Sports Rankings REU Final Report 2012: An Analysis of Pairwise-Comparison Based Sports Ranking Methods and a Novel Agent-Based Markovian Basketball Simulation

Daniel Barrow, Ian Drayer, Peter Elliott, and Garren Gaut
Faculty Mentor: Braxton Osting

August 10, 2012

1 Introduction

There are numerous methods to rank sports teams. Each of these methods incorporate different components to determine a ranking system. For example, the Rating Percentage Index (RPI) method considers the weighted average of a team's winning percentage and its opponents winning percentage [16]. The Elo method assigns each team a rating and when two teams play each other, the rankings change depending on the rating of their opponent and the outcome of the match [10]. A Random Walker method casts a vote for a team to be the winner and the ordering of the number of votes deduces a ranking system [4]. A Hodge Rank determines the ranking of teams by minimizing the error of wins and losses [14][12]. The Bradley-Terry model uses a probabilistic approach to deduce a team ranking[2]. These various ranking methods are generally consistent with differentiating the best teams from the worst teams, but are less consistent when ranking teams whose rankings are closer together.

There still remain a number of unanswered questions about the relation between different ranking methods: which ranking methods most accurately predict the true rankings of teams in a given sport and how much influence do scheduling and number of games played have on the accuracy of each ranking method? The REU report is organized as follows. In Section 2 we discuss several commonly used ranking methods. In section 3 we use cross-validation to analyze the accuracy of the ranking methods discussed in Section 2. In Section 4 we describe the BODGE model and BODGE simulation, an agent-based Markovian basketball model and simulation. In Section 5, we discuss a probabilistic geometric interpretation inconsistencies and champions of a tournament. In Section 7 we invent a new baseball statistic that measures the offensive value of a batter. Section 8 we describe a future direction that we would like to take our research.

2 Ranking Methods

In our meritocratic society the concept of rank is paramount. Consumers seek the best product, sports fans want to know the best sports team and search engines need to know the best website suited for a particular need. The need for rankings in various fields has led to the development of many different algorithms for each field and particular need. In our paper we apply many of these methods to the ranking of sports data. However, it is important to understand some of the impetuous for the development of each individual ranking method.

Ranking methodology is directly linked to situational factors, including but not limited to the number of the items to be ranked, availability of comparison data, number of comparisons between items and time period over which items are ranked. The least squares method on a graph [12] uses pairwise data to minimize the difference in rating and observed pairwise comparisons to obtain a ranking, resulting in a fast ranking method which always has an exact analytical solution. Various least squares methods have been developed that differ in the calculation of the pairwise comparison data [18] [5]. Other methods use different data to create a ranking, such as the PageRank [21] developed by Larry Page at Google. PageRank uses a random walker to analyze the structure of links between websites resulting in a ranking of website by time the walker spends at each website, called PageRank. Arpad Elo, a professor and avid chess player developed the Elo ranking method [8] for chess which rates players based on the strength of their opponent by calculating changes to a rating as a function of the difference in rating between two players. Microsoft built upon Elo method by changing the distribution from which changes to ratings are drawn and applied a new method, TrueSkill [11], to multi-player on-line gaming. Rating Percentage Index (RPI) [23] is a sports ranking method based on winning percentage; the RPI score is a weighted function of a teams' winning percentage, teams' opponent's winning percentage, and teams' opponents' opponents' winning percentage.

In this section, we discuss each of these methods in detail. For simplicity and illustration we assume for each method that we are ranking sports data.

2.1 Least Squares

Least squares ranking is equivalent to solving a least squares minimization on a graph where each node is a team and each edge is a pairwise comparison between teams. Let B be the edge incidence matrix where

$$b_{ij} = \begin{cases} 1 & : i > j \\ -1 & : j > i \\ 0 & : otherwise \end{cases}$$

Let ϕ be a rating of teams which will generate a ranking and y be a vector of pairwise comparison data. Unweighted least squares is equivalent to solving the minimization

$$\min_{\phi} ||B\phi - y||_p$$

where p indicates the l_p vector norm and $B\phi$ is a vector. Teams can be also be weighted by a matrix W where w_{ij} represents some form of weight between teams i and j. Possible weights include number of games played, time point in the season or home/road game weights. Then our least squares problem is

$$\min_{\phi} ||B\phi - y||_{p,W}$$

where W indicates that we minimize using

$$\min_{\phi} (y - B\phi)^t W (y - B\phi),$$

for $p = 2$, or

$$\min \sum_{ij} w_{ij} |B\phi - y|_{ij}^p$$

for arbitrary p.

Various least squares ranking methods can be implemented by varying the choice of norm and method for calculation of pairwise comparison data. In our implementations we use the euclidean, absolute and infinity norms and calculate pairwise comparison data by calculating the season score differential between teams. Other pairwise comparison calculations include log score differential, win percentage[18] and win percentage with prior information[5].

2.2 PageRank

Page Rank [21] calculates a ranking by walking between nodes of a graph and finding the node at which the random walker spends the most time. At any given node i, the walker has probability p_{ij} of transitioning to node j and the probability p of transitioning to a random node (drawn from a uniform distribution). In sports, we represent each team as a node [4]. Let i and j be two teams. Let s_{ij} be the points i scores against j divided by the total number of points scored in the game. Let p_{ij} be the probability the walker transitions from team i to team j. For all i we calculate the probability a walker remains stationary,

$$p_{ii} = \sum_j \frac{s_{ij}}{o_i}$$

where o_i is the number of opponents of team i. We calculate each p_{ij} by solving for K in the equation:

$$K(\sum_j \frac{s_{ij}}{o_j}) = 1 - p$$

and setting $p_{ij} = K\frac{s_{ij}}{o_j}$. We create a transition matrix T where $t_{ij} = p_{ij}$ and take it to the m th power where m is large to get a stationary distribution of time spend at each node.

2.3 Elo

The Elo method [8] initializes each players rating at 1500. Ratings range from 100 to over 2800. For game between two players or teams, Elo calculates an expected win percentage. Suppose we have two teams A and B with ratings r_A and r_b and game outcomes s_A and s_B. We calculate team A's expected score as

$$E_A = \frac{1}{1 + 10^{(r_B - r_A)/400}}$$

We then update player's scores from the formula

$$r_{Anew} = r_{Aold} + K(s_A - E_A)$$

where K is a constant. In chess K varies from 16 to 32.

2.4 TrueSkill

Microsoft's TrueSkill algorithm [11] ranks players in multi-player online gaming by assigning each person a skill level modeled as a normal distribution. Player performance levels are also modeled from a normal distribution. In each game a teams performance level is the sum of individual player performance level. After a game has been played, TrueSkill updates player skill levels based on a function of the game outcome. In our implementation, we modified TrueSkill only to account for teams rather than individual player performances which resulted in a greatly simplified model. We initialized each team with a rating of 25 and a variance of $(\frac{25}{3})^2$ so that 99% of the ratings are within three initial standard deviations of initial ratings. Suppose we have two teams A and B each with a mean skill levels μ_A and μ_B and skill variances ${\sigma_A}^2$ and ${\sigma_B}^2$. After each game we update the winner's and loser's mean skill and skill variance as follows. Let μ_W and μ_L and σ_W and σ_L represent the ratings and standard deviations of the winning and losing teams in a game respectively. Let

$$v(t,e) = \frac{\phi(t-e)}{\Phi(t-e)} \text{ and } w(t,e) = v(t-e) * [v(t-e) + t - e]$$

where $\phi(x)$ is the standard normal distribution and $\Phi(x)$ is the standard normal cumulative distribution. Then we update rating means and variances as follows:

$$\mu_{Wnew} = \mu_W + \frac{{\sigma_W}^2}{c} * v\left(\frac{(\mu_W - \mu_L)}{c}, \frac{\epsilon}{c}\right)$$

$$\mu_{Lnew} = \mu_L - \frac{{\sigma_W}^2}{c} * v\left(\frac{(\mu_W - \mu_L)}{c}, \frac{\epsilon}{c}\right)$$

$${\sigma_{Wnew}}^2 = {\sigma_W}^2 * [1 - \frac{{\sigma_W}^2}{c^2} * w\left(\frac{(\mu_W - \mu_L)}{c}, \frac{\epsilon}{c}\right)]$$

$${\sigma_{Wnew}}^2 = {\sigma_L}^2 * [1 - \frac{{\sigma_L}^2}{c^2} * w\left(\frac{(\mu_W - \mu_L)}{c}, \frac{\epsilon}{c}\right)]$$

To calculate a ranking, we calculate the rating of each team over an entire season of games.

