

An Analysis of the Basketball Endgame: When to Foul When Trailing and Leading

Franklin H. J. Kenter
Rice University
Houston, TX, U.S.A., 77005
Email: franklin.h.kenter@rice.edu

Abstract

A common tactic near the end of a basketball game is for the trailing team to foul in order to gain an advantage by forcing the opponent to shoot free throws. While this tactic is widely used at almost all levels of play, deeper investigation into if and when a team should foul is nearly absent. In this paper, we model basketball as a combinatorial game to provide, for the first time, a well-supported quantitative description of when to foul. The results are surprising: not only should trailing teams foul earlier and more often than they actually do, but also, the *leading* team should foul more often than the trailing team. Using play-by-play data from NBA games, we illustrate the potential impact of this model.

1 Introduction

Statistical analysis alone cannot lead to understanding rare events or uncovering obscure strategies; it is limited to observable data. Consequently, in the context of analytics, statistics can only help *improve* performance, it cannot *optimize* it. To optimize performance, one needs to go beyond statistics. We turn to *combinatorial game theory* to address this issue.

Combinatorial game theory studies games as a series of simple, alternating moves, and then analyzes these moves to determine how to optimize results, a process known as “solving the game.” Simple examples are tic-tac-toe and backgammon. More complicated examples are chess and the game go. To apply this theory to basketball, we model basketball as a combinatorial game in which teams make alternating choices regarding when to pass, when to shoot, and when to foul. Then, using mathematical and algorithmic tools, we solve this game to determine the optimal move for each team for each possible game state. As a result, we are able to provide for the first time, a well-supported, concrete determination regarding when to foul. Additionally, we determine, under optimal play, how long each possession should take and whether the offense should aim for 2- or 3-point shots.

The tactic of fouling by the trailing team towards the end of a basketball game has grown to become a typical phenomenon since the permanent implementation of the 3-point shot in the 1980s. Since then, it has been pivotal for several notable comebacks and is utilized today at all levels including high school, college, and professional play. It is generally accepted as “part of the game.” The objective of intentional fouling by the trailing team is to create possible advantage by forcing the leading team to shoot free throws. While free throws are generally an easier method for scoring points, fouling can produce two important advantages: stopping the game clock and limiting the total number of points scored in a possession. These two advantages, together, can improve the prospects for the trailing team to make a last-minute comeback with quick successive scores through increased possessions.

Despite the tactic’s widespread use for over 20 years, detailed investigation about if and when a team should foul has been nearly absent. Only recently has some discussion emerged and only in the specific case when the defense is *leading* by three points during the last seconds. In this scenario, fouling severely limits the trailing team’s ability to score the three points needed to force the game into overtime. The consensus here is that the leading team should foul [1, 2]; however, there are differing opinions [3]. Even so, the general case remained unresolved. Given the growth in applying statistics to sport, one would expect this problem would be easily resolved. However, as it turns out, it is not so simple – and requires more than statistical analysis alone.

The results of the combinatorial game-theoretic model are extensive and challenge existing beliefs on several fronts. First, we demonstrate that trailing teams should begin to foul far earlier and far more often than they actually do.

2015 Research Paper Competition
Presented by:

Additionally, we show that not only should the leading team foul when leading by three during the final seconds, but also, earlier with larger leads as well. The leading team should, in fact, foul more often than the trailing team. Finally, and perhaps most important, these results demonstrate that statistical analysis is not necessarily the best tool for evaluating and optimizing aspects of sport. In some cases especially, such as this one, statistics may not have a central role at all. In our case, the application of statistics is limited to determining some model parameters and testing how the model fits observation.

In this paper, we do the following:

- (1) describe our model for basketball as a combinatorial game,
- (2) demonstrate that this combinatorial model provides a good fit by comparing it to observed data,
- (3) apply the results of the combinatorial game to determine when teams should foul,
- (4) implement adaptations of our model to determine the potential benefit for NBA teams, and
- (5) provide discussion regarding the implications of these results.

2 Modeling Basketball as a Combinatorial Game

We aim to model basketball as a simplified combinatorial game similar to chess or backgammon, where the players make alternating moves based on the current game state. We can then analyze the many rapid decisions that occur during the game by recursively solving the game using tools from algorithmic game theory. In particular, once we model basketball as a combinatorial game, we can determine the optimal move under perfect play. Hence, we will be able to determine exactly when teams should pass, shoot and foul, and also, how long each possession should last based on the current score, the game clock, and the shot clock.

The challenge in modeling basketball as a combinatorial game is to realize a balance between two competing forces. On one hand, it is important to capture the overall feel and essence of the meaningful choices in basketball. On the other, the number of such choices needs to be limited so that the game can be solved in a reliable manner. In particular, if there are too many choices, then the number of possible game sequences will grow exponentially, beyond the realm of computability, and no meaningful conclusions can be extracted from the model.

