the density in each block as being continuous we can set the offense and defense densities to have discrete levels. If we set there to be 14 levels ($N_{Bins} = 14$) with 0 indicating the absence of density, 5 indicating an individual in the center of a bin that is far from anyone else, and 13 indicating three defense (or offense) individuals standing close each other, we can start obtaining a tractable number of parameters for basketball.

In particular, by treating players as generating a density influence and discretizing the densities and the court, we can determine 'similar' positions in a very general sense. If two positions are identical, they would generate the same densities for each block in the half-court. If one player is moved by, for example, ~foot in any direction this would instigate a *slight* change in densities in the surrounding blocks near the player, reasonably identifying such a ~foot change as constituting a different but *very* similar position.

1.2. Model 1: Modeling Player Densities Probabilistically

With the joint density (offense and defense) in each block allowed to take one of N_{bins}^2 values, if we wanted to model the game in an entirely probabilistic sense we would need to determine $N_{blocks} \cdot N_{bins}^2 \approx 20\,000$ parameters for the half court. In other words, when a desired set of conditions are satisfied—such as ball location, time on clock, shot outcome, *etc.*—we can quantify the probability that block b has an offense density given by n_o and a defense density given by n_d,

$$P_b(n_o, n_d).$$

For example, when the shot will be taken within the next 3 seconds, will miss, and the ball is located at [-15.5, 22], we can obtain a probability distribution for densities in block b based on when these conditions were satisfied in games. Consider block b to be centered at [7.5. 37.5], with a schematic shown in Figure 2.

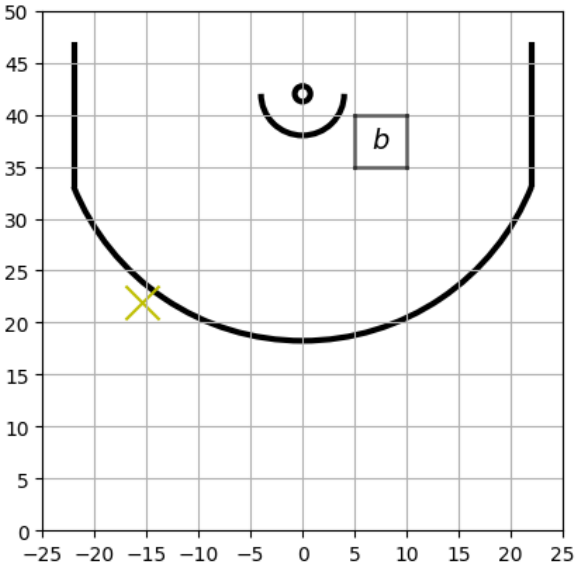


Figure 2: Schematic of considered block.

By obtaining statistics for when these conditions are satisfied, the distribution of densities in block b is shown in Figure 3.

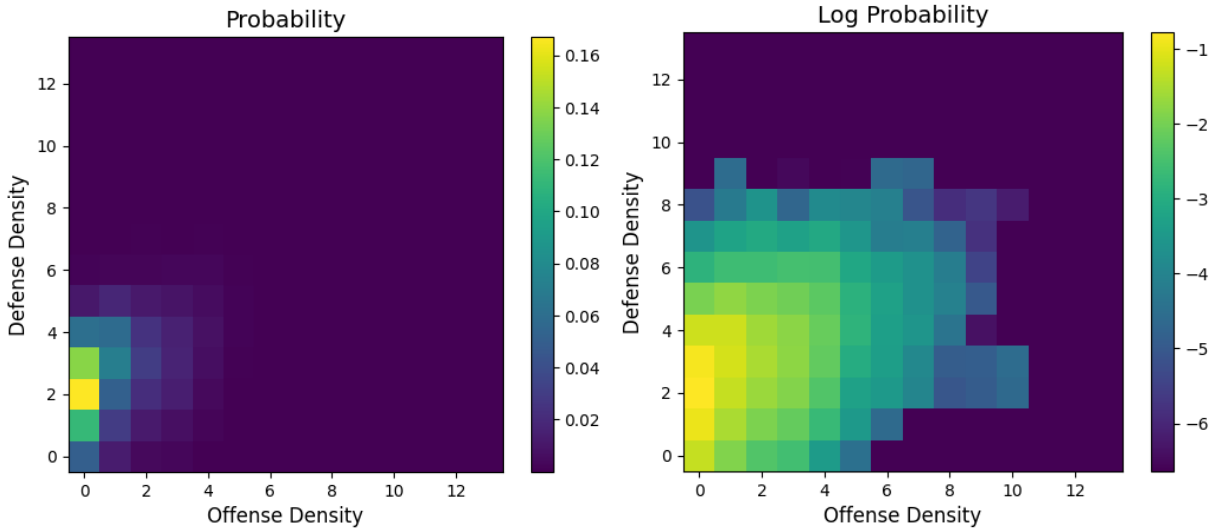


Figure 3: Probabilities of densities in block at [7.5, 37.5]

From these densities we see, appropriately, that the defense density in this block tends to be greater than the offense density. The average defense density is 2.5, suggesting that at least 1 defense player is often present near this bin, in good position to get a weak-side rebound (recall that 5 corresponds to a density when a player is at the center of the bin, and for reference the average across all bins is ~ 1), while the offense average defense density in this bin is 0.8, suggesting that it would be common to encounter no offense players nearby. Furthermore, note that the offense and defense densities are correlated, which is more visible by viewing $\log\big(P_b(n_o, n_d)\big)$. In particular, note that an offense density of 6 with a defense density of 0 was never observed in the data, while an offense density of 6 with a

defense density of 4 is not completely improbable; demonstrating that the bin is more likely to contain offense *and* defense players nearby rather than only offense.

If we were to consider a specific snapshot of a play, we can then determine the densities associated with this snapshot in each bin and determine how likely these densities are to correspond with certain conditions (*e.g.,* a quantified $P_b(n_o, n_d)$ from data when the result of the plays was a miss). A common measure to evaluate the likelihood of the overall position (across all blocks) is the so-called log-likelihood, which can be evaluated for a snapshot of play by using

$$lnL = \sum_b \ln\left(P_b\left(n_{o,b}, n_{d,b}\right)\right),$$

where $n_{o,b}$ and $n_{d,b}$ are the offense and defense densities in block b generated by the snapshot of the play. We will demonstrate in the results section how to correlate this likelihood score with more intuitive metrics.

1.3 Model 2: Modeling Player Densities with DFFT

Given that there is tracking data that contains many thousands of plays, and each position provides N_{blocks} worth of data points (one per block), if we were to treat players of the same team as being indistinguishable (as we have been doing so far) we may have sufficient data to consider the situation in an entirely probabilistic sense. However, if we want to distinguish between different players, account for the ball position, or generally be able to work with smaller data sets it is often necessary to reduce the number of parameters, for which we can construct standard DFFT models.

