Ajay Andrew Gupta*

A new approach to bracket prediction in the NCAA Men's Basketball Tournament based on a dual-proportion likelihood

Abstract: The widespread proliferation of and interest in bracket pools that accompany the National Collegiate Athletic Association Division I Men's Basketball Tournament have created a need to produce a set of predicted winners for each tournament game by people without expert knowledge of college basketball. Previous research has addressed bracket prediction to some degree, but not nearly on the level of the popular interest in the topic. This paper reviews relevant previous research, and then introduces a rating system for teams using game data from that season prior to the tournament. The ratings from this system are used within a novel, four-predictor probability model to produce sets of bracket predictions for each tournament from 2009 to 2014. This dual-proportion probability model is built around the constraint of two teams with a combined 100% probability of winning a given game. This paper also performs Monte Carlo simulation to investigate whether modifications are necessary from an expected value-based prediction system such as the one introduced in the paper, in order to have the maximum bracket score within a defined group. The findings are that selecting one high-probability "upset" team for one to three late rounds games is likely to outperform other strategies, including one with no modifications to the expected value, as long as the upset choice overlaps a large minority of competing brackets while leaving the bracket some distinguishing characteristics in late rounds.

Keywords: brackets; college basketball; maximum likelihood; statistics.

DOI 10.1515/jqas-2014-0047

1 Introduction

For 68 National Collegiate Athletic Association (NCAA) Division I teams, the men's college basketball season culminates in a 68-team tournament known often simply as the NCAA Tournament. The NCAA Tournament has had at least 64 teams since the 1985 tournament, and currently has 68. It is single-elimination, so each team plays based on a prescribed set of possible opponents, with the winner of each game advancing to the next round of the tournament and the loser exiting the entire tournament. The rounds, in order, are the Round of 68, Round of 64, Round of 32, Sweet 16, Elite 8, Final 4, and the national championship game.

The NCAA has a selection committee that decides which teams to include, other than those who qualify by winning their conference's tournament or the Ivy League's regular season championship. The committee also groups teams into four "regions," and assigns "seeds" within each region. The seeds are integers from 1 to 16, with lower seed numbers indicating a higher-quality team. The NCAA selection committee has various criteria for deciding on the inclusion and seeding of teams. Arguably the most heavily applied is the ratings percentage index (RPI), a weighted combination of win-loss proportions involving a team and other connected teams. Another measure used is a set of team ratings produced by sports statistician Jeff Sagarin, which are his estimated mean points per game for the team (Sagarin 2014). Sagarin also includes a strength-of-schedule metric, based on the ratings of opponents played.

Beginning in 1977, fans have competed in "bracket pools," contests to predict the NCAA Tournament correctly (Geiling 2014). Participants generally pay an entry fee and submit a set of predictions for each game's winners known as a "bracket." The entry fees form a prize fund for the player with the highest points from correct predictions. Second- or third-place finishers may share part of the fund. In bracket predictions, players earn full points for a given game by predicting the winner, even if the predicted loser was eliminated in a previous round. These pools became more popular when the NCAA Tournament included 64 teams, especially among co-workers in an office. Such office pools often carried pool-specific rules such as an "upset bonus" which awarded more points for correct picks when the selected team had a worse seed. In recent years, the system has changed to large Internet-based administrators which handle many different groups with standardized sets of rules. Now, the vast majority of

***Corresponding author: Ajay Andrew Gupta,** Statistics, The Florida State University, 117 N. Woodward Ave. P.O. Box 3064330, Tallahassee, FL 32306, USA, e-mail: ajgupta@stat.fsu.edu

pools have potential points per game that double with each round, and exclude extra features such as upset bonuses. With the changes, bracket pools have grown. They now have an estimated $1 billion in off-book wagers and 60 million Americans participating (Geiling 2014).

But how does one compete shrewdly in such a pool? Many participants have seen few (if any) of the tournament team's regular-season games that year, so useful methods generally should not involve extensive knowledge of college basketball. Because of this, the "Pick the Seeds" strategy is very prevalent. In this, a player chooses the better-seeded team to win each game within the region. For the Final 4 and championship game, the winners are selected by judgment.