In order to achieve this balance, we limit the choices for each team. In our game, the teams take alternating turns making one of two choices. On the defense's turn, they may either "foul" or "defend," and on the offense's turn, they may either "shoot" or "pass." Here, we assume notions like player position or individual contribution do not have a primary impact and can be excluded. In essence, our model game reduces to the decisions a coach may make from the sidelines. All other aspects of the game including whether a shot is made, who rebounds the ball, and whether a turnover occurs are modeled by chance. The probabilities of these events are determined by the various parameters of the model and can vary depending on the level of play one wishes to analyze. These parameters include shot clock duration, two-point percentage, three-point percentage, free throw percentage, offensive rebound rate, and turnover rate. We will discuss our specific choice of parameters later.

The specific rules of our modeled game are as follows. The game is divided into a sequence of possessions where one team is on offense and the other is on defense. During each possession, the two teams alternate turns, with each turn representing one second of gameplay. On each possession, the defensive team takes the first turn, deciding whether to "foul" or "defend." If the defense "fouls," the offense shoots two free throws. Otherwise, if the defense "defends," the offensive team takes the next turn, choosing whether to "shoot" or to "pass" with the exception that the offense cannot "shoot" during the first 8 seconds of the possession (as this represents the amount of time needed to bring the ball down the court). When the offense "passes," there is a small chance of a turnover based on the turnover rate parameter ($= 2 / \text{turnover rate} \times 100\%$). When the offense "shoots," they choose to attempt a 2-point or a 3-point play. Whenever the offense takes a shot, it is made with a probability equal to the corresponding parameter (i.e., free-throw percentage, two-point percentage, or three-point percentage). If a shot fails, then the offense has a chance to retain possession with a probability corresponding to the offensive rebound percentage (or half the rebound percentage in the case of a missed second free-throw), in which case, the offense restarts with a new possession. For each possession, the teams alternate turns until a foul, shot (without an offensive rebound) or turnover occurs, or the shot clock expires. In which case, the possession ends, and a new possession begins with reversed roles. A sequence of possessions continues until the game clock expires, at which point the team with the most points wins. For the purposes of our model, if there is a tie, the game ends and the winner is determined

randomly each with a 50% chance of winning.

In order to analyze various levels of play, we will vary our parameters depending upon the application. For our purposes, to model typical NCAA Division I men's basketball, we utilize a 35-second shot clock with parameters based on the median values of the corresponding statistics, including 70.0% free throw percentage, 47.5% two-point percentage, 33.0% three-point percentage, 95s turnover rate, and 31.0% offensive rebound rate. Similarly, for NBA basketball, we use a 24-second shot clock, 75.5% free throw percentage, 48.0% two-point percentage, 36.0% three-point percentage, 100s turnover rate, and 27.0% offensive rebound rate. These values are simply based on the median season-long statistics for the corresponding levels of play, so that we can analyze typical play. For specific applications, one can vary these parameters to their needs.

3 Solving the Game and Comparison to Observed Data

The modeled game described may appear to lack typical features of basketball such as dribbling, shooting, and individual actions of players. But the analytical and strategic aspects are strongly preserved. The model focuses on the choices of the strategies utilized by the team. We will demonstrate this by comparing the model output with in-game observed data.

In order to compare our model to actual basketball, we need to *solve* the combinatorial game. Here, "solving" the game means determining the optimal strategy at any point in time based on the exact game state, including score, possession, time remaining, and shot clock remaining. In turn, the solved model will determine various important pieces of information, including: the in-game probability of winning, when to shoot, when to foul, what type of shot to make, and also, how long each possession should last.

We can solve this game by reverse engineering each move to determine for each possible game state which of the options is best. This is done by creating a large game tree through the probabilistic minimax theorem (or "expectimax" theorem) from artificial intelligence and game theory [4]. This approach is similar to mapping out all possible tic-tac-toe games in order to determine that tic-tac-toe is always a draw under optimal play. Except in our case, there is a probability for each step and there are many more possible game states. In fact, the challenge with our modeled game is that the game tree is locally exponential, as each possession has over 100 potential outcomes. Therefore, determining the optimal outcome over several possessions can quickly become computationally intensive. For these reasons, we will typically limit our analysis to the final minutes of a game.