So far, we have been treating the distribution in each block as being distinct, evaluated completely separately from the densities observed in other blocks. However, there are a lot of similarities between blocks: the court is symmetric along the central axis, and we should expect contiguous blocks to have similar density distributions. A common assumption in DFFT is to consider the average density as being a distinguishing factor between blocks, an assumption that seems particularly salient for basketball given that the average density for different parts of the court vary substantially, as shown in Figure 4.

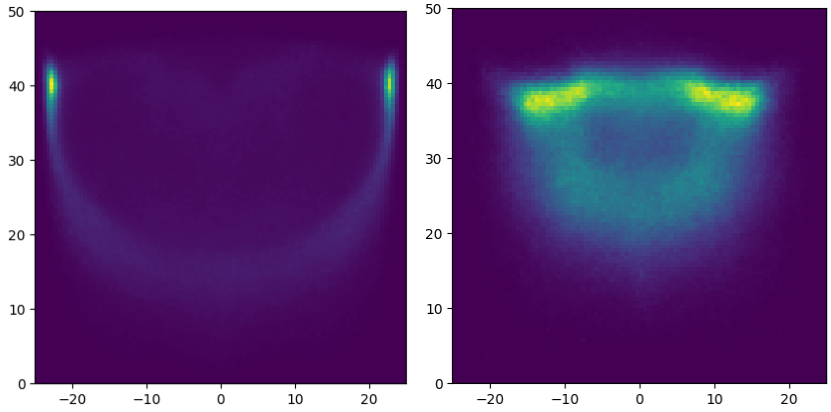


Figure 4: Average offense and defense positions. Note the non-uniformity for different parts of the court. The offense is often situated along the 3-point line.

Suppose, then, that we were to consider the average density as being the distinguishing factor between different blocks. To construct the appropriate model, suppose that we have quantified the parameters for one block, block b, and want to use the resulting values as a starting point for what should be expected in a different block, block b', but subject to the constraint of known average densities (offense and defense) in b'. To construct a model for our expected distribution in block b', we can minimize the KL-divergence between the distribution we know, $P_b(n_o, n_d)$, and our model distribution for block b', $\widetilde{P_{b'}}(n_o, n_d)$, given that $\widetilde{P_{b'}}(n_o, n_d)$ satisfies the known average offense and defense densities in block b', $\sum_{n_o,n_d} n_o \, \widetilde{P_{b'}}(n_o, n_d) = \langle n_o \rangle_{b'}$ and $\sum_{n_o,n_d} n_d \, \widetilde{P_{b'}}(n_o, n_d) = \langle n_d \rangle_{b'}$. The result of performing this minimization over all possible distributions that satisfy the constraints, $\rho(n_o, n_d)$,

$$\widetilde{P_{b'}}(n_o, n_d) = \underset{\rho(n_o,n_d)}{\operatorname{argmin}} \sum_{n_o,n_d} \rho(n_o, n_d) \ln \frac{\rho(n_o, n_d)}{P_b(n_o, n_d)},$$

can be re-written in terms of Lagrange multipliers to

$$\widetilde{P_{b'}}(n_o, n_d) = \frac{1}{Z} P_b(n_o, n_d) e^{-n_o v_o - n_d v_d},$$

where v_{o} and v_{d} need to be optimized so that the model distribution for block b', $\widetilde{P_{b'}}(n_o, n_d)$, satisfies the average density values we know in block b'.

In literature, minimizing the KL-divergence is sometimes known as minimum discrimination information or minimum cross-entropy. To provide an interpretation of this procedure, the KL-divergence itself is a common measure of 'distance' of one distribution from another. Finding the distribution that minimizes the KL-divergence is equivalent to identifying the conditional distribution $P_b(n_o, n_d | \langle n_o \rangle_{b'}, \langle n_d \rangle_{b'})$, in the sense that if $P_b(n_o, n_d)$ is sampled repeatedly, and the resulting sampled distribution satisfies the target average values, then the sampled distribution is expected to be given by $P_b(n_o, n_d | \langle n_o \rangle_{b'}, \langle n_d \rangle_{b'})$. Furthermore, the optimized model distribution $\widetilde{P_{b'}}(n_o, n_d)$ is guaranteed to be 'closer' to the true distribution at block b', $P_{b'}(n_o, n_d)$, in the sense of having a smaller KL-divergence value,

$$\sum_{n_o,n_d} P_{b'}(n_o, n_d) \ln \frac{P_{b'}(n_o, n_d)}{\widetilde{P_{b'}}(n_o, n_d)} \leq \sum_{n_o,n_d} P_{b'}(n_o, n_d) \ln \frac{P_{b'}(n_o, n_d)}{P_b(n_o, n_d)},$$

with equality holding only when $P_b(n_o, n_d)$ has the same average densities as block b'.

This approach is quite general and can be used to introduce other constraints, and for each constraint a new parameter (akin to v above) would be introduced. In this work we will limit ourselves to consideration of average values between blocks as being their distinguishing factor as this minimizes the number of parameters required, $N_{bins}^2 + 2\, N_{blocks} \approx 400$, and we find this model to perform well. The model for each block is then given by

$$P_b(n_o,\, n_d) = \frac{1}{Z}\, e^{-n_o v_{b,o} - n_d v_{b,d} - f(n_o,\, n_d)},$$

where $v_{b,o}$ and $v_{b,d}$ are parameters that identify the offense and defense average densities in each block and $f(n_o,\, n_d)$ is a non-parametric block-independent function containing N_{bins}^2 parameters. Note that $f(n_o,\, n_d)$ arises because we no longer consider block b as being special compared to b', but the form is exactly mathematically equivalent to the form we obtained by minimizing the KL-divergence, *i.e.*, it is equivalent to considering a form $\frac{1}{Z}\, g(n_o,\, n_d)\, e^{-n_o v_{b,o} - n_d v_{b,d}}$, where $g(n_o,\, n_d) \equiv e^{-f(n_o,\, n_d)}$, and in the minimizing of the KL-divergence procedure $g(n_o,\, n_d) = P_b(n_o, n_d)$.

1.3 Data

Using Second Spectrum player tracking data from a portion of the 2022-23 NBA season, we derive optimized model parameters and play outcome likelihoods for a subset of tracking datapoints. We also calculate selected player-specific offensive and defensive values, in particular, values quantifying the effect of offensive players on the density of defensive players, and values quantifying the probability that specific defensive players will be in good or bad position to prevent 2-point or 3-point scoring outcomes.