2.5 RPI

RPI [23]uses a set of winning percentages to calculate a teams RPI. Let *wp*, *owp* and *oowp* be a team's winning percentage, a team's opponent's winning percentage and a team's opponent's opponent's winning percentages respectively. We calculate a team's RPI as follows:

$$RPI = 0.25 * wp + 0.5 * owp + 0.25 * oowp$$

Winning percentage is calculated as the ratio of games won to games played.

The availability of various ranking methods leads us to question which methods are best in which scenarios. In this paper we examine the performance of each method when applied to sports datasets using cross validation.

3 Cross-Validation of Ranking Methods

The primary facet of performance that we try to evaluate is the predictive accuracy. To do this, we need to choose a notion of error for a ranking. In general, there are many possible ways of doing this. For example, if the goal of the ranking is to predict margins of victory, error could be defined as the distance from the predicted margin to the actual margin in a game. In our case, we simply try to predict who will win each game, so we define an error as a game where a lower ranked team wins. Then the total predictive error is the proportion of games for which the lower ranked team wins. The best performing ranking method is the one that has the lowest expected predictive error.

We use cross-validation to estimate the expected predictive error for each method described in Section 2 . In particular, 10-fold cross-validation is used to partition the data into 10 subsets, and each subset is used once to test the accuracy of a ranking generated from the other nine subsets. The average error across the 10 subsets is used as a cross-validation score. This process is repeated 100 times to get a distribution of scores. Another popular cross-validation method is the holdout. For that method, the data is divided into only two subsets, with one used for ranking and the other used to test the results. Holdout cross-validation is less computationally expensive than k-fold (the general case of 10-fold), but results show that it is less accurate [1].

3.1 Results

We have used cross-validation to test the results of each ranking method on 10 seasons of results from Major League Baseball (MLB), 6 seasons from the National Basketball Association (NBA), 10 seasons from NCAA Division 1-A Football (NCAAF), and 1 season from NCAA Division 1 Basketball (NCAAB). These results can be displayed with a boxplot for each method and each season showing the distribution of cross-validation scores (Figures 1 through 4). Observations and conclusions are discussed in the Hypothesis Testing section.

3.2 Hypothesis Testing

After collecting cross-validation scores for each method, we test whether the results indicate a statistically significant difference in performance between ranking methods. This requires two tests: the Friedman test and the post-hoc Nemenyi test.[6]

The Friedman test is used first to check the hypothesis that the distributions of the scores are the same for all of the ranking methods. This is necessary because performing multiple