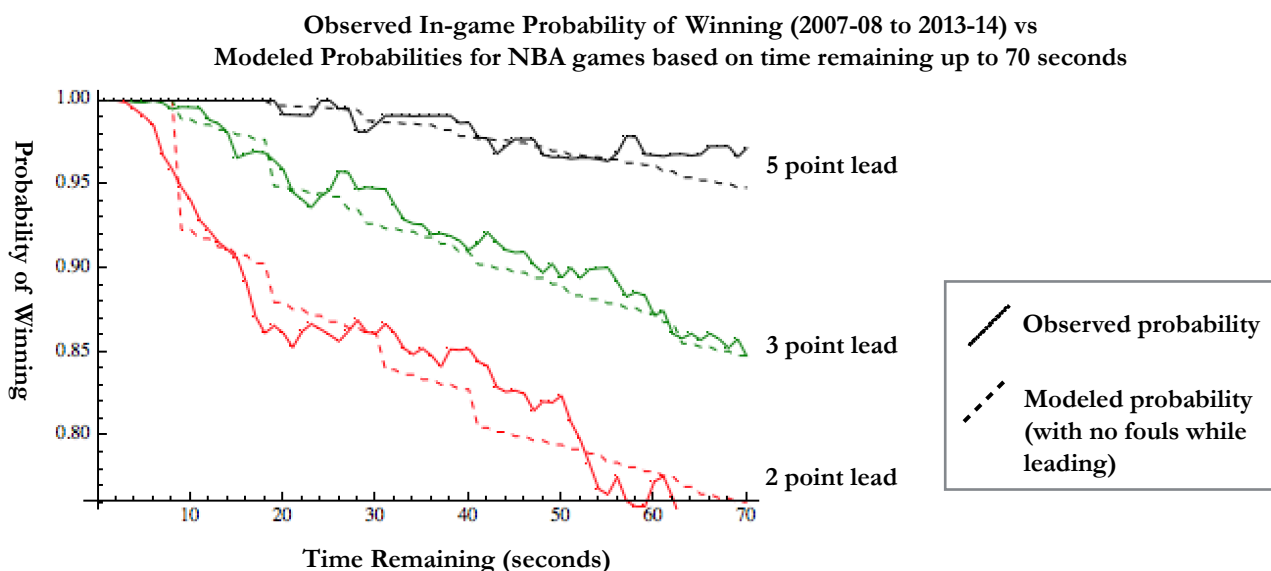


Figure 1: A comparison between observed in-game probabilities for winning and the modeled probabilities of winning based on point lead and time remaining. Here, modeled probabilities use an adapted combinatorial game which disallows the leading team from fouling.

After solving the game, we can take the in-game probabilities of winning from the solved model and compare them with observed data. Using widely available play-by-play data from the NBA seasons 2007-08 through 2013-14, we can compute observed in-game probabilities of winning based on the time remaining and score difference. The comparison to our solved model is quite remarkable. In Figure 1 above, we make this comparison for various scenarios. Here, the solid lines correspond to the observed data and the dashed lines correspond to our solved model. As can be seen, the observed data very closely follow the solved model predictions. Notably, the model is able to capture sudden and sharp drops in and around 15 to 20 seconds. Hence, the solved model indeed closely reflects true play.

There are two important things to note. First, the solved game in this comparison is modified by not allowing the leading team to foul, reflecting how NBA teams currently play. As we will show later, the leading team can improve its probability of winning by fouling at select points in time. Additionally, the modeled probabilities are not statistical regressions nor are they based on observations at all; rather, they are output of the simple combinatorial model. This demonstrates an advantage of combinatorial methods, as the model is able to capture both the short-term and long-term behavior of the win probabilities in basketball.

4 When To Foul

Common knowledge says it is typically unwise to foul the offense intentionally in order to force the free throws. This is true in the model as well. Depending on the parameters of the model, the average number of points per possession is 0.95 to 1.08 when shooting, but 1.40 to 1.51 when free throws are awarded. So, on average, free throws are worth approximately 50% more. However, at some point, time becomes the enemy, as the trailing team may not have enough time to mount a comeback. This is where fouling comes into play. By fouling, the trailing team limits the leading team's possession time. In effect, a single foul trades 0.5 points, on average, for an additional approximate 20 seconds of possession time (or 30 seconds in NCAA play).