2. Results

If the DFFT model (Model 2) is successful, then we should be able to compute anything using the DFFT model which can be obtained from a probabilistic model (Model 1) but by using relatively few parameters, greatly reducing the amount of training data required. This means that by 'training' the DFFT model on data sets satisfying different conditions (teams, ball location, outcome, *etc.*), we can ask questions such as what are the chances that these different conditions are met based on a snapshot of a play. This allows us to determine where the defense is likely to be based on the positions of the offense, or how much greater the defense density is around specific players compared to what is expected for an average player.

2.1 Locating Players

As a first example, suppose that we know the location of 9 players (5 offense + 4 defense) as well as the ball, and we are trying to determine the location of the last defensive player. To determine which positions for this player are likely, we can evaluate the loglikelihood lnL when the player is placed at different potential positions, which will give a score of how likely a player is to be at each of the considered hypothetical locations. For example, training Model 1 (the probabilistic model) on data when the ball is near [-18, 22] and the result was a missed shot attempt, we get the prediction shown in Figure 5.

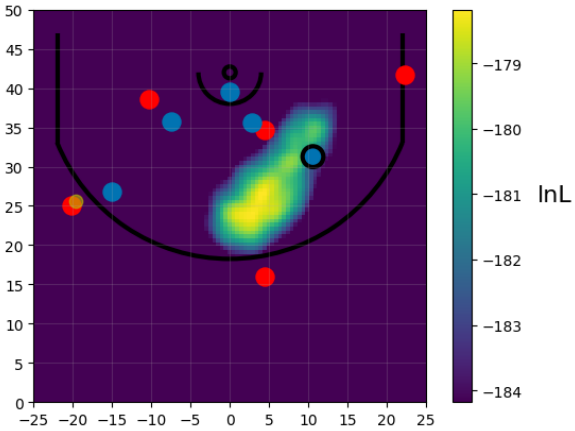


Figure 5: Blue circle with black outline represents the hidden player and the heatmap demonstrates the lnL values.

It is further useful to consider the predicted locations for this player when the model is trained on different play outcomes (0, 2, or 3), which means the model is trained on entirely different data sets and is therefore a good test of the generalizability of the model. Figure 6 shows the predicted locations of this play with the probabilistic and DFFT models.

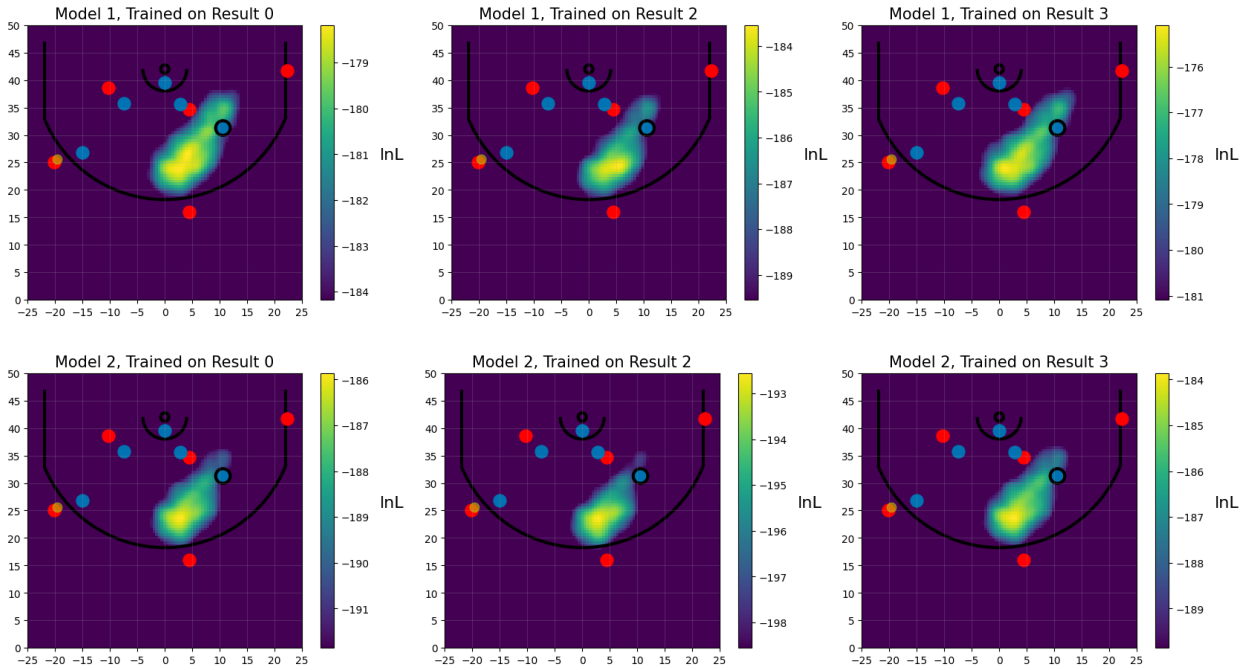


Figure 6: Predictions based on Model 1 (probabilistic) and Model 2 (DFFT). The predictions using the models and different training data show only slight differences, demonstrating that the models are able to generalize to positions not in their training set.

Considering the different model and training data sets shown in Figure 6, it becomes apparent that where a player is expected to be does not depend very strongly on the outcome of a play. Furthermore,

it is found that the hidden player is standing a bit closer to the corner offense player than what would typically be expected for this position. Finally, the *lnL* values can be compared when the same model is used, which demonstrates that both models believe the result of this play is likely to be a 0 or a 3, rather than a 2 (indicated by the great lnL values). This agrees with our expectation that the player with the ball may pass to his teammate at the top of the key, who may attempt a 3-point shot.

By considering over 1000 randomly selected plays scattered throughout the season, Figure 7 demonstrates how frequently we were able to localize the 'hidden' player to a portion of the half-court. The results of this procedure find that the player is localized to within 1% of the half-court (25 square feet) around a quarter of the time and to within 5% of the half-court (125 square feet) around 60% of the time. This demonstrates that both models are able to localize the location of a hidden player quite well and that the DFFT model performs only marginally worse than the probabilistic model despite using two orders of magnitude fewer parameters.