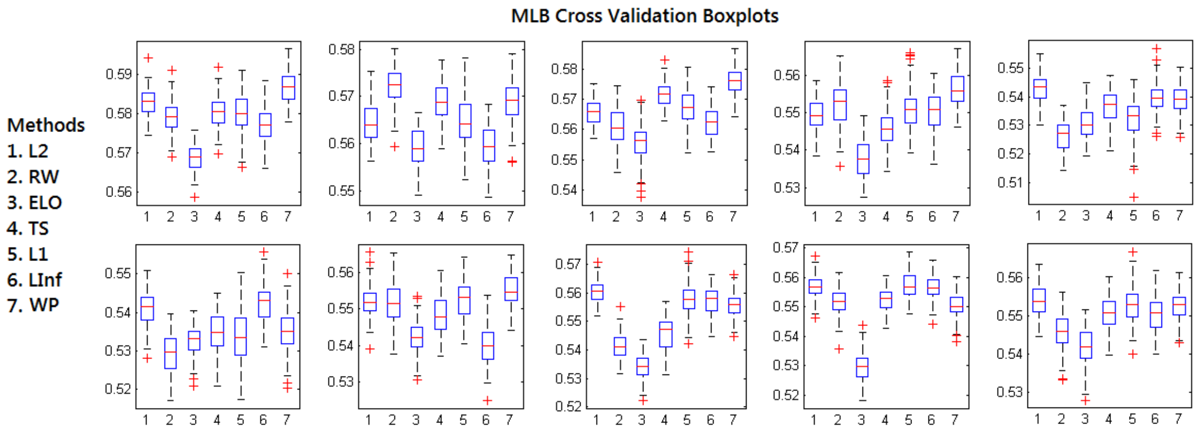

Figure 1: MLB Cross Validation Scores

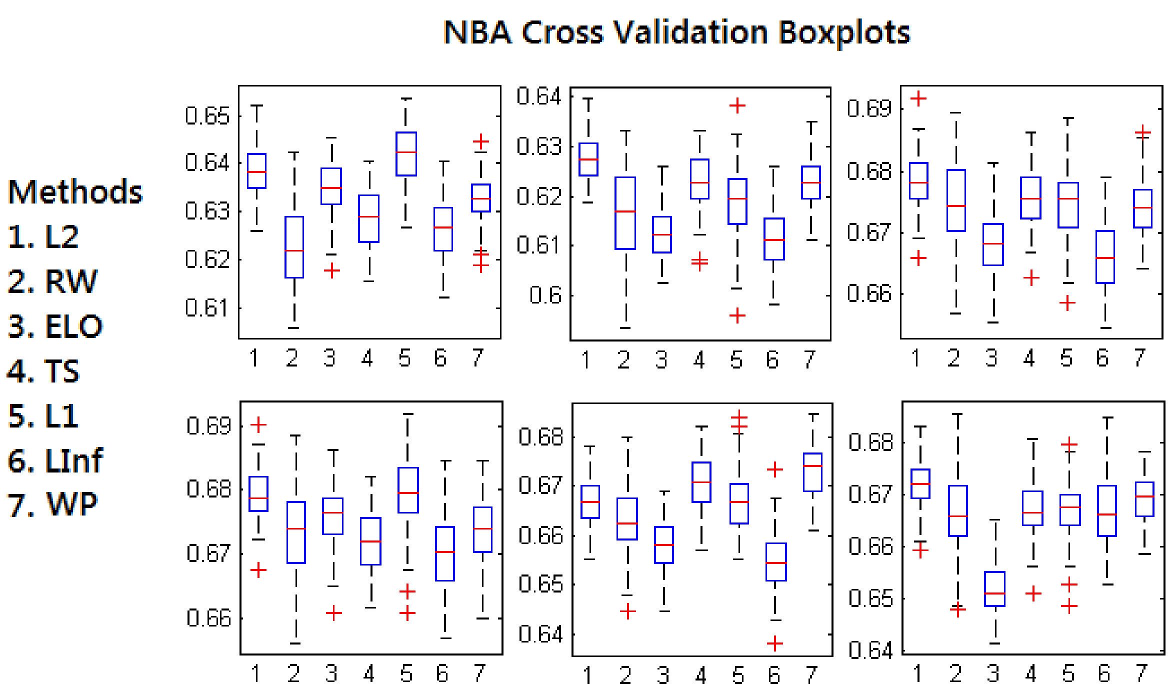

Figure 2: NBA Cross Validation Scores

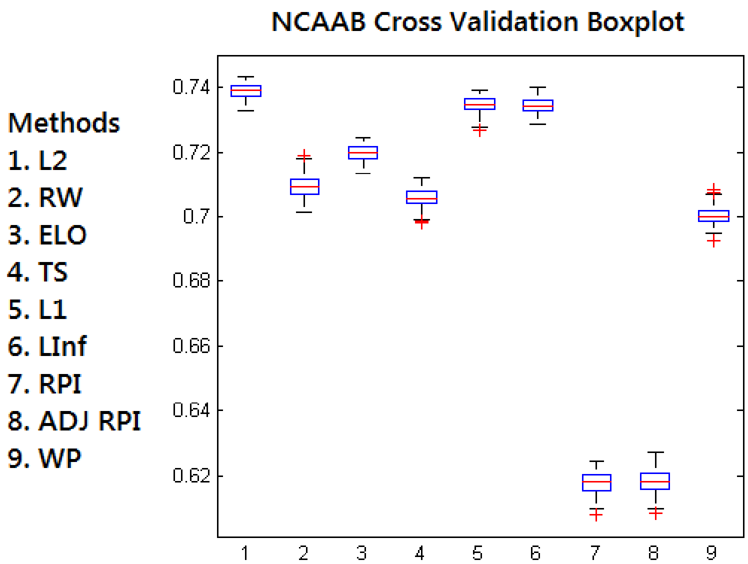

Figure 3: NCAA Basketball Cross Validation Scores

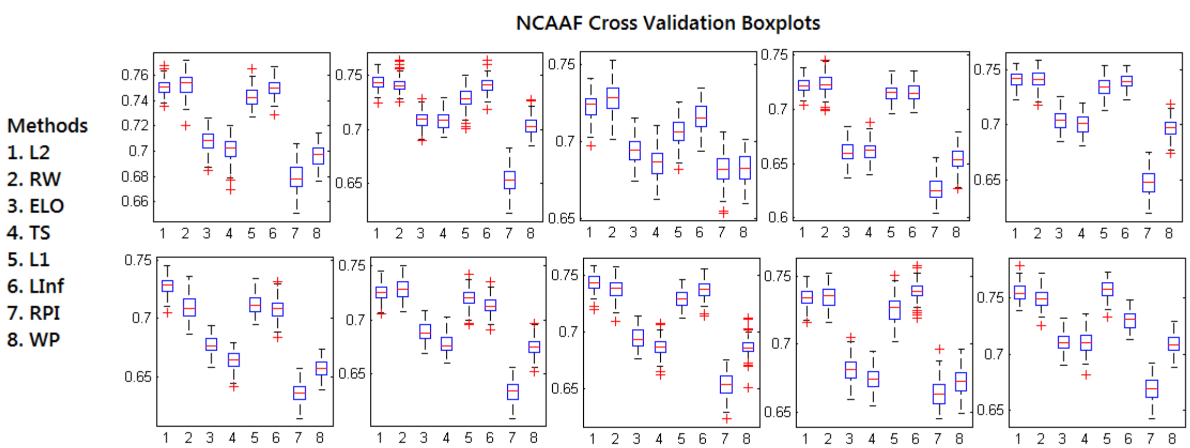

Figure 4: NCAA Football Cross Validation Scores

comparisons simultaneously can significantly elevate the probability of at least one type I error occurring (assuming that the null hypothesis is true). The Friedman test works by ranking the methods across each partition, finding the average ranking for each method, and calculating the following statistics (where k is the number of methods and N is the number of repetitions):

$$\chi_F^2 = \frac{12N}{k(k+1)}\left[\sum_j R_j^2 - \frac{k(k+1)^2}{4}\right] \tag{1}$$

$$F_F = \frac{(N-1)\chi_F^2}{N(k-1)-\chi_F^2} \tag{2}$$

F_F follows the F distribution with $k-1$ and $(k-1)(N-1)$ degrees of freedom. If F_F is larger than the critical value for the chosen significance level, then we can reject the null hypothesis.

If the Friedman test rejects the null hypothesis then we move on to the Nemenyi test, which allows us to make pairwise comparisons between each ranking method. We define the critical difference (CD) by:

$$CD = q_{\alpha,k}\sqrt{\frac{k(k+1)}{6N}}$$

Where $q_{\alpha,k}$ is drawn from the Studentized Range Distribution and depends on the significance level α as well as the number of methods that are compared. We can conclude that two methods have a significant difference in performance if the difference between their average rankings from the Friedman test is greater than CD. These two tests gave the results in figures 5 through 8.

These results show that the ℓ^2 ranking is in all cases either the best or tied for best method. Conversely, the RPI is significantly worse than all other methods. This is interesting because the RPI is an official ranking method used by the NCAA. NCAA selection committees in multiple sports consider the RPI when choosing teams for postseason play [1]. We suggest that they use other more reliable methods for ranking.