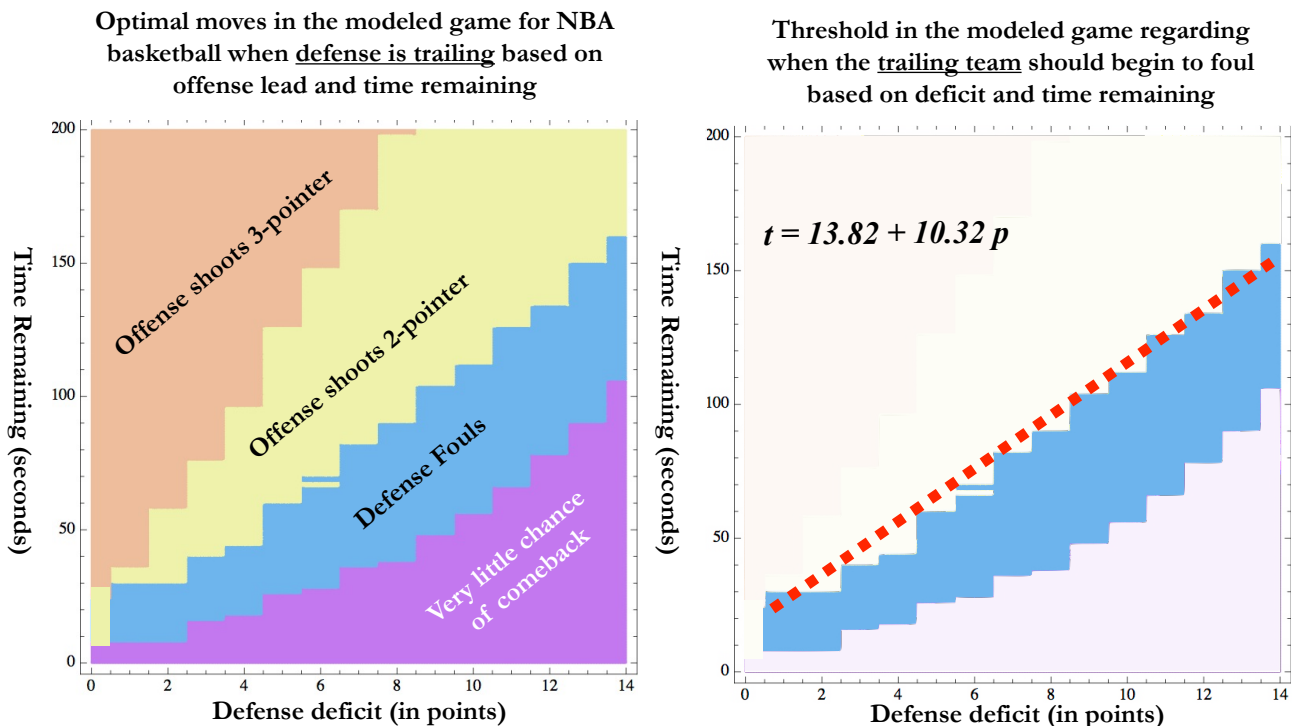


Figure 2: Plots indicating the optimal plays while defense is trailing for NBA basketball based on score and time remaining. Observe the near-linear threshold (right) for the earliest point at which a team should foul.

We can get insight into this issue by looking at the results of the solved game. Initially, we considered only the probability of winning the game at any one point under optimal play. However, upon solving the game, we can also determine which move is best at any point in time. As a result, we can determine when it is optimal to foul.

The results of this analysis are summarized in Figure 2 above. The threshold for when to foul in typical NBA play is very pronounced in the model and is a near-linear function of the point spread. In fact, for NBA play, a team trailing by p points should foul with approximately $t = 13.82 + 10.32 p$ seconds remaining. This result is achieved by applying a linear regression to the data in Figure 2, and it provides an extremely good fit for larger deficits. However, it does not necessarily apply to close games where the point difference is 1 or 2, in which case, the trailing team should foul with 30 or fewer seconds remaining.

In context, a team down by 5 points should foul when there are roughly 45 seconds remaining in the game, and a team trailing by 8 points should begin to foul with approximately 90 seconds remaining. This is much earlier than most NBA teams ever consider fouling.

Additionally, the combinatorial game indicates there is also a near-linear threshold for when teams should focus on 2-point shots over 3-point shots, as seen in Figure 2. This is consistent with previous work by Goldman and Rao, who arrive at a similar conclusion using a different approach [5].

As may be expected, the results change slightly when considering other levels of play. For typical NCAA Division I men's basketball, teams should foul earlier. This is largely due to the longer shot clock, but is also attributed to the slightly lower 3-point shooting percentage. Using the techniques above, we can determine that for NCAA basketball, a team behind p points should foul when there are roughly $t = 10.73 + 16.15 p$ seconds remaining. However, for close games with a 1- or 2-point difference, the trailing team should foul when there are 40 or fewer seconds remaining.

To estimate the marginal benefit of fouling early, we compare the model with an adapted one. For the adapted model, we modify the combinatorial game to disallow fouling when there are more than 30 seconds remaining. We then can compare the in-game probabilities of winning between the two models.

As shown in Figure 3 below, employing a strategy which fouls while behind earlier than current practice can

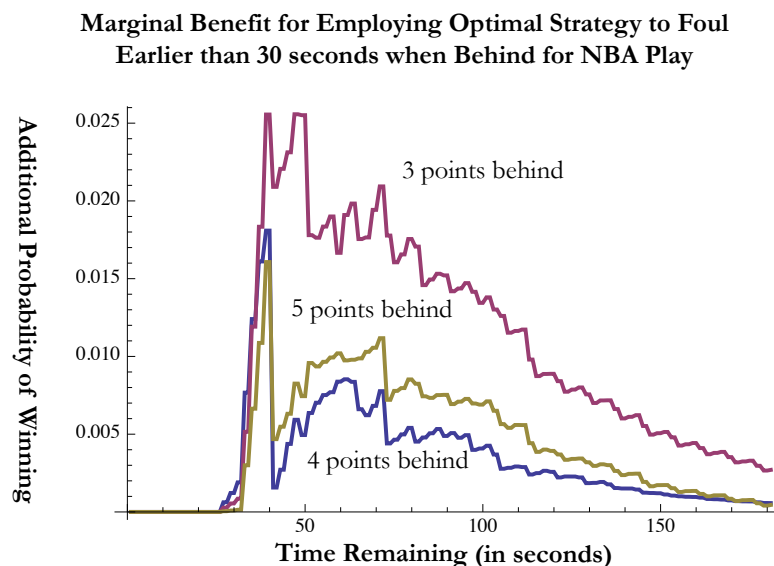
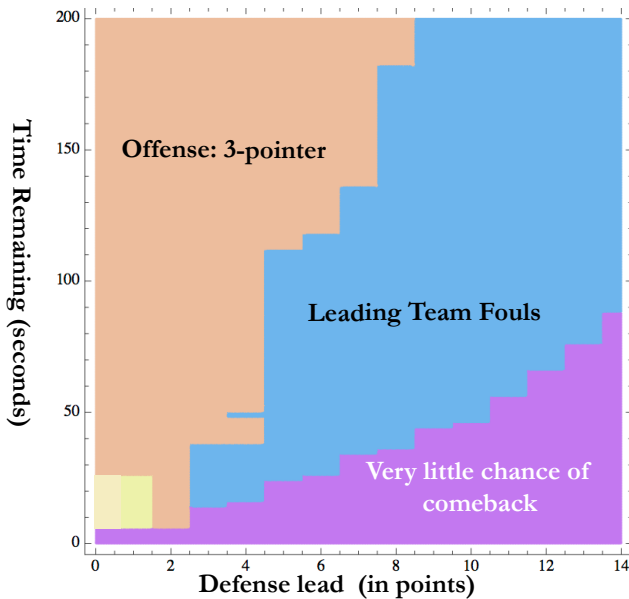