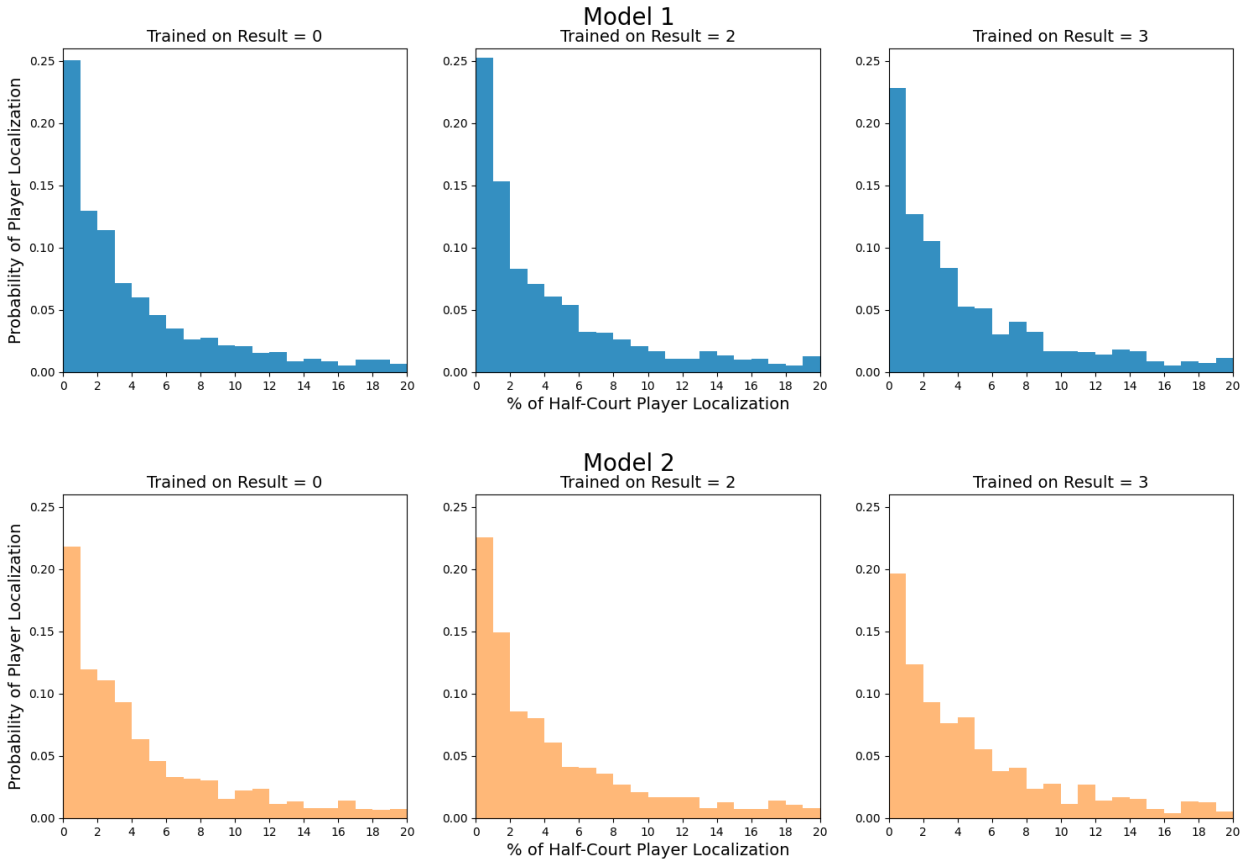


Figure 7: Distribution of of hidden player localization with different models and training data based on over 1000 randomly selected basketball snapshots.

As a direct extension, we can determine how the location of a player is expected to change when some aspect of the play is altered. For example, if the offense player in the corner had the ball instead, the distribution for the hidden player would be given by the heatmap shown in Figure 8 (using Model 1, trained on Result = 0). Future work can explore altering specific aspects of the game to see which alterations lead to most significant changes or more generally simulate a play based by having players continually move to likely locations in increments.

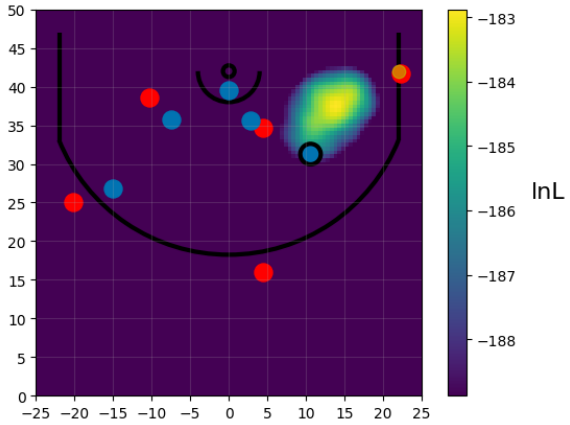


Figure 8: Changing which offense player has the ball.

Finally, we note that we are not limited to predicting the location of a single player and can determine the entire density distribution of a team. For example, the average defense density based only on the location of the offense players and the ball, is shown in Figure 9.

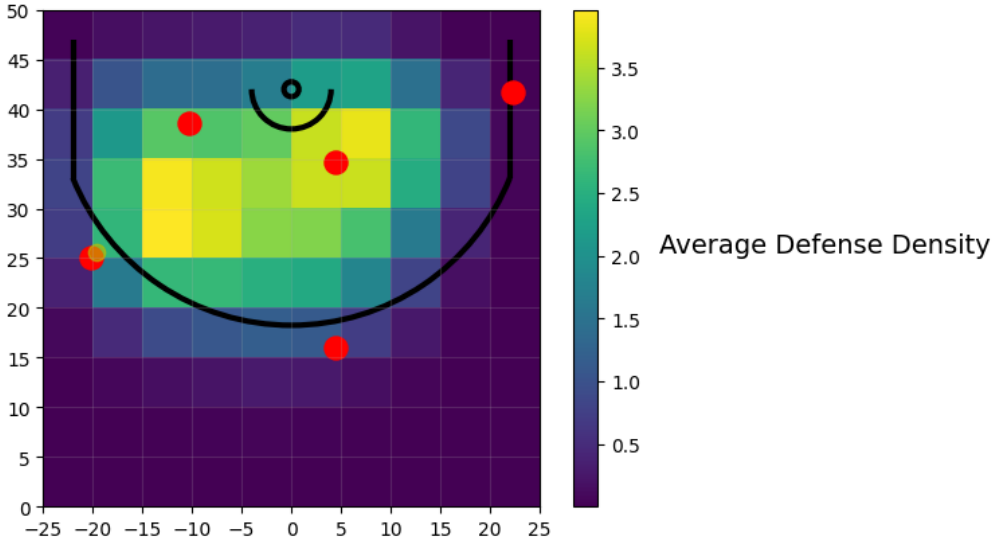


Figure 9: Forecasting defense density based on location of the offense players and the ball.

2.2 Correlating loglikelihood with %outcomes

As was mentioned in the previous subsection, it is possible to determine which conditions (score outcome of play, *etc.*) are likely to occur based on lnL values from a model trained on data sets for the various outcomes. For example, extending our work on identifying the hidden player, we can ask which of the potential positions for this hidden player would be more likely to lead to a 0-point, 2-point, or 3-point outcomes for the play, as shown in Figure 10.