Modeling of the tournament in full is difficult. The single-elimination structure makes outcomes highly variable. There is also a sample size problem. For a given year's tournament, many teams could realistically play each other. Factors such as one team's uncharacteristically good or bad shooting impact which potential games end up in historical data. We only end up with data from one game for each possible pairing, though, or zero if either team gets eliminated before they could play. The next year, the structure remains, but the teams included are different. Even returning players have different players and opponents.

This paper does not attempt to discover the unknown tournament. It has a smaller focus, creating an optimal bracket for the NCAA Tournament. My intent was to introduce a methodology for bracket prediction that is transparent and effective against common competitors, such as Pick the Seeds. This implies restrictions on predictor variables, though. If this methodology is to outpredict Pick the Seeds, it cannot reproduce Pick the Seeds. I want to avoid using the seeds as predictors or using variables such as RPI which are used by the selection committee and have very high correlation with the seeds. Also, to have transparent methodology, I cannot use predictors that are not transparent themselves. Therefore, I also avoid opinion-based measures such as poll rankings, expert opinions, and betting information, because one can know the results for the predictors but not the full rationale. I also avoid predictors that are statistical measures with proprietary methodologies, such as the Sagarin ratings, because this essentially delegates the modeling to a second type of model with unknown details.

Some researchers have intended to make statistically simple methods that could be used by introductory statistics students (Carlin 1994; Schwertman, Schenk and Holbrook 1996). In this method, however, an average bracket-maker would not produce the predicted bracket himself of herself. An authority in statistics would implement the methodology each year, advising many people like Jeff Sagarin's ratings do. Unlike predictions made by basketball experts, these would be algorithmic, not subjective. Unlike the Sagarin ratings, this method would produce predictions for bracket winners and probabilities of winning. This methodology would be available, too, even if many would not have the background knowledge to implement it. It is debatable whether many people in unrelated fields would want to implement a statistical method for their brackets, anyway. A simple enough method may also not be adequately accurate considering the modeling challenges of the tournament.

2 Existing literature

2.1 Seed distributions

Quantitative work on the NCAA Tournament has covered areas with varied degrees of utility for bracket prediction. Some work has focused on the probability distributions of the seed numbers of teams winning at certain stages of the tournament. Schwertman, McCready and Howard (1991) and Schwertman et al. (1996) proposed regression-based models to predict the probability of a better-seeded team defeating a worse-seeded team as a function of the teams' seeds. More recently, Jacobson et al. (2011) estimated the seed distribution with a truncated geometric distribution.

2.2 Normal approximations

Carlin (1994) proposed a method to predict the 1994 NCAA Tournament using gambling point spreads for the Round of 64, and 1.165 times teams' difference in Sagarin ratings for later rounds. Point spreads are casinos' estimates of one team's score minus another. His estimated probabilities were cumulative distribution functions (CDFs) for a Gaussian with a mean of 0 and standard deviation of 10. Breiter and Carlin (1996) used the same general model, and turned these estimated probabilities into predicted brackets ending with the Elite 8.

Kaplan and Garstka (2001) also examined bracket prediction, using the 1998 and 1999 NCAA Tournament and National Invitational Tournament (NIT). Two of their methods modeled the probability that team *i* defeats team *j*, two of which used Gaussian CDFs, with a mean of 0 and standard deviation of $\sqrt{\lambda_i+\lambda_j}$, evaluated at $\lambda_i-\lambda_j$. One method set the λ_i values to be teams' Sagarin ratings. The

other method deduced λ_i from gambling point spreads and point totals. Point totals are like point spreads, but for the sum of teams' scores instead of the difference.

2.3 Regression methods

Kaplan and Garstka (2001) had a third method which used regular-season wins and losses. This assumed teams followed a Bradley-Terry model. Koenker and Basset (2010) modified a paired comparison model which was also based on the regular season. The paired comparison model calculated a team's expected score using an offensive strength parameter and defensive strength parameter for each team, as well as one home-court adjustment for all teams. They modified the paired comparison model to handle dependence by copula methods. They also replaced each strength parameter and the home-court adjustment with a 199-quantile function. This model can make bracket predictions for some games, but suffers from extreme overfitting. For an average team, they attempted to fit 199 quantiles from 12.67 data points. It is not included in the comparison models for Section 7 because the overfitting prevents it from predicting several games.