Figure 3: The difference in the in-game probability of winning between two adapted models of the combinatorial game: the original model and one disallowing fouling with more than 30 seconds remaining.

Optimal moves in the modeled game for NBA basketball when *defense is leading* based on defense lead and time remaining



Threshold in the modeled game regarding when the *leading team* should begin to foul based on lead and time remaining

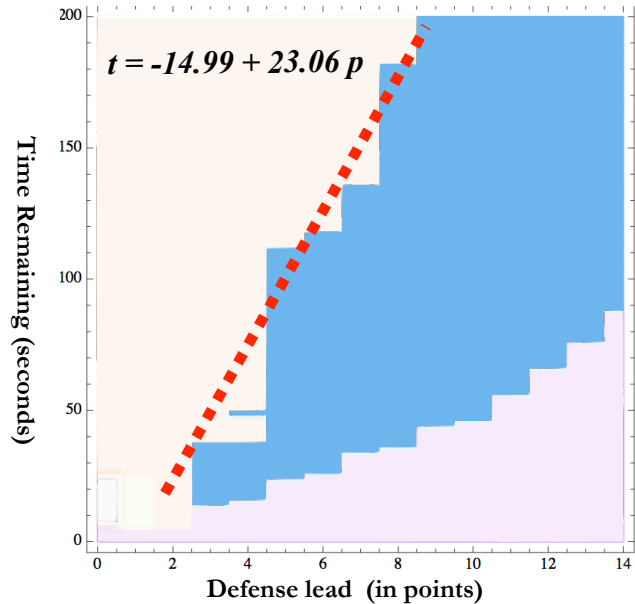


Figure 4: Plots indicating the optimal plays while defense is *leading* for NBA basketball based on score and time remaining. Observe the near-linear threshold (right) for the earliest point at which a leading team should foul.

substantially increase a team's chance of winning. For example, if one is willing to foul with more than 30 seconds remaining, the chance of winning with a 3-point deficit with one minute remaining increases by approximately 1.8%. This may not seem substantial, but when considering that the observed probability of winning in such a situation is only 23%, a 1.8% increase is, indeed, substantial. For emphasis, this does not mean that one should foul when down by 3 points with one minute remaining. In fact, Figure 2 indicates one should not foul in such a situation. Rather, if a team is down by 3 points but *is willing to foul* in any later scenario that justifies it (and not just within 30 seconds remaining), it retains a higher chance of winning.

One of the most surprising outcomes of our model demonstrates that the *leading* team should foul **more often** than the trailing team. Recent discussion among sports analysts supports the idea of fouling when ahead by three points during the final seconds. However, our model goes further, suggesting that teams foul earlier with larger leads as well. Using the same techniques as before, we not only determine when the leading team should foul, but also its benefits.

The results of our analysis with regard to the leading case are summarized in Figure 4 above. As with the trailing case, the threshold for when the leading team should foul is also a near-linear function of the point difference. For typical NBA play, a team leading by p points should foul with approximately $t = -14.99 + 23.06 p$ seconds remaining. While this indicates that the leading team should not foul when only leading by 1 or 2 points, it means that for larger leads, the leading team should begin to foul before the trailing team.

To see the benefit of fouling while leading, we can compare the original model with an adapted one that does not allow the leading team to foul. This is similar to our comparison before with fouling earlier while trailing. The benefits here are indeed enormous and summarized in Figure 5 below. For instance, if a team is willing to foul when leading by 3 points, it can increase its chances of winning by slightly more than 10%. This may sound far-fetched, but it is in line with previous analysis which estimates a 9-10% gain from fouling when up by 3 points [1].

The concept of fouling when ahead may be counterintuitive. However, toward the end of the game, the main goal of the trailing team is to increase the total variance in order to widen the window of possibilities that win the game.