[1] http://www.ncaa.com/sites/default/files/files/BracketPrin-Proc10-5-10.pdf

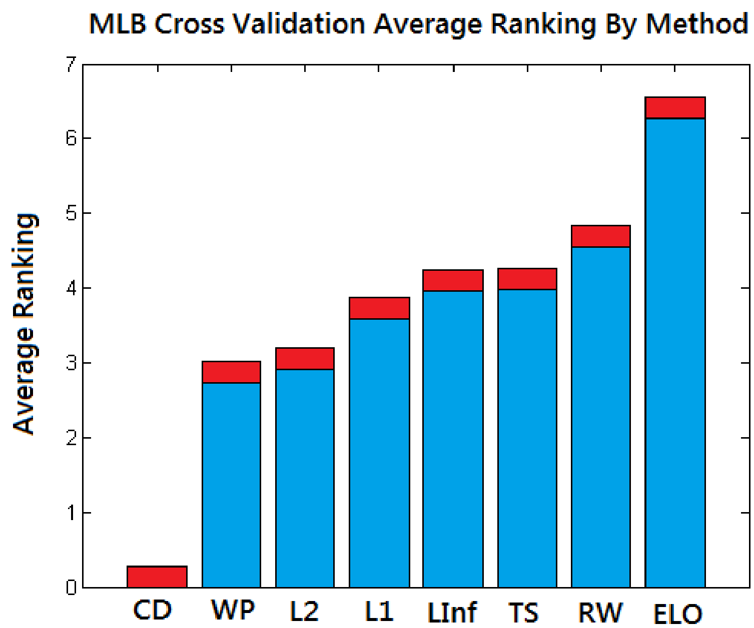

Figure 5: MLB Average Rankings

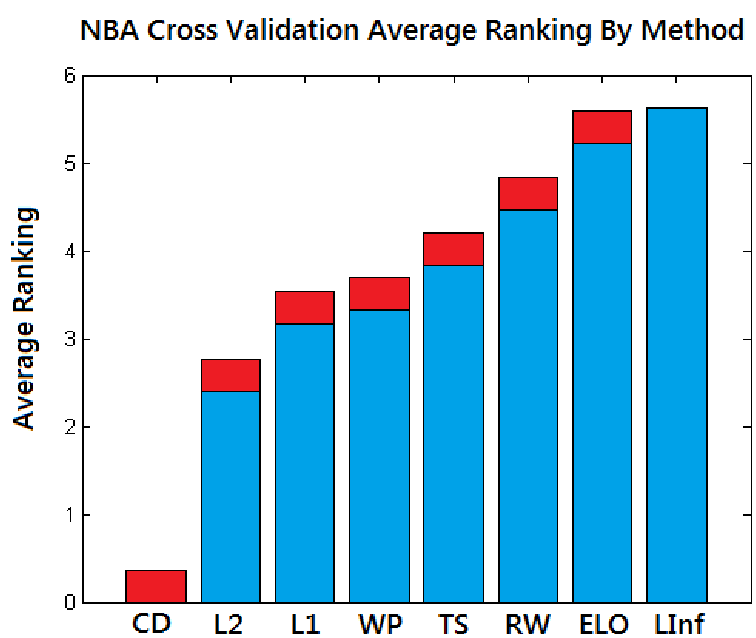

Figure 6: NBA Average Rankings

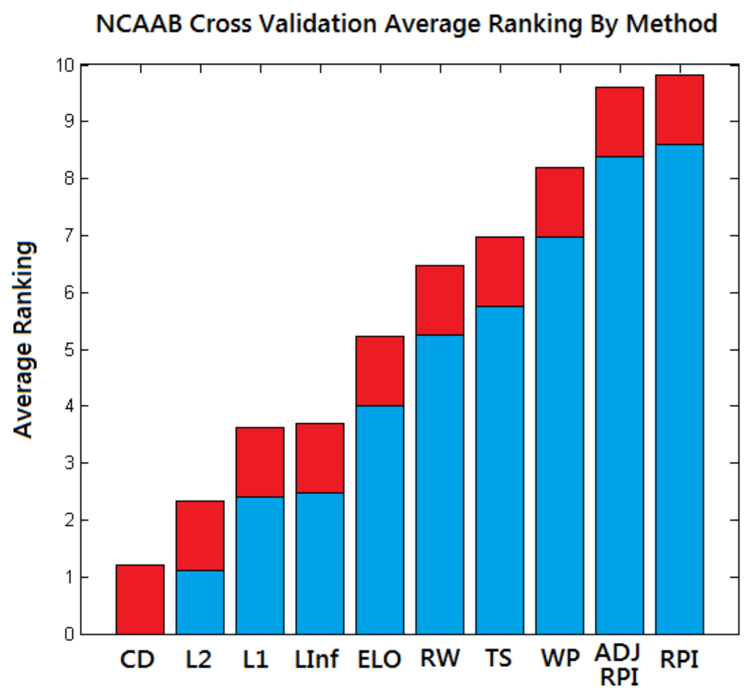

Figure 7: NCAA Basketball Average Rankings

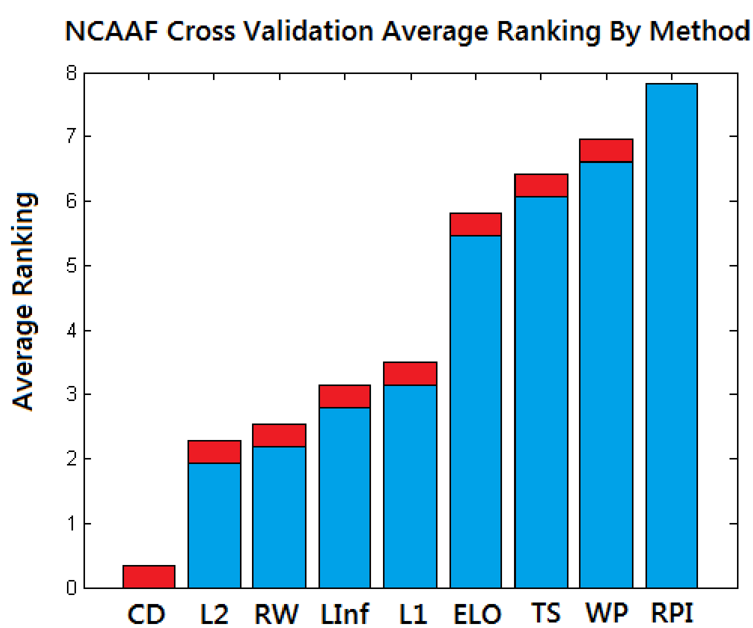

Figure 8: NCAA Football Average Rankings

4 The BODGE Model

4.1 Markov Chain Player Induced Team Rankings

In 2002, the Oakland Athletics organization conducted an experiment in which they relied on baseball statistics to build a baseball team, as opposed to the status quo of relying purely on scouts [17]. To many people's surprise, the Oakland Athletics experiment worked as they won over one hundred games during the regular season. Since then, sports organizations have considered hiring mathematicians and statisticians to do similar analysis as the $A's$.

We believe the main reason the Moneyball approach worked is due to the fact that baseball can be viewed as a series of discrete events relying on individualistic statistics. Since baseball consists of pitches, which lead to hits, which lead to plays, we can analyze each part of the game. It is much harder for most other sports to be broken into a series of discrete events as easily; when they are decomposed in such a way, several important information is lost. Essentially we want to know how well the Moneyball idea can be extended to sports other than baseball.

People have tried to model basketball teams with similarly methodology as Moneyball. Recently, researchers have been using Markov chains models to induce team rankings. In [22], Markov chains were used to rank teams, where each change of state implied a hypothetical voter that cast a vote for which team is the best . This ratio should converge to the expected probability that a team beats its opponent. Additionally in [22], Markov chains were used to model transition probabilities. In [9], a possession based Markov chain model was introduced. This model simulates possessions and relates team statistics to the transitions states. This play-by-play possession induces transition matrices between two teams, which in return provides team summary statistics. The problem with this model is the Markov chain overestimates the skills of the weaker teams' overall summary statistics.

One criticism which may result from team ranking methods studied in Section 2 is that these rankings only use macroscopic results. When analyzing from the macro perspective, we are only looking at the game outcomes such as wins versus losses, point differential, and team results. A different approach would be to analyze game data on the micro scale, i.e. to look at the smaller details that compose a game, such as player movement on the court or the outcome of each possession. Consequently, our paper will now switch its focus from team rankings based team statistics to team rankings induced from player statistics.

We now introduce a method for using player statistics and player tendencies to induce basketball rankings, which we call the **BODGE model**. The BODGE model allows for a simulation of a basketball game, in which we analyze the micro aspects of the game at discrete subdivisions of time. We then use a continuous Markov chain to link each of these micro details in each possession. A unique aspect of the BODGE model is that it models player motion via Markov chains. Currently, the BODGE does not incorporate continuous motion, but modeling with continuous motion could eventually be added through the use of image processing [13]. The goal of the BODGE model is to provide insight into a team's success and explain why a team with less raw talent could outperform another more talented team.

4.2 Features of the BODGE Model

The BODGE model requires 38 pieces of data to be collected for each player. Some data has already been collected and is easily accessible to the public, while other data must be collected by watching film footage and taking specific notes about how the player behaves on the court. We also invent two new statistical ideas, which should be implemented in such a model.

The data that we must collect for each player that is easily available to the public are:

1. normalized number of points scored
2. normalized field goal attempts
3. normalized free throw attempts
4. normalized free throws made
5. normalized assists
6. normalized rebounds
7. normalized turn overs
8. normalized number of shots blocked or contest
9. assist percentage
10. expected minutes played per game.

We must also collected data involving player motion on the court. Subdivide the basketball court into 5 regions, as shown in Figure 1. For each player, watch a sufficiently large random sample of minutes played on the court to collect:

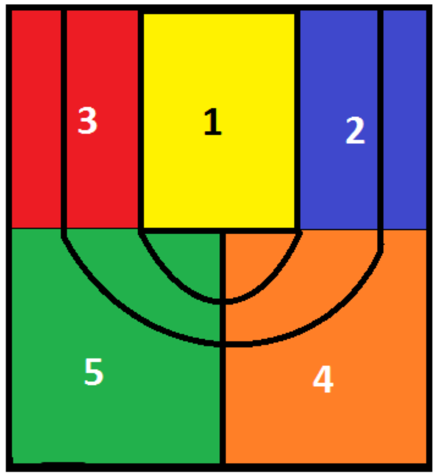

Figure 9: The 5 regions that the basketball court is divided into for the BODGE model in the construction of the player movement matrix T.

1. The 5×5 player motion matrix, whose rows represent the region a player starts in and the columns represent the region of the court that the player cuts to. The entries of this matrix correspond to the rate at which the player moves between regions. Twenty pieces of data are extracted from this matrix.

2. The probability that player starts each possession in each region.

3. The probability that the player will start with the ball at the beginning of each possession.

4. The amount of time between catching the ball and passing the ball.

5. The normalized number of times the ball is caught.

We now introduce some new statistics which is required by the BODGE model. Since we model player motion, it is essential to model how much a player demands the ball. Hence, we create a **demand ball factor (DF)**, defined by

$$\text{DF} = (\text{Expected Minutes/Game})[1.5(\text{normalized points}) + (\text{normalized assists})].$$

The BODGE model requires us to calculate a player's, **relative demand factor (RDF)**, which is defined by

$$\text{RDF} = \frac{\text{the player's demand coefficient}}{\Sigma\ (\text{all player's demand coefficient})}.$$

The BODGE model requires us to find a way to model the how well a player can pass the ball. To obtain a standard formula, we define this statistic to be the **quality of pass (QOP** to be

$$\text{QOP} = \frac{(\text{player's assist percentage})}{\Sigma(\text{all player's assist percentage})}.$$

The formula for QOP can still be improved, but it is difficult to quantify how *good* a pass is in a game.

The final features of the BODGE model do not directly involve collecting data for each player. Instead, we need to find a way to track aspects of the game that involve location on the court at an arbitrary point in time. In particular, we track:

1. Which player has the ball at each point of time.
2. An identification label assigned to each player to tell us which region of the court a player is in at each point of time.