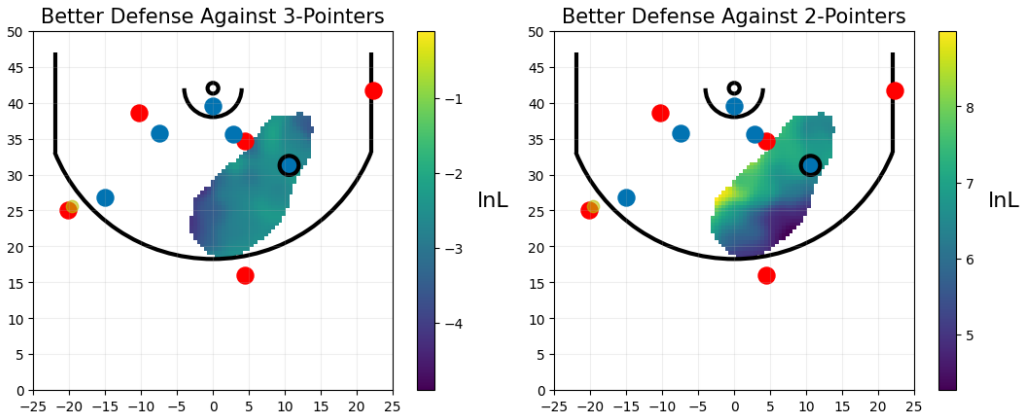


Figure 10: Identifying better and worse positions for the hidden player for resulting in play outcomes. Determined by the difference in heatmaps between a miss result and the respective outcome.

The lnL values demonstrate that small changes in the hidden player's position are not likely to have a substantial effect on the outcome of the play when it comes to 3-point outcomes; however, if this defense player was substantially closer to the player with the ball a 3-point outcome would become even more likely (a lnL difference of 1 compared to his actual position). By a similar consideration for defense against 2-pointers, we see that the overall position of the defense is already highly defending against 2-point results (lnL values of around 8), however we see that fairly small changes in the position of the hidden player have substantial effects. If the hidden player moved closer to the player at the top of the key a 2-point shot would become more likely by a lnL difference of 2, while moving closer to the player with the ball would slightly improve defense against 2-pointers. Overall, this analysis demonstrates that the defensive player is standing in a fairly strong position, there are no nearby positions that would substantially result in a better team defense against 3-pointers (the current threat for the team) and he is also in decent position to help defend against 2-pointers.

So far we have been using lnL values, but it beneficial to relate lnL measure with more intuitive measures, such as the %of various outcomes. To this end, after partitioning the data into training and testing sets, we can determine how our lnP values correlate with the actual play outcomes on the testing sets, as shown in Figure 11. What we then see is that lnL values are fairly linear with %outcome between different play results, and suggesting that a lnL difference of 1 corresponds roughly to %outcome difference of 2%. So the lnL difference of 2 for the hidden player (that we observed for 2-point vs 0-point results) equates to a 2-point outcome being around 4% more likely than a 2-point result compared to his current location (do note, however, that a 2-point outcome for this play is still unlikely).

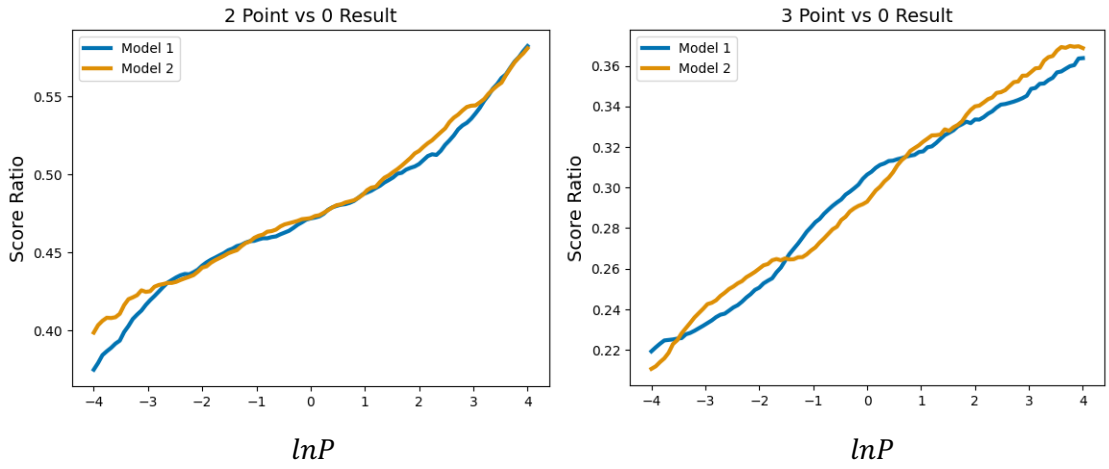


Figure 11: correlation of lnP with shot outcomes. For 2 point result (left), Score Ratio = $\frac{\#2\text{points}}{\#2\text{points}+\#0\text{pointers}}$, similarly for 3-point results (right). Therefore, for every point increase in lnP the difference in the score ratio is around 2%.

2.2 Defense Prowess

Now we can evaluate if a particular player is doing a good or bad job positioning himself to prevent a certain outcome in his immediate vicinity. This evaluation takes into account the location of all other players and the ball. We can then consider hundreds of positions for each player to see which players are consistently in good positions to prevent certain outcomes. Figure 12 shows when the offense ends up being successful in the play (2-point or 3-point result), while Figure 13 shows when the defense is successful (0-point result). The lnL values correspond to the difference of the actual position compared to average over likely positions for the player (plays when the considered player was not in a likely position were excluded) using the DFFT model (Model 2).