Marginal Benefit of Employing Optimal Strategy to Foul when Ahead for NBA Play

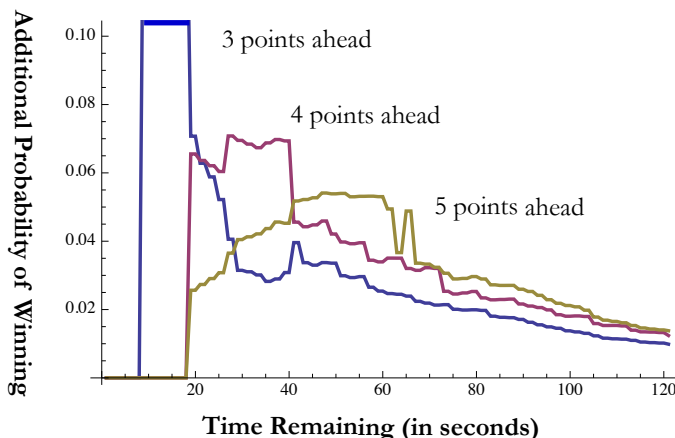


Figure 5: The difference in the in-game probability of winning between two adapted models of the combinatorial game: the original model and one disallowing fouling while ahead. Fouling while ahead can increase the leading team's chance of winning by up to 11%.

One main component in this wider variance is the riskier 3-point shot. The trailing team can limit this variance by fouling. the leading team may give up points, on average, but limit the trailing team to 2 points per possession. This decreases the total variance and, with a sufficient lead, increases the leading team's chances of winning.

5 Impact for NBA Teams

Is there an actual example to show that implementing the proposed strategy would result in substantial benefit? Further, are these strategies beneficial over the long term or are they specific to rare scenarios?

To answer these questions, we use play-by-play data from various NBA seasons to estimate the additional number of wins a team can earn that season. We apply the results of our model by considering the game state of each individual game with two minutes remaining. By using the model to determine the probability of winning at that point in time, we can estimate how many total games, on average, each team can win by playing an optimal endgame. We then compare this quantity to the actual number of wins that season.

The answer to the questions is that the long-term impact of optimal endgame play can be substantial. One interesting example is the 2009-10 NBA regular season where the Indiana Pacers missed the playoff by 9 games and the Toronto Raptors missed the playoffs by 1 game. That season, the Pacers lost many close games. In fact, based on the scores at two

Estimated Additional Wins under Optimal Play over the 2009-2010 NBA Season for Teams Near .500 Win%

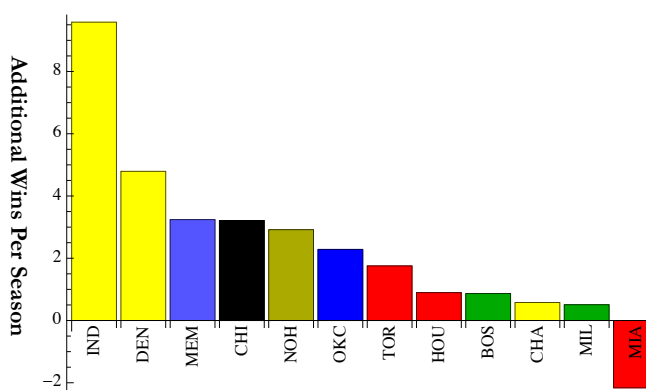


Figure 6: In most cases, the impact of implementing optimal endgame play over the course of an NBA season will add at least 1 game win per season, and, in many cases, more.

minutes remaining, the Pacers could have won an expected 9.58 additional games under optimal play. The Raptors could have won 1.70 additional games. In both cases, the additional wins would have been enough games to earn a playoff spot. Even for many other “bubble” teams, using optimal play adds an expected 1-2 wins per season. These results are summarized in Figure 6.

To put this in perspective, the typical NBA player is valued around 1 to 2 wins per season above a replacement player. However, using an optimal strategy does not count toward the five players allowed on the court. Hence, employing an optimal endgame strategy is akin to having a sixth player on the court.

6 Why Not Just Statistics?

Note that our approach does not use statistical methods in a conventional manner. In fact, the main conclusions of this paper are based on combinatorial game theory, as opposed to statistical theory. Even so, one may want to validate these conclusions using statistical methods and hypothesis testing. However, doing so is easier said than done.

From a statistical standpoint, to see whether or not the optimal combinatorial strategy is valid for actual play, one would need sufficiently many games with perfectly timed incidental fouls. Such games are extremely rare. The main advantage of the leading team fouling is to limit the total number of points the trailing team can score. Fouling once simply gives the trailing team more points on average without limiting the potential for a comeback. Fouling successively places a severe limit on the number of points the trailing team can score in the time remaining. Hence, in order to take full advantage of fouling while ahead, it must be done in a consistent and methodical manner. The strategy must be taken and analyzed as a whole. Therefore, we must conclude that the optimal strategy presented here most likely would have never been uncovered using only statistical methods.

Beyond describing when to foul in basketball, a key takeaway of this article is that statistics cannot necessarily uncover all optimal strategies in sport. This is especially true for strategies such as the one presented in this article.