4.3 The BODGE Simulation

The BODGE simulation models a game through a series of continuous Markov chains. The simulation begins by having ten players on the court, five on each team. To simulate a jump ball, there is a $\frac{1}{2}$ probability that each team will have the first possession. After this has been decided the continuous Markov chain begins. The team that starts with the ball will pass the ball to a player to begin the possession by one of the features in our model. At each point in time, every player on the court has the option to change regions on the court. Additionally, the player who starts with the ball has three more options in the continuous Markov chain. They are

- the player with the ball shoots
- the player with the ball passes
- the player with the ball causes a turnover.

Each possession is bounded by a 24 second shot clock that has been implemented into our model. We add a condition such that the player with the ball will shoot with probability one if all 24 seconds expires and there has not been a shot or a turnover. After a shot or a turnover, the possession ends and the shot clock restarts. If the possession ends with a turnover, the other teams begins the next possession. If the shot was made, the other team starts with the ball. If the shot was missed, both teams have the opportunity to get the rebound. If the ball misses there is a set probability that the ball will fall in each region. Once the region has been determined, the simulation uses the rebounding feature to determine which player gets the rebound. This player's team will have the ball for the next possession. Furthermore, our simulation subdivides the game into four 12 minute (720) second periods, which represents a quarter of a basketball game. After the 48 minutes (2880 seconds) have passed, the game ends producing a final score. Note that if two teams play twice, it is highly improbable that the outcome of the game will be identical.

A useful feature of agent based models such as the BODGE simulation is that we can extract player statistics from these theoretical games. For each player, the BODGE simulation tells us an individual player's number of

- shot attempt sequences (whenever the ball is shot by a player including when the shot results with free throws, and any "and one" is still considered one shot attempt sequence)
- shot attempt sequences where points are scored
- turnovers
- points
- passes to a player that scores (which can be interpreted as assists).

Defense is implemented into the BODGE simulation. Recall that a feature of the BODGE model is the number of shots blocked or contested category. Whenever an offensive player shoots the ball, all defenders in the region independently have a given probability to affect the offensive player's field goal percentage. If one player contests the shot, the offensive player's field goal percentage is multiplied by 0.9; if two players contest a shot, the offensive player's field goal percentage is multiplied by 0.75; if three players contest a shot, the field goal percentage is multiplied by 0.5; and if four or five players contest a shot, the field goal percentage is multiplied by 0.25.

A useful application of the BODGE simulation is we can simulate games to predict the pairwise compassions between teams. Simply, just run a sufficiently large number of games between each pair of teams to find the probability that team i beats team j, which is denoted as p_{ij}. Given enough games played, this value of p_{ij} will converge. We now can use the Bradley-Terry model with each p_{ij} to obtain our player induced team rankings.

4.4 BODGE Convergence

To test the speed of the convergence of the BODGE simulation to a stationary ranking, we generated 30 teams of players synthetic players and simulated repeated round robin tournaments. An updated ranking is generated after each new round robin, and we calculate the Kendall-Tau distance from each ranking to the final ranking.

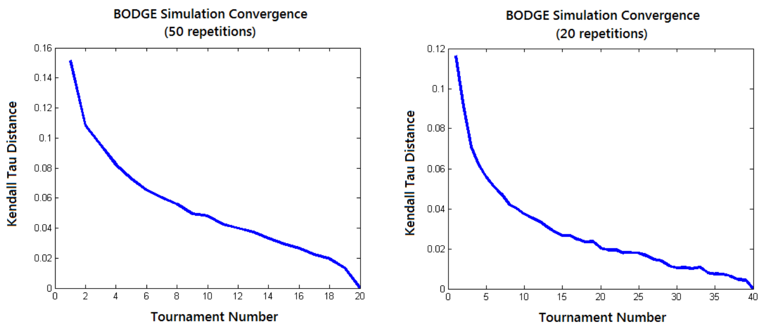

Figure 10: 50 repetitions of 20 Round Robin Tournaments

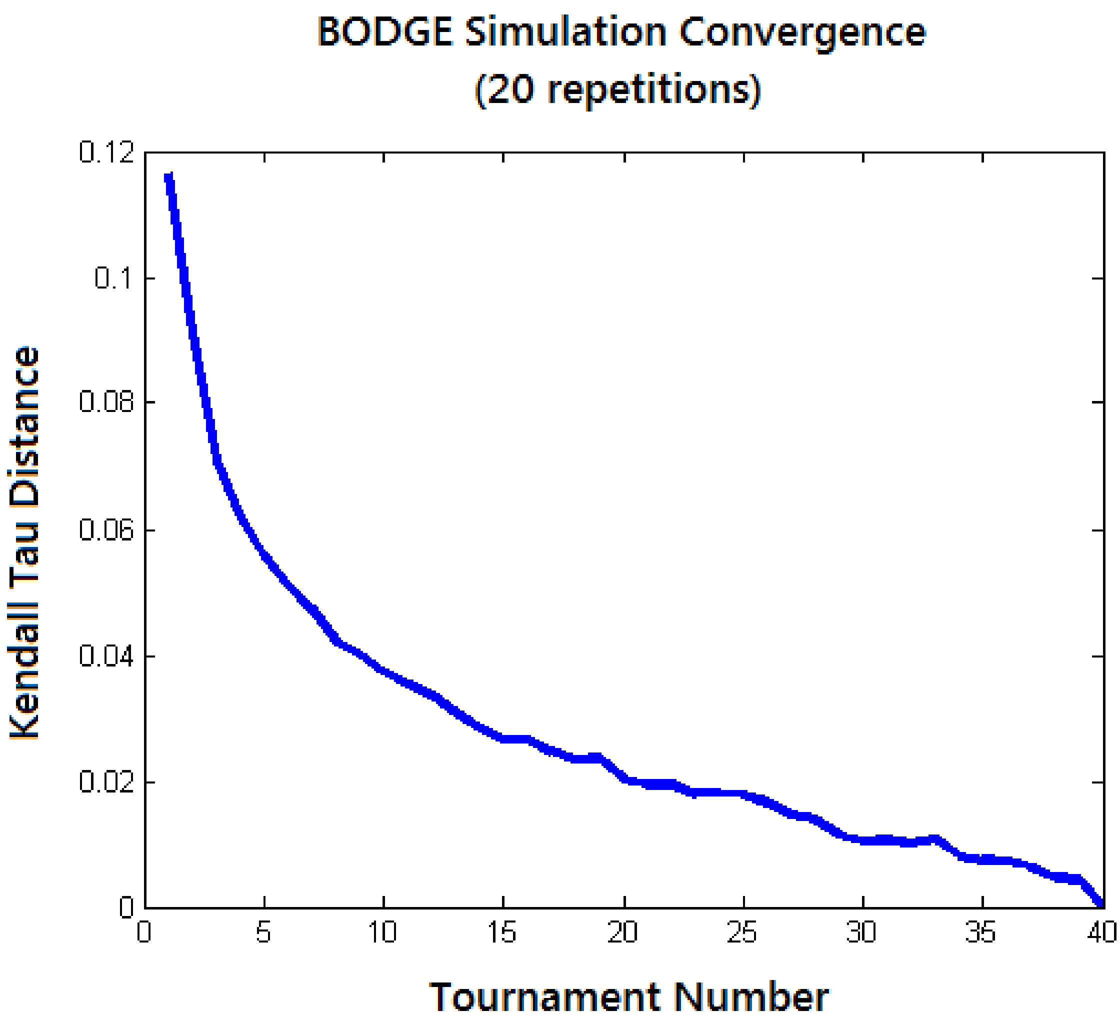

Figure 11: 20 repetitions of 40 Round Robin Tournaments

The results of the simulations show that it does not quite converge after 40 Round Robins. More simulations will be needed to determine exactly how long is typically needed.

4.5 Applications of the BODGE Model

One benefit of the BODGE is its ability to account for injuries. For example, during the 2012 playoffs, Derrick Rose suffered an ACL injury [7]. This injury resulted in the number 8 seed, the Philadelphia 76'ers, to beat the top seeded Chicago Bulls in the playoffs. Using the BODGE model, we could project the difference made by the loss of Rose. Furthermore, this application of the BODGE model could be a tool that is useful analysts in determining the outcome of series in the playoffs.

Another application of the BODGE for teams is the analysis of how beneficial a trade could potentially improve both teams' ability to win. It has been well known that Dwight Howard wants to leave the Orlando Magic as soon as possible, but the Magic are waiting for an acceptable offer[25]. Using the BODGE model, teams could see what trade combination would be most beneficial to their team. Once again, this is based off of a team's potential, not based off of team chemistry, so this may not always determine the best possible trade.