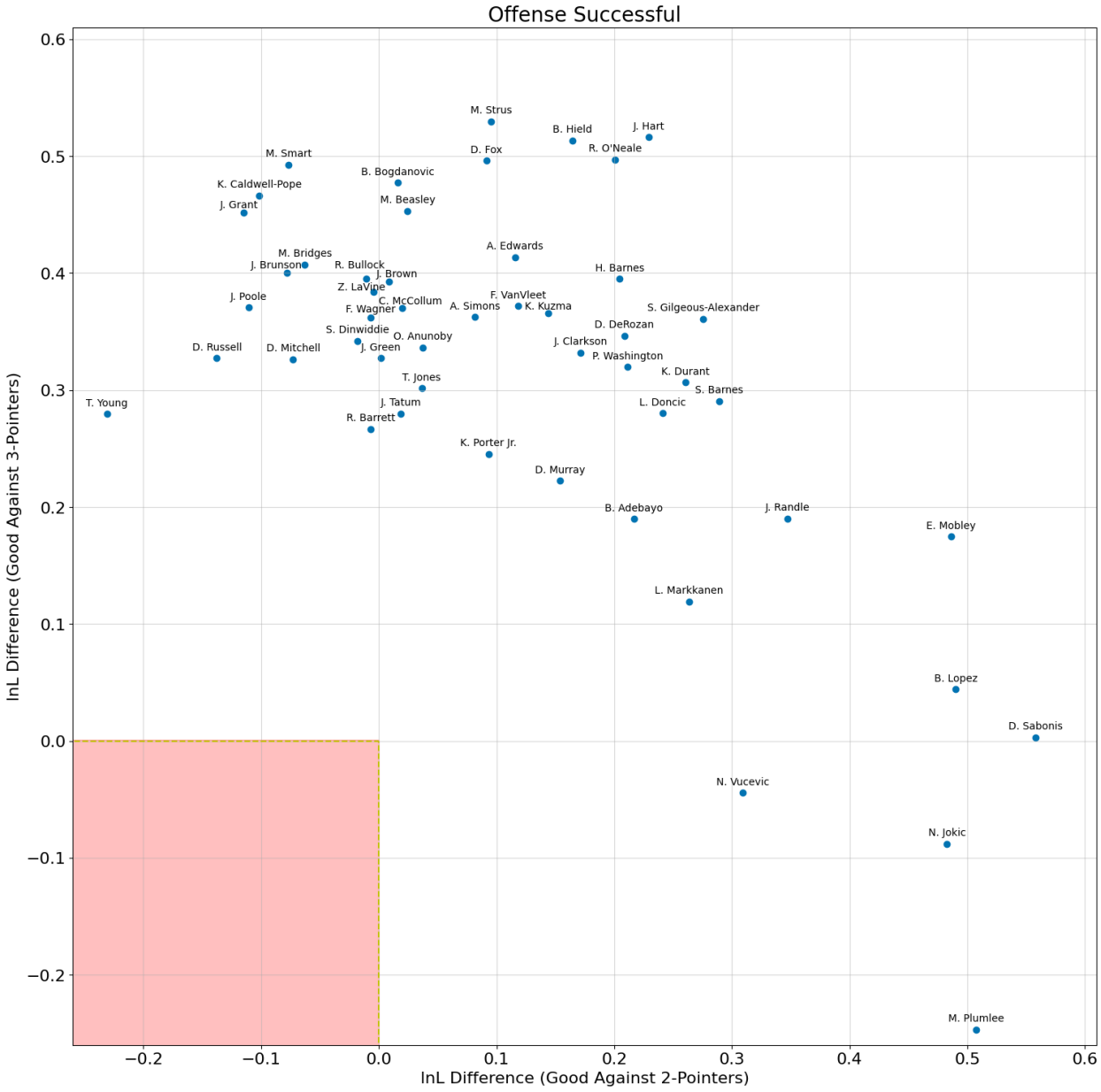


Figure 12: Average defense prowess of 50 of the most frequent players in the data set when offense is successful.

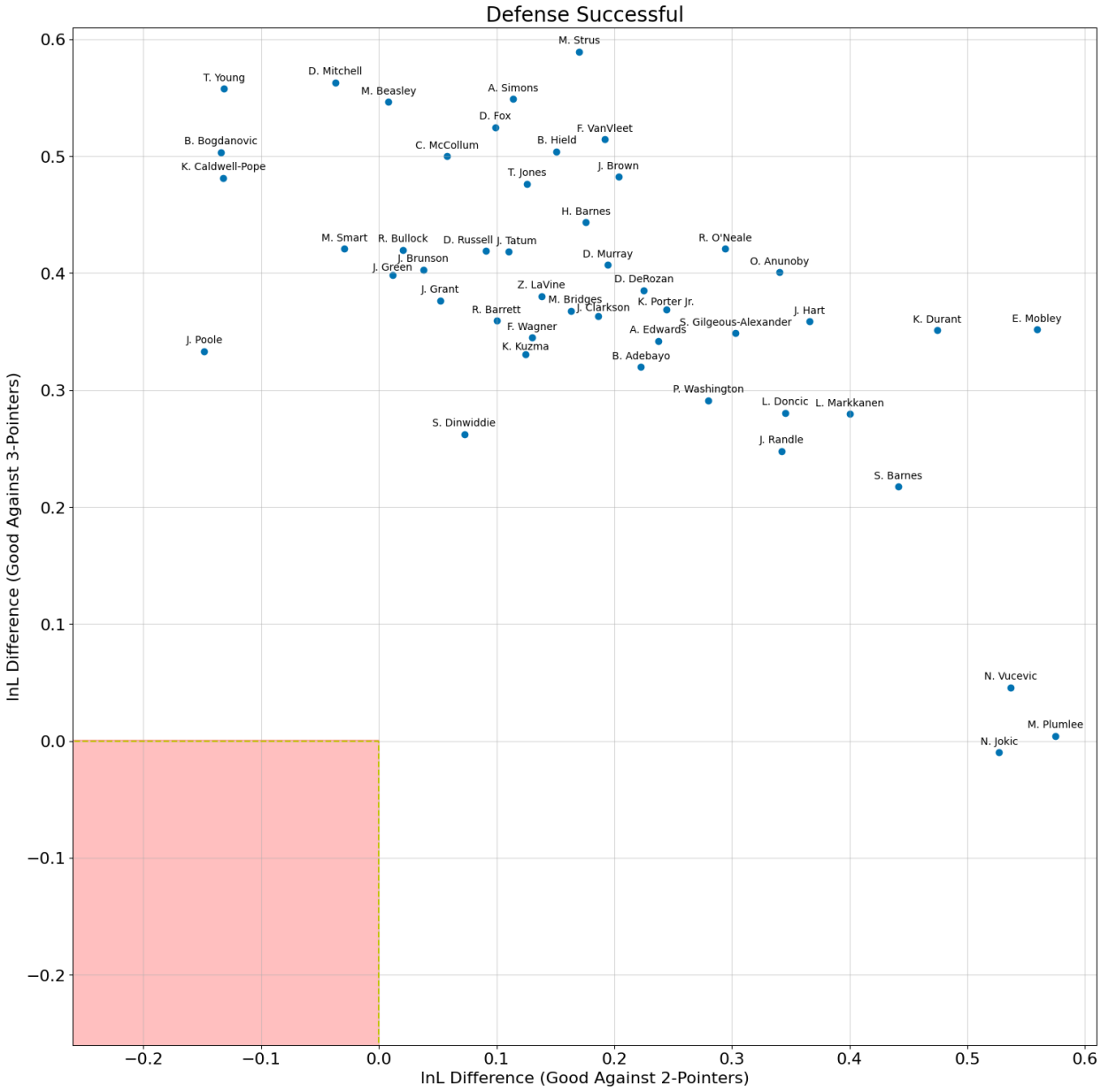


Figure 13: Average defense prowess of 50 of the most frequent players in the data set when defense is successful.