7 Discussion

In recent times, the NCAA has been considering changes in order to limit or eliminate the foul-centric basketball endgame [6]. While there are several ideas and proposals to change the game, in order for the game to become less about fouling, there are two things to consider. First, when behind, the advantage of fouling comes from stopping the clock at the expense of giving up points to the opposing team. Second, when ahead, the advantage of fouling comes from limiting the opponent from scoring three-points in one possession. If the aim is to eliminate the foul-based endgame, we offer the following suggestion. If a team shoots free throws as a result of being fouled, that team should have an accessible opportunity to earn three points, perhaps by earning a third free throw if it makes the first two. In fact, “bonus” free throws were introduced early in the history of basketball as a way of punishing a team for excessively fouling [7]. However, at that point in time, there was no three-point shot. With the addition of the three-point shot, it is reasonable to allow the offense to score all three available points if fouled. In fact, using a modified combinatorial game incorporating the suggestion above, the leading team should almost never foul, and the trailing team should foul approximately half as often as we conclude in Section 4.

8 Conclusion

In this article, we not only have modeled basketball as a combinatorial game and demonstrated this model closely mimics the strategic choices in basketball, but also used these results to determine when teams should foul during the endgame. The results suggest a new improved strategy involving fouling earlier, more often, and also while in the lead. Further, this strategy could not have been discovered with statistical methods.

The techniques used here are not limited to basketball and may be translated to most other sports. For example, future work could apply this approach to determine when football teams should perform an onside kick, when hockey teams should pull their goalie, or even when tennis players should attempt a “winner.”

9 Acknowledgments

The author would like to explicitly thank Christine Lew and Waldemar Stronka for their comments on this paper.

10 References

- [1] D. H. Annis, "Optimal end-game strategy in basketball." *Journal of Quantitative Analysis in Sports* vol. 2, no. 3, 2006.
- [2] K. Pomroy, "Yet another study about fouling when up 3," kenpom.com.
- [3] J. Ezekowitz, "Up Three, Time Running Out, Do We Foul? The First Comprehensive CBB Analysis," Harvard Sports Analysis Collective.
- [4] S. J. Russell and P. Norvig, *Artificial Intelligence: A Modern Approach*, 2009.
- [5] M. Goldman and J. M. Rao, "Live by the Three, Die by the Three? The Price of Risk in the NBA." 7th Annual MIT Sloan Sports Analytics Conference, 2013.
- [6] R. Glier, "Does the End Need Rewriting?" *New York Times*. March 24, 2010.
- [7] J. Naismih, *Basketball: Its Origin and Development*, U. of Nebraska Press, 1941.

A Additional Figures

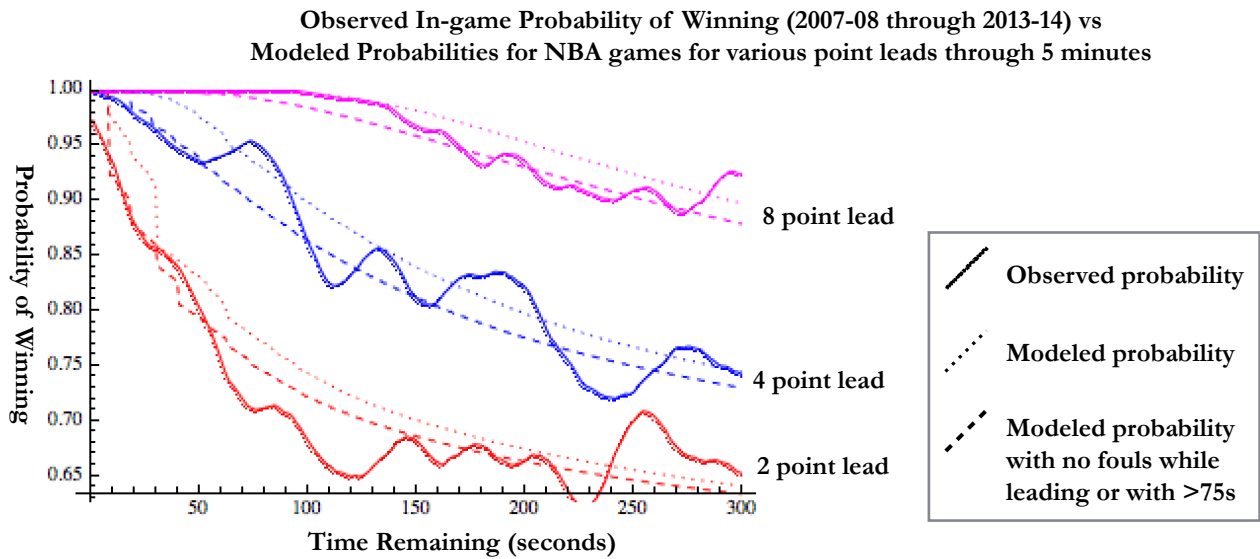


Figure A.1: A comparison between observed in-game probabilities for winning and the modeled probabilities of winning based on point lead and time remaining. Included are the optimal modeled probabilities and the modified modeled probabilities when the combinatorial game is adapted disallowing the leading team from fouling.