West (2006) used ordinal logistic regression to model the number of NCAA Tournament games a team would win. His predictors were the team's proportion of games won, cumulative points scored minus points allowed, Sagarin strength-of-schedule metric, and number of wins against teams with top-30 Sagarin ratings.

2.4 Competitive strategy

Metrick (1996) examined adjustments one should make for competitors' brackets. He created four logistic regression models for the probability of a bracket choosing a #1 seed as its tournament champion. These showed evidence of participants choosing #1 seeds less often as pool size decreases and choosing teams more often if particpants lived in the school's metropolitan area. He also calculated a mixed-strategy Nash equilibrium on an approximated version of bracket scoring. The assumed probabilities were translated gambling odds of winning the tournament. From this, Metrick concluded that players have a consistent tendency to select #1 seeds as their champions too often. He suggested #2 and #3 seeds instead.

This topic is valuable, but Metrick's assumed probabilities fit poorly. The three teams he considered most overpicked all made the Final 4. None of the three teams he considered most underpicked made the Final 4. More importantly, his approximated scoring system was oversimplified. Metrick looked only at the selected champion. The approximation picked a winner randomly among those who correctly predicted the champion, or among the entire pool if no one predicted the champion. Metrick largely ignored the case in which no one correctly predicts the tournament champion. He only mentioned that someone had correctly picked the tournament winner in all 24 pools he studied. His median pool size was 54, though. The case of an unpredicted champion is more important with today's smaller pools. Also, in the tournament he studied, the most heavily selected team to win the championship actually won. This scenario is very different from 2014, in which #7 seed Connecticut defeated the #8 seed Kentucky in the championship game.

2.5 Contribution of this work

This work uses much more data in training and validating the model than typical in past work. Seed distribution work has tended to use many seasons because its only required data are seed numbers and NCAA Tournament wins. Among previous papers that used more than seed information, the best data usage was two seasons in Kaplan and Garstka (2001) and four seasons in West (2006). Other work has generally included only one season (Schwertman et al. 1991, 1996; Carlin 1994; Breiter and Carlin 1996; Metrick 1996; Koenker and Basset 2010). This produces a very small sample considering that tournament outcomes are highly variable. "Upsets," in which a game defies expectations (usually seed-based), occur with high frequency. I use six seasons, in hopes of avoiding a fit that is specific to one tournament. I also use a six-fold cross-validation procedure, which uses the data more efficiently than West's independent validation set design. It produces six different years of validation sets instead of West's one.

This paper examines the task of bracket prediction directly. By contrast, West intended his work to produce ratings used by the tournament's selection committee. The seed distributions predict probabilities of seed-based upsets, but all of these probabilities are <50%, making the predicted bracket identical to Pick the Seeds. Breiter and Carlin (1996), Metrick (1996) and Kaplan and Garstka (2001) all address pools. This work, unlike those three, uses the now-common bracket scoring system. The others all increase points more slowly by round, and some include an upset bonus.

I will also revisit the topic of adjustments for human competitors. Unlike Metrick (1996), this will use simulations from various pool sizes similar to those common

today. It will also use the full bracket scoring system common today, instead of an approximation only using the selected champion.

3 Data

This work uses data from free online resources for the 2008–2009 through 2013–2014 seasons (Sports-Reference 2014; ESPNCB 2014). I incorporate games from teams' regular seasons, conference tournaments, and NCAA Tournament games before the Round of 64 in my team ratings. Bracket pools begin with the Round of 64, treating the previous winners as given, so these data would be available. I treat regular-season and conference tournament data differently, and include NCAA Tournament games before the Round of 64 as conference tournament games. I included every game in which at least one team was in Division I, and treated any non-Division I teams in these data as one combined team. Including the combined non-Division-I team, the numbers of teams by season were 349, 348, 347, 345, 348, and 352, in chronological order.

For Section 8, I created a distribution of competitors' bracket selections using people's 2013 and 2014 tournament bracket selections from ESPN's Tournament Challenge, because it was the largest source, and had available data (ESPNWPW 2014; Quintong 2014). It included a distribution of proportions picking teams at a certain round, not a joint distribution including entire brackets. Consequently, I needed to assume that each game was picked independently from the overall distribution, conditional on the teams playing the game. The assumption is not correct, because the same beliefs and strategies can impact multiple selections in one's bracket. However, I did not have appropriate data to model the dependence more accurately.