NBA analysts could use the BODGE model to explain why teams with a lot of raw talent may not be the best teams. Consider the 2012 New York Knicks and "Linsanity". Before Jeremy Lin became a starter, the Knicks had two NBA all stars on their team (Amare Stoudemire and Carmelo Anthony). When Jeremy Lin made his breakthrough, both Stoudemire and Anthony were injured [24]. Using the BODGE model, we can assess which Knicks team was the best team. Additionally, the BODGE model could explain why the Knicks are an average NBA team, even though they have significant talent on their team, as modeled by the motion of players and player tendencies. Furthermore, one could use the BODGE model to assess if the Knicks made the wrong decision by letting Jeremy Lin go as a free agent.

Up until this point, we have discussed the BODGE model for the NBA. Since BODGE models basketball, it can be used in NCAA or International Basketball Federation (FIBA) basketball, as well as high school and Amateur Athletic Union (AAU) basketball. The current version of the BODGE model simply models man-to-man defense, but could easily be extended to model zone defense. The changes on the defensive side would have each player be responsible for a zone as opposed to an individual man.

Finally, the USA could assemble the ideal Olympic team using the BODGE model. The greatest strength of the BODGE model is that it relies on player statistics and player's ability to play as a team. Since most other Olympic teams play a FIBA style of basketball, one could assemble the American team by using the BODGE model.

4.6 Future Improvements for the BODGE Model

A disadvantage of the BODGE model is that it is impossible to gather information for rookies prior to the season starting. One way to get around this problem is to take a sample of each statistic we need in the model for all players that made the transition from NCAA or FIBA to the NBA, and plot the scoring and other statistics in NCAA or FIBA play against its paralleled statistic in the NBA and perform a regression analysis to predict the expected statistic in NBA play. We would still have to predict the players potential motion, which could be determined by a coach or an analyst.

Another problem with the BODGE model is that it models a team's theoretical potential. Our model cannot explain intangible factors such as team chemistry. However, using player tendencies and teamwork based statistics, theoretical chemistry can be implied in the model. Similarly, a new coach may utilize players differently by implementing a new offense. If this occurs, our model would depict a theoretical team that does not exist. Our model, therefore, is ideal for coach-free or minimal coaching scenarios.

Another current problem with the BODGE model is that substitutions and foul trouble have not been implemented into the simulation. Moreover, it does not take into account for player stamina. There is room for improvement in the model to incorporate when and how often substitutions are made and modeling it with the continuous Markov chains. Also the question arises of which a bench player gets put into the game and how long a star player should stay out when he gets benched. These are decisions that should be made by the coach.

Lastly, the BODGE model attempts to break a continuous game into discrete events. By doing this, it is easier to analyze a continuous game. The problem is that we lose information about some details in the game. For example, it is impossible to account for a player setting a pick and how the defender will react to this pick. All the BODGE model will tell us is what the probability of the player switching regions is and what the probability that he will be open when he changes regions is. The BODGE model assumes that these probabilities can capture the likelihood that a pick will be set up, but it is not guaranteed. Also, the BODGE model does not account for set plays in close games; this is another category that would rely on a model for the coach.

4.7 Future Goals of the BODGE Model

The first goal of improvement on the BODGE model is to determine a way to model motion accurately. Currently, we have been collecting the data for motion based off of watching film footage and recording the location of players on the court and the time it takes a player to switch regions. We are hopeful that image processing applied to film footage, as described by [13] can accurately and precisely provide the statistics that are needed for the BODGE. With the use of image processing on film footage, it would become possible to subdivide the basketball court into smaller partitions or consider continuous player movement models.

The main weakness of the BODGE model is determining how to accurately model defense. Currently, the only defensive aspect of the BODGE model is how often a shot is contested. It seems possible that a help-side defense statistic could be developed, but we have not yet decided on how it should be derived. Using the techniques of image processing of film footage, we could measure the expected distance between a defender and the man he is guarding, and combining this with a defender's size and athleticism, we could furthermore improve a player's defense. It is difficult to measure how valuable of a defender a player is because there are few defensive statistics available and no team defensive statistics induced from player statistics.

5 A geometric interpretation of sampling tournaments

Consider a set of n teams, denoted $V = \{i\}_{i=1}^n$ and consider a game for which there are no ties. Let p_{ij} be the probability that team i beats team j, i.e.,

$$p_{ij} := \mathrm{pr}(i \text{ beats } j) \qquad i, j \in V \times V. \tag{3}$$

The statement that there is a winner at the end of each game can be written $p_{ij} + p_{ji} = 1$.

Define the set of distinct team pairs, $A = \{(i,j) \in V \times V \colon i > j\}$. Denote by $N := |A| = \binom{n}{2}$ the number of distinct team pairs. Given p_{ij}, we define a *tournament* on V to be a random variable $y \colon A \to \{0,1\}$ where $y_{ij} = 1$ with probability p_{ij}. If $y_{ij} = 1$, we say that team i beat team j in the tournament y and write $i \succ_y j$. A sample tournament y is *inconsistent* if there exists three teams i, j, and k such that $i \succ_y j$, $j \succ_y k$, and $k \succ_y i$. A team j is called *champion* if $j \succ_y i$ for all $i \in V \setminus \{i\}$.

A sample tournament on V can be represented by a directed graph where nodes represent teams and arc ij is present if $y_{ij} = 1$. The tournament is consistent if and only if its corresponding directed graph is acyclic. A team is champion if and only if its corresponding node contains no incoming arcs.

In this section, we develop a graphical representation for the set of all tournaments on V generated by p_{ij}. Our representation gives geometric interpretations for the probability for which a sample tournament will be inconsistent or yield a champion. We begin in §5.1 by considering a tournament among three teams and generalize the idea to n teams in §5.2.

5.1 A geometric interpretation of sampling 3-team tournaments

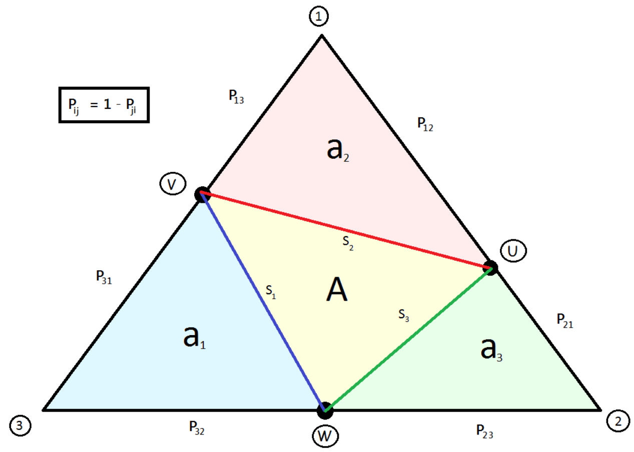

Figure 12: A geometric interpretation of 3-team tournaments. See §5.1

Consider the set of $n = 3$ teams, V, and let y be a tournament on V generated by p_{ij}. The probability that y is inconsistent is given by

$$\begin{aligned} \mathrm{pr}\{y \text{ inconsistent}\} &= \mathrm{pr}\{1 \succ 2 \text{ and } 2 \succ 3 \text{ and } 3 \succ 1\} + \mathrm{pr}\{2 \succ 1 \text{ and } 3 \succ 2 \text{ and } 1 \succ 3\} \\ &= p_{12}\, p_{23}\, p_{31} + (1 - p_{12})(1 - p_{23})(1 - p_{31}) \end{aligned} \tag{4}$$

As drawn in Fig. 12, consider an equilateral triangle, S, where each vertex represents one of the three teams. Let each edge of S has length 1, so that the area of S is given by $|S| = \frac{\sqrt{3}}{4}$. On each edge of the triangle, $\overline{ij}$, we consider the point which is a distance p_{ij} from vertex i and a distance p_{ji} from vertex j. In Fig. 12, these points are labeled U, V, and W. Let A be the triangle with vertices U, V, and W and denote the area of A by $|A|$. Furthermore, let B_i be the triangle closest to vertex i.