Optimal Moves in NCAA Basketball under Proposed Rule Change allowing up to 3 Potential Free Throws per Foul

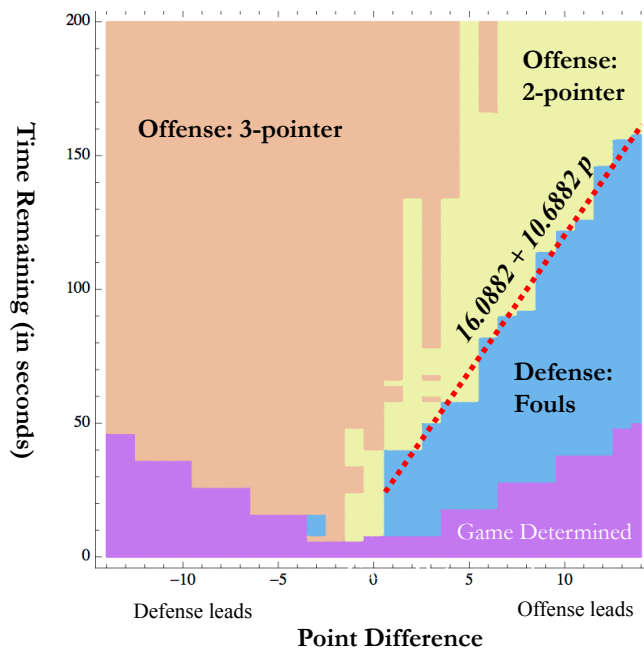


Figure A.2: Plots indicating the optimal plays for a modified NCAA Division I basketball where the fouled team gets a third free throw if they make the first two. In this case, the leading team should almost never foul, and the threshold for fouling is $16.0882 + 10.6882p$. For larger p , this threshold is approximately 50-60% of the threshold mentioned in Section 4 ($10.73 + 16.15p$) under the current rules.

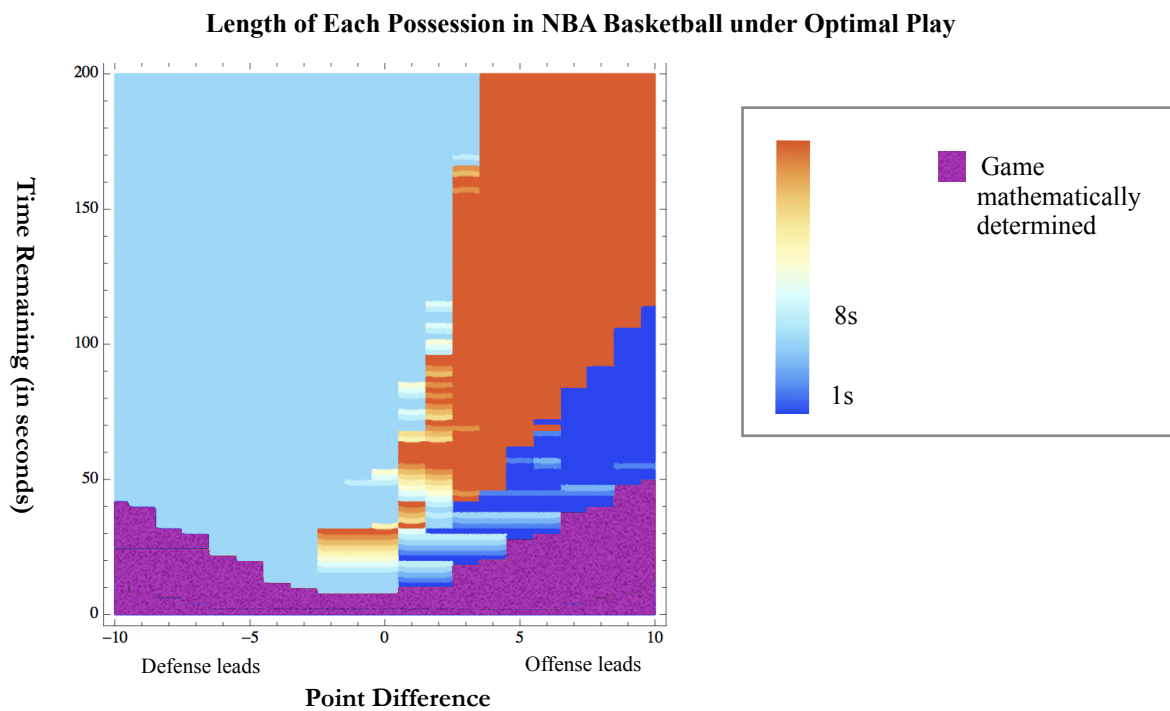
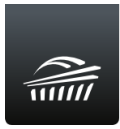


Figure A.2: A density plot indicating the length of each possession within the combinatorial model under optimal play. Blues indicate faster play, reds indicate slower play, and purple indicate that the outcome of the game is effectively determined.