4 Probability model

For NCAA basketball teams with numbers t, define strength parameters $s(t)$, and a generic home-court strength parameter s_{HC} which is added if a team plays at home. Thus, for each game g, there is an effective strength $s_{adj}(t, g)$ equal to $s(t)+s_{HC}$ if game g is on team t's home court, and $s(t)$ otherwise. The work that follows will also use functions $w(g)$, $h(g)$, and $l(g)$, which yield the team numbers of game g's winner, higher-effective-strength team, and lower-effective-strength team, respectively. For tied $s_{adj}(t, g)$, it does not matter which team is assigned in each role, because both probabilities are 0.5.

The probability model in Eqn. 1 will be referred to as the dual-proportion model. This describes the probability that team t wins game g. In this,

$$d(g)=s_{adj}(h(g),g)-s_{adj}(l(g),g).$$

$$\mathrm{P}(w(g)=t)=\begin{cases}1-0.5e^{-d(g)}, & h(g)=t\\ 0.5e^{-d(g)}, & l(g)=t\end{cases} \tag{1}$$

Eqns. 2 and 3 show the likelihood $\mathcal{L}$ and log-likelihood ℓ for games numbered 1 to G. In these, $\boldsymbol{D}$ is observed data about the games. $\boldsymbol{\Omega}$ is the collection of all the model's parameters, which includes two types. Each team t has a strength parameter $s(t)$. Higher values indicate that team t is more likely to win a game. There is also a home-court strength s_{HC}. A positive value indicates that a team is more likely to win if it plays at home. The effect of the game location increases as s_{HC} gets further from zero. The home-court adjustment is on the same scale as the team strengths $s(t)$. In Eqns. 4–7, $\boldsymbol{\Omega}$ also includes a scaling parameter β that applies to big wins, meaning at least 10 points. A positive value implies that a team more likely to win overall (than an average team) is also more likely to win big. A magnitude of β greater than one implies that distinctions in team quality are larger with regard to winning big than with regard to winning at all. A magnitude less than one implies that distinctions in team quality become smaller when predicting more likely teams to win big.

$$\mathcal{L}(\boldsymbol{\Omega}|\boldsymbol{D})=\prod_{g=1}^{G}[0.5e^{-d(g)}]^{\mathbf{1}_{(w(g)=l(g))}}[1-0.5e^{-d(g)}]^{\mathbf{1}_{(w(g)=h(g))}} \tag{2}$$

$$\ell(\boldsymbol{\Omega}|\boldsymbol{D})=\sum_{g=1}^{G}([\ln 0.5-d(g)]\mathbf{1}_{(w(g)=l(g))}+\ln(1-e^{-d(g)})\mathbf{1}_{(w(g)=h(g))}) \tag{3}$$

Unfortunately, the log-likelihood lacks a continuous derivative like that of the Bradley-Terry model. However, its shape is better for bracket prediction. The Bradley-Terry model does not have a linear shape, but still does not tend to fit as steep a curve as the dual-proportion model does. The dual-proportion model better reflects how some NCAA Division I basketball teams are of very comparable strength, but some have large differences. Some schools recruit much better athletes because of scholarships, reputations, and larger student populations. In bracket-making, identifying games that can be treated as near-certainties lets one see which teams are more likely to be the winners of later rounds because of an easy early-round schedule.

In work with one proportion, one typically uses a logistic model. I wanted to reflect that each basketball game has two proportions involved. Each of the two teams in a game has a probability of winning the game, and these must add up to 100%. An optimization can add this as a separate constraint. However, I prefer to incorporate the existence of two dependent proportions into the model because it is a fundamental element of the basketball being modeled. One can still add a logistic structure for a portion of the model, but the advantages are lost. The model already constrains the proportion to the 0–1 range, and would not have a continuous derivative either way. This version of the dual-proportion model is less complicated, in mathematical form and by avoiding interpretation of the log-odds ratio.