Proposition 5.1. *The probability that team i is the champion is given by $|B_i|/|S|$. The probability that y is inconsistent is given by $|A|/|S|$.*

Proof. Assume without loss of generality that team 1 is the champion of a tournament consisting of two other teams. Then the probability that team 1 is the champion is $p_{12}(p_{13})$, where $p_{12} = \mathrm{pr}(1 \succ 2)$ and $(p_{13}) = \mathrm{pr}\ (1 \succ 3)$, using elementary probability. The area of triangle a_2 can be calculated by using the formula $\frac{1}{2}p_{12}(p_{13}) \sin \frac{\pi}{3} = \frac{\sqrt{3}}{4}p_{12}(p_{13})$. Observe

$$\frac{|B_A|}{|S|} = \frac{\frac{\sqrt{3}}{4}p_{12}p_{13}}{\frac{\sqrt{3}}{4}} = p_{12}p_{13}.$$

Hence, the probability that team i is the champion of a tournament consisting of three teams can be given by the formula $|B_i|/|S|$.

$\square$

Proof. First we will show what is the probability that there will exist some inconsistency in a tournament of three teams. We will quickly define $p_1 = p_{31}, p_2 = p_{12}, p_3 = p_{23}$. Either $(1 \succ 2 \ \wedge \ 2 \succ 3 \ \wedge \ 3 \succ 1)$ or $(2 \succ \ \wedge 3 \succ 2 \ \wedge \ 1 \succ 3)$. Moreover,

$$\begin{aligned}
P(cycle) &= \prod_{i=1}^{3} p_i + \prod_{i=1}^{3}(1-p_i) \\
&= p_1p_2p_3 + (1-p_1)\cdot(1-p_2)\cdot(1-p_3) \\
&= \cancel{p_1p_2p_3} + (1 - p_1 - p_2 - p_3 + p_1p_2 + p_1p_3 + p_2p_3 - \cancel{p_1p_2p_3}) \\
&= 1 - \sum_{i=1}^{3} p_i + \sum_{i<j=1,1}^{3,3} p_ip_j
\end{aligned}$$

Now we will show that the probability that an inconsistency occurs is equal to $|A|/|S|$. Observe:

$$\begin{aligned}
\frac{|A|}{|S|} &= \frac{4 \cdot Area(A)}{\sqrt{3}} \\
&= \frac{4}{\sqrt{3}}(\frac{\sqrt{3}}{4} - a_1 - a_2 - a_3) \\
&= \frac{4}{\sqrt{3}}(\frac{\sqrt{3}}{4} - \frac{sin\frac{\pi}{3}\cdot p_1 \cdot (1-p_3)}{2} - \frac{\sin\frac{\pi}{3}\cdot p_2\cdot(1-p_1)}{2} - \frac{\sin\frac{\pi}{3}\cdot p_3\cdot(1-p_2)}{2}) \\
&= \frac{4}{\sqrt{3}}(\frac{\sqrt{3}}{4} - \frac{\sqrt{3}}{4}\cdot(p_1 - p_1p_3 + p_2 - p_1p_2 + p_3 - p_2p_3)) \\
&= 4(\frac{1}{4} - \frac{1}{4}\cdot(\sum_{i=1}^{3} p_i - \sum_{i<j=1,1}^{3,3} p_ip_j)) \\
&= 1 - \sum_{i=1}^{3} p_i + \sum_{i<j=1,1}^{3,3} p_ip_j
\end{aligned}$$

Therefore, the probability that y is inconsistent in a tournament consisting of three teams is precisely $|A|/|S|$.

□

5.2 A geometric interpretation of sampling n-team tournaments

Consider the set of n teams, V, and let y be a tournament on V generated by p_{ij}. The probability that i is the champion is given by

$$\mathrm{pr}\{i \text{ is the champion}\} = \prod_{j \neq i} p_{ij}. \tag{5}$$

The generalization of our representation to n-teams uses the notion of polytopes and simplexes. Recall that an *n-polytope* is an n-dimensional region enclosed by a finite set of half-spaces and that a *n-simplex* is an n polytope which is the convex hull of $n+1$ vertices. A *regular n-simplex* is an n-simplex with equidistant vertices. Let S_n denote the regular n-simplex where the distance between vertices is 1 denote the volume of S_n by $|S_n|$. Using determinants, it can be shown [3] that

$$|S_n| = \frac{1}{n!}\sqrt{\frac{n+1}{2^n}}. \tag{6}$$

Note that $|S_2| = \frac{\sqrt{3}}{4}$ as computed in §5.1.

Let S_{n-1} represent an n-team tournament so that each vertex represents a team. On each edge of S_{n-1}, $\overline{ij}$, we consider the point which is a distance p_{ij} from vertex i and a distance p_{ji} from vertex j. Let B_i be the $n-1$ simplex embedded inside S_{n-1} that contains vertex i.

Proposition 5.2. *The probability team i is the champion is given by $|B_i|/|S_n|$.*

Proof. From equation (6), we can use the methods presented [3] to find the volume $|B_i|$. We can observe that if Vertex i is fixed at the origin the sides extending from it are of length $p_{ij}, i \neq j$. We see through the determinant that

$$|B_i| = \frac{1}{n!}\sqrt{\frac{n+1}{2^n}} \prod_{j \neq i}^{n+1} p_{ij} \tag{7}$$

and furthermore

$$|B_i|/|S_n| = \prod_{j \neq i}^{n+1} p_{ij} \tag{8}$$

which is the probability that team i wins all its games and is the champion.

□

So given an acyclic tournament $\{a_j\}_{j=1}^{n+1}$, the $\prod_{j>n}^{n+1} p_{a_k a_j}$ is the probability of team k beating all teams ranked beneath it. Observe that the probability that a cycle does not occur is

$$\begin{aligned}
P(\text{no cycle}) &= \sum_{\substack{\{a_j\}:a_j \text{is a} \\ \text{permutation of}\{i\}_{i=1}^{n}}} \prod_{k=1}^{n} \prod_{j>k}^{n} p_{a_k a_j} & (9) \\
&= n! \times \prod_{k=1}^{n} \prod_{j>k}^{n} \frac{1}{2} & (10) \\
&= n! \times \left(\frac{1}{2}\right)^{\sum_{k=1}^{n} n-k} & (11) \\
&= \frac{n!}{2^{\frac{(n-1)^2-(n-1)}{2}}} = \frac{n!}{2^{\binom{n-1}{2}}} & (12)
\end{aligned}$$

As $n \to \infty$, $P(\text{no cycle}) \to 0$ by Stirling's approximation, and therefore the tournament will be inconsistent.

We can note that

$$P(\text{any acyclic tournament}) = \sum_{\text{all acyclic tournaments}} P(\text{a given acyclic tournament}) \quad (13)$$

and that for a given acyclic tournament $\{a_j\}_{j=1}^{n+1}$ to occur, team a_j must beat all teams a_k where $k > j$, which will occur with probability $p_{a_k a_j}$ for each pair a_j, a_k. We then find that

$$P(\text{a given cyclic tournament}\{a_j\}_{j=1}^{n+1} \text{ occurring}) = \prod_{k=1}^{n} \prod_{j>k}^{n+1} p_{a_k a_j} \quad (14)$$

. Therefore, the probability that no cycle will occur is given by the formula

$$P(\text{No Cycle Occurring}) = \sum_{\substack{\{a_j\}:a_j \text{is a} \\ \text{permutation of}\{i\}_{i=1}^{n+1}}} \prod_{k=1}^{n} \prod_{j>k}^{n+1} p_{a_k a_j} \quad (15)$$

We can note that each face of a n-simplex is an $(n-1)$-simplex and thus once a champion has been determined we can eliminate that one vertex and consider the $(n-1)$-simplex formed by the n remaining teams. We hope to be able to extend this concept to a geometric interpretation to the simplex representing n teams in a tournament.

6 A Measure of Offensive Value in Baseball

Each year, the Baseball Writers Association of America selects one player from the American League and one player from the National League to receive the award of Most Valuable

Player [20]. Currently these writers use a variety of statistics including Batting Average and RBIs for batters and wins and ERA for pitchers in an attempt to measure the contribution a player has to his team's success. We have devised a new aggregated statistic to help identify offensive productivity based upon the ability of a batter to move runners along the bases into scoring position and across the plate. We further incorporate the leverage index [26] to identify players that are able to move runners in high-pressure situations.