We see immediately that the model accurately discriminates between backcourt defensive players, who generally prioritize preventing 3-point shots (y>x), and frontcourt defensive players, who generally prioritize preventing 2-point shots (x>y). Furthermore, we can see which players are most successful in positioning themselves well defensively (x+y is large) and which players could improve their positioning (x+y is small).

2.3 Player Gravity

Perhaps a particularly natural application of these DFFT ideas is the consideration of player gravity. When considering how offense players attract defense players, the situation is complicated because there could be other offense players nearby. To determine play gravity it is then important to disentagle a specific player's gravity from the gravity of other offensive players, and from their typical position on the court. To account for this complication, we will use the DFFT model and train it on when a specific player (*e.g.,* Luke Doncic) is in a particular location. We will then have a separate parameterized DFFT model per player and a parametrized DFFT model for when any offensive player is present at the respective location to obtain what should be expected for the average player. We will then consider the same set of 100 plays and use the positions of the offense players to generate the expected average density for the defense, similar to Figure 9, using the player-specific parametrized models. Finally, we will compare the density at the location being considered between the player-specific and average-player generated defensive densities. The primary goal of this procedure is to hold the locations of the other four offensive players constant and therefore isolate the effect of the offensive player of interest on the defensive player density. We repeat this procedure at 8 different locations for each player and calculate the "local" offensive player gravity for these locations, i.e. the effect of the player on the defensive presence in his immediate vicinity. We weight these results by how often each specific player is present at each location, and plot the averaged results in Figure 14.

We note that each offensive player has different local gravity at different locations depending on his skills, shown in Figure 15. Highly versatile scorers such as Doncic and Durant have high gravity at all locations, while other players may have high gravity at specific locations due to their skill driving toward the basket (DeRozan, Fox), their skill shooting 3-pointers (Hield, Poole), or their handedness (Russell).

In addition to affecting the average defensive presence in their immediate vicinity, offensive players may affect the defensive presence in other parts of the court through their passing ability; i.e. a "nonlocal" gravity effect. Comparing Jokic to Markkanen, we see that Jokic has a very pronounced nonlocal effect on the defensive player density, shown in Figure 16. Unlike the local gravity effect, this nonlocal effect is not easily captured in a single-number metric. However, it is equally valuable for understanding the unique ways in which star players dictate the game.

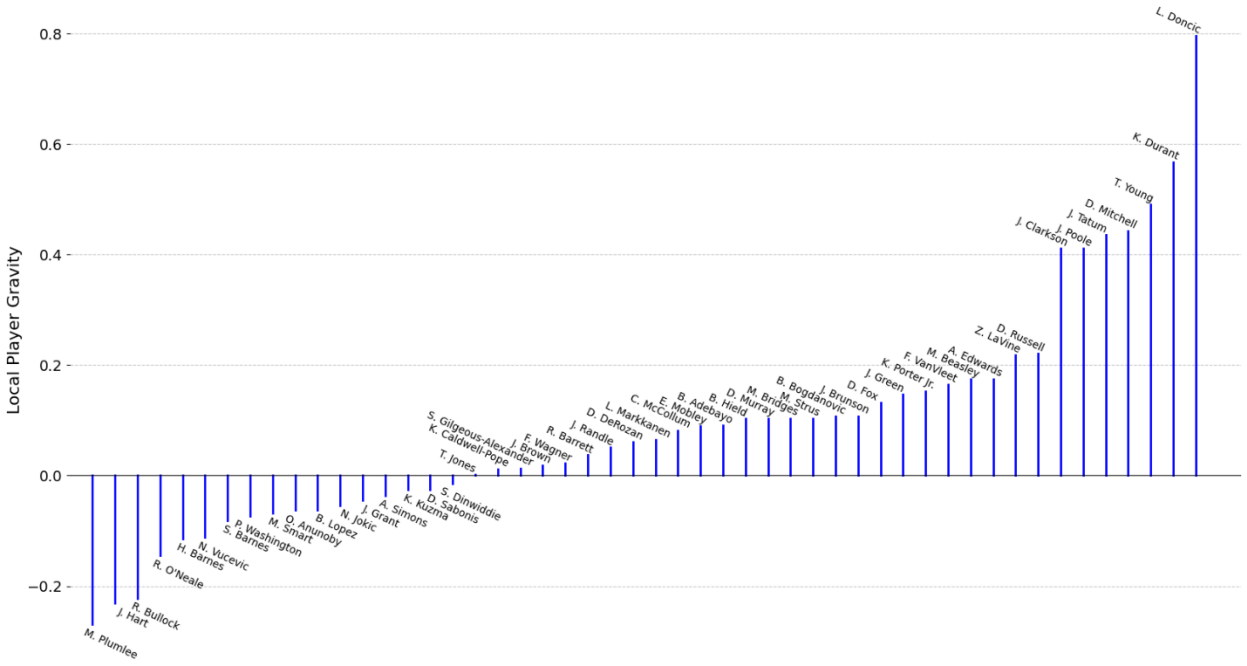
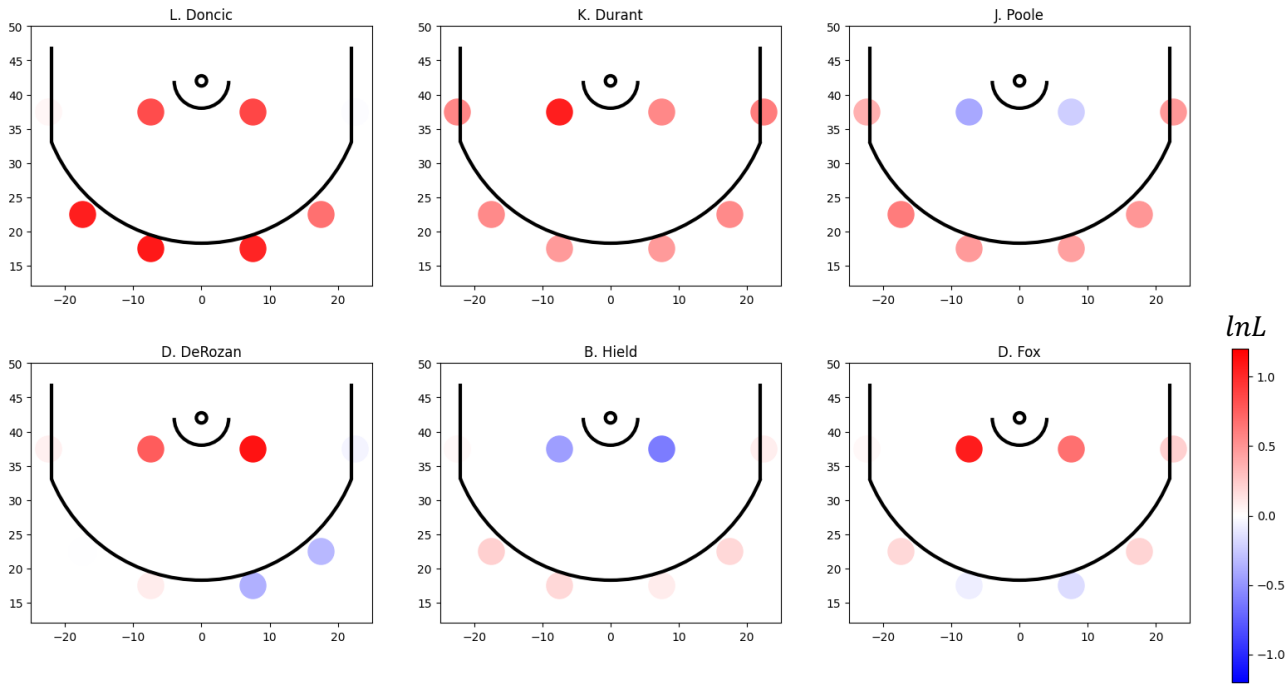


Figure 14. Local player gravity as determined by effect on defense density at the location of the player.



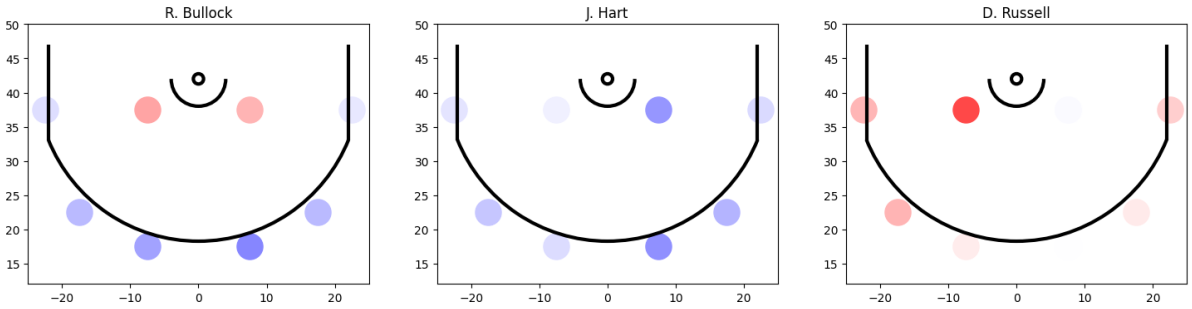


Figure 15: Local gravity for different players at different locations on the court. Red colors indicate higher-than-average local defensive presence when the player is at this position, and blue colors indicate lower-than-average defensive presence when the player is at this position.

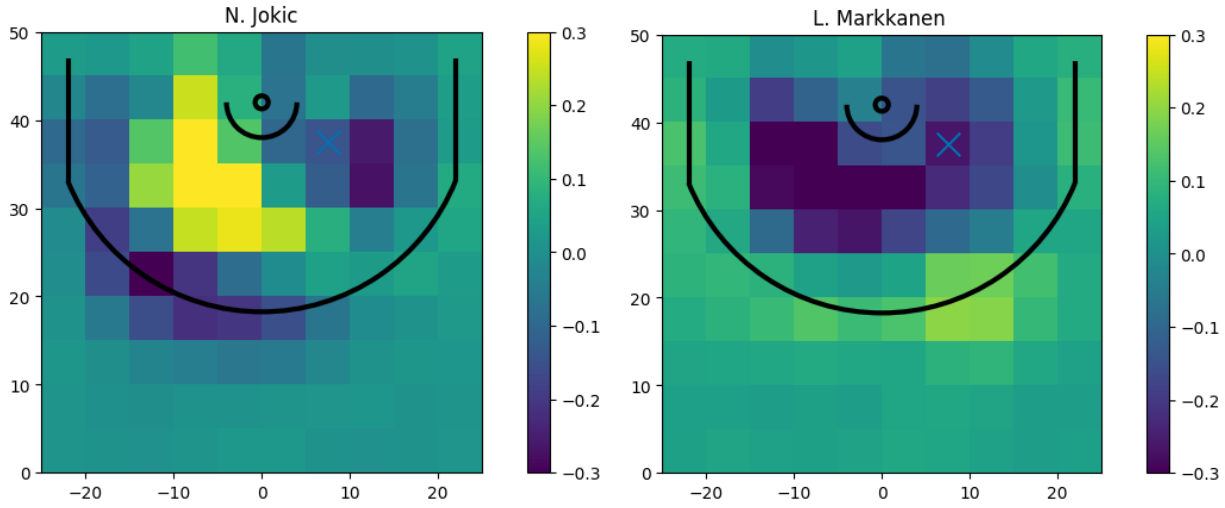


Figure 16: Effect of Jokic and Markkanen on the defensive player distribution when they are on the right block (location indicated by the X). Yellow colors indicate an enhanced defensive presence, and blue colors indicate a reduced defensive presence.

3. Conclusion

DFFT quantifies the subtitles that underlie offensive and defensive strategies in NBA basketball, while requiring orders of magnitude fewer parameters than a probabilistic model. This provides a new way of statistically evaluating player positioning and a detailed view into how the "gravity" of star players changes the game. We show that defensive players significantly affect shot outcome through their positioning, effectively identify which players preferentially defend against 2-pointers or 3-pointers, and quantify the extent to which different players affect the odds of successful defense purely through their positioning. We show that certain offensive players, especially the league's leading scorers, have "gravity" that attracts a larger defensive presence than is seen by typical offensive players. This effect varies depending on the position of the player in good agreement with their known scoring tendencies.

We show that an exceptional playmaker such as NBA Finals MVP Nikola Jokic can have a unique effect on the defensive player density that extends far away from his position, reflecting the defense's response to his unique passing ability. Overall, our new framework for analyzing NBA player tracking data allows for a more detailed understanding of how players respond to one another and how they influence play outcomes on a continuous basis throughout NBA games.

References

[1] Méndez-Valderrama, J. Felipe, et al. "Density-functional fluctuation theory of crowds." *Nature communications* 9.1 (2018): 3538.
[2] Chen, Yuchao, et al. "Small-area Population Forecast in a Segregated City using Density-Functional Fluctuation Theory." *arXiv preprint arXiv:2008.09663* (2020).
[3] Barron, Boris, et al. "Extending the use of information theory in segregation analyses to construct comprehensive models of segregation." *arXiv preprint arXiv:2212.06980* (2022).