6.1 Methodology

In order to calculate the new statistic, which we term **batting efficiency**, we require the number of outs before and after the at-bat, the status of the bases before and after the at-bat, and the number of runs that scored during the at-bat. In order to compute the **leverage adjusted batting efficiency** we also collect the leverage index. All data was collected for these calculations using a php script to scrape http://www.baseball-reference.com for the 2011 season. In order to compute the batting efficiency for a single at-bat we use Table 13.

Base Status	- - -	1 - -	- 2 -	- - 3	1 2 -	1 - 3	- 2 3	1 2 3
Score	5	9	8	7	12	11	10	14

Figure 13: The various potential value for the base situations both before and after a player's at-bat

In order to determine the score of the at-bat we use the following formula:

$$\textbf{AB Score} = \text{Score}_{final} - \text{Score}_{initial} + 5 \times \text{RBIs}$$

To determine a player's season batting efficiency we sum the batter's score and divide it by the batter's possible score, i.e.:

$$\textbf{Batting Efficiency} = \frac{\displaystyle\sum_{\text{all at-bats}} \text{score achieved}}{\displaystyle\sum_{\text{all at-bats}} \text{score possible}}$$

When we consider the leverage adjusted batting efficiency, we insert the leverage index and compute as follows:

$$\textbf{L.A.B.E.} = \frac{\displaystyle\sum_{\text{all at-bats}} \text{score achieved} \times \text{leverage index}}{\displaystyle\sum_{\text{all at-bats}} \text{score possible} \times \text{leverage index}}$$

6.2 Results

We computed both the Batting Efficiency (figure 14) and Leverage Adjusted Batting Efficiency (figure 15) on all at-bats from the 2011 season and removed all players who had fewer than 200 plate appearances. We note that Ryan Braun, the winner of the National League MVP award, ranks first in both B.E. and L.A.B.E., while other players such as Mike Napoli perform significantly better in low leverage situations, as he is ranked second in Batting Efficiency while he is ranked 22^{nd} in Leverage Adjusted Batting Efficiency.

Batting Efficiency			
Player	Bases Achieved	Bases Possible	Efficiency
Ryan Braun	782	4430	0.1765
Mike Napoli	540	3065	0.1762
Jose Bautista	797	4606	0.1730
Matt Kemp	865	5055	0.1711
Miguel Cabrera	848	4981	0.1702
Prince Fielder	853	5030	0.1696
Matt Downs	279	1653	0.1688
Adrian Beltre	636	3835	0.1658
Carlos Gonzales	623	3759	0.1657
Joey Votto	826	4985	0.1657

Figure 14: Top 10 in Batting Efficiency of players with at least 200 plate appearances for the 2011 season

7 Future Directions

We would still like to investigate a way to combine rankings, such as performing rank aggregation experiments. Rank aggregation allows us for novel ranking methods through the combination of other existing ranking methods. We programmed several rank aggregation methods ranging from Borda Count[19] which calculates the sum of rankings to an optimized rank aggregation method which solves a linear programming problem that minimizes the Kemeny difference of an aggregated rank to each of its composite ranks [15]. Time restrictions inhibited the analysis of these methods, but in the future we would like to study the properties of rank aggregation. In particular we would like to see how well a rank aggregation performs in cross validation compared to how well its composite ranks perform in cross validation. Additionally we would like to perform more experiments to test the accuracy of the BODGE model. We hope to find connections between the geometric interpretation of sampling tournaments and Hodge decomposition.

Batting Efficiency			
Player	Adj. Bases Achieved	Adj. Bases Possible	Leverage Adjusted Efficiency
Ryan Braun	961.22	4723.49	0.2035
Travis Hafner	648.19	3188.82	0.2033
Josh Hamilton	706.66	3493.7	0.2023
Prince Fielder	1092.58	5591.93	0.1954
Lance Berkman	955.33	4976.15	0.1920
Robinson Cano	1066.95	5564.03	0.1918
Matt Kemp	1112.83	5811.14	0.1915
Jose Bautista	1049.31	5482.25	0.1914
Miguel Cabrera	1074.06	5639.29	0.1905
Joey Votto	1116.68	5943.94	0.1879

Figure 15: Top 10 in Leverage Adjusted Batting Efficiency of players with at least 200 plate appearances for the 2011 season

References

[1] Avrim Blum, Adam Kalai, and John Langford. Beating the hold-out: bounds for k-fold and progressive cross-validation. In *Proceedings of the twelfth annual conference on Computational learning theory*, COLT '99, pages 203–208, New York, NY, USA, 1999. ACM.

[2] Ralph Allen Bradley and Milton E. Terry. Rank analysis of incomplete block designs: I. the method of paired comparisons. *Biometrika*, 39:324–345, 1952.

[3] Ralph H. Buchholz and Michael J. Smith. Volume of an n dimensional simplex. 1996.

[4] Mucha Peter J. Callaghan, Thomas and Mason A. Porter. Random walker ranking for ncaa division i-a football. 114:761–777, 2007.

[5] W. N. Colley. Colley's bias-free college football ranking mehtod: The colley matrix explained, 2002.

[6] Janez Demšar. Statistical comparisons of classifiers over multiple data sets. *J. Mach. Learn. Res.*, 7:1–30, December 2006.

[7] Mike Dodd. Rose tears acl to spoil bulls' game 1 victory, 2012.

[8] A.E. Elo. The new u.s.c.f. rating system. *Chess Life*, 16:160–161, 1961.

[9] Petar Vracar Erik Strumbelj. Simulating a basketball match with a homogeneous markov model and forecasting the outcome. *International Journal of Forecasting*, 28:532–542, 2011.

[10] Mark E. Glickman. A comprehensive guide to chess ratings. *American Chess Journal*, 3:59–102, 1995.

[11] Minka T. Graepel, T. and R. Herbrich. Trueskill(tm): A bayesian skill rating system. In Platt J. Scholkopf, B. and T. Hoffman, editors, *Advances in Neural Information Processing Systems 19*, pages 569–576. MIT Press, 2007.

[12] Kalyanaraman Kaushik Hirani, Anil N. and Seth Watts. Least squares ranking on graphs. *arXiv*, 1011.1716:1–43, 2010.

[13] Tim McGarry Jarome Bourbousson, Carole Seve. Space-time coordination dynamics in basketball. *Journal of Sports Sciences*, 28, 2010.

[14] Lim Lek-Heng Yao Yuan Jiang, Xiaoye and Yinyu Ye. Statistical ranking and combinatorial hodge theory. *Mathematical Programming*, 127:203–244, 2011.

[15] Yoshitsugu Yamamoto Kathryn E. Pedings, Amy N. Langville. A minimum violations ranking method.

[16] Paul Kvam and Joel S. Sokol. A logistic regression/markov chain model for ncaa basketball. *Wiley InterScience*, 2006.

[17] Michael Lewis. *MoneyBall*. W. W. Norton & Company, 2003.

[18] Kenneth Massey. Statistical models applied to the rating of sports teams. Master's thesis, Bluefield College, 1997.

[19] University of Alabama. The borda count method.

[20] Baseball Writers' Association of America. Voting faq, August 2012.

[21] Lawrence Page, Sergey Brin, Rajeev Motwani, and Terry Winograd. The pagerank citation ranking: Bringing order to the web. Technical Report 1999-66, Stanford InfoLab, November 1999. Previous number = SIDL-WP-1999-0120.

[22] Joel S. Sokol Paul Kvam. A logistic regression/markov chain model for ncaa basketball. *Naval Research Logistics*, 53:788–803, 2006.

[23] David Pickle and Bruce Howard. Computer to aid in basketball championship selection. *NCAA News*, page 4, February 1981.

[24] Ben Reiter. Two weeks of jeremy lin's rise a period that we'll never forget, 2012.

[25] Brad Ross. Nba trade rumors: The biggest conflicting reports surrounding dwight howard, 2012.

[26] Lichtman Mitchel Tango, Tom and Andrew Dolphin. *The Book: Playing the Percentages in Baseball*. Potomac Books, Inc., 2007.