5 Team ratings

I modify the dual-proportion model to produce regular-season ratings, which are the most important predictor variable in my method for bracket predictions. Maximizing Eqn. 3 on the regular season will not produce strengths that maximize Eqn. 3 for the NCAA Tournament, because of differences between the population of regular season games and that of NCAA Tournament games. Examples include the NCAA Tournament's pre-game festivities, larger audiences, and use of neutral courts only. Methods using unadjusted regular-season likelihood of winning tend to pick relatively minor-conference ("mid-major") teams with very strong regular-season records to perform better than their popular expectations. For example, Kaplan and Garstka's Bradley-Terry model picked Gonzaga of the West Coast Conference to win the 2013 NCAA Tournament. They had a 31-2 record entering the tournament, but lost in the Round of 32.

An important modeling question is whether to use a binary win-loss response or to use a point difference response and translate it to the binary variable. Some wins are more indicative of future wins than others are. A 1-point win may imply two teams of similar qualities in which one happened to prevail, whereas a 25-point win would imply two teams of completely different calibers. Then again, modeling point differences introduces a loss function other than that for the intended use and stops someone from using patterns that exist only on a binary level. Point differences can also be deceptive, because the goal for the basketball players and coaches is to win the game, not to maximize the score margin. The point difference gets impacted by late-game decisions in which a leading team slows its offense or in which a trailing team commits personal fouls to stop the game clock.

My rating system addresses the strength-of-schedule issue and the binary and point difference responses' weaknesses with the same adjustment. I include two binary variables in the regular-season likelihood function being optimized. One, which is part of likelihood for all games, is about whether a team won or lost. The other is also about winning or losing, but only applies in games considered a big win by one team. I defined a big win as being a win of 10 or more points, with no overtimes. I intended 10 points to be a threshold at which it would be implausible for a game to become a big win when within one shot of a tie in the last minute.

The strength-of-schedule adjustment works by reducing the strengths of the non-tournament teams in a weak conference. In a strictly win-loss likelihood system, one or two losses by a team can be overwhelmed by many wins in a relatively weak conference, producing less decrease in strength than several losses in a very strong conference. The weaker teams in a weak conference are likely to have big out-of-conference losses, which get amplified by the big-win term. This lets the best teams in a weak conference have an average strength parameter and still have large probabilities of beating their conference teams. The NCAA Tournament teams in the strongest conferences are likely to have small in-conference losses, which get dampened by their exclusion from the big-win term. Squaring the likelihood term for the big wins creates an adequate balance between the big and small wins.

The likelihood function for ratings also amplifies the effect of conference tournament games. The rationale is that the population of conference tournament games is more similar to that of the NCAA Tournament games than the population of regular-season games is. Reasons include the single-elimination structure and heavier television coverage. The likelihood function is shown below.

$$\mathcal{L}_R(\boldsymbol{\Omega}|\boldsymbol{D}) = \prod_{g=1}^{G} p_{\mathrm{BW}}(g|\boldsymbol{\Omega})^{2c(g)}\, p_W(g|\boldsymbol{\Omega})^{c(g)} \tag{4}$$

Above, the function $c(g)$ is 1.8 for a conference tournament game and 1 for a regular-season game. The 1.8 here and the power of 2 for the big-win term in the likelihood function were tuning parameters chosen by trial-and-error testing of several integer multiples of 0.2. For a potential $c(g)$ and big-win exponent, I trained a model on a year's regular-season data and tested the model on the same year's tournament data. The values chosen appeared to have reasonable results.

The term $p_W(g|\mathbf{\Omega})$ in Eqn. 4 is the dual-proportion model's probability that team $w(g)$ wins game g, where team $w(g)$ is the actual winner of game g. The term $p_{BW}(g|\mathbf{\Omega})$ is the probability that team $w(g)$ has a big win in game g, given that game g is a big win for one team. Both of these are detailed below.

$$p_{BW}(g|\mathbf{\Omega})=\begin{cases}[0.5e^{-\beta d(g)}]^{\mathbf{1}_{(w(g)=l(g))}}[1-0.5e^{-\beta d(g)}]^{\mathbf{1}_{(w(g)=h(g))}}, & b(g)=1\\ 1, & b(g)=0\end{cases} \tag{5}$$

$$p_W(g|\mathbf{\Omega})=[0.5e^{-d(g)}]^{\mathbf{1}_{(w(g)=l(g))}}[1-0.5e^{-d(g)}]^{\mathbf{1}_{(w(g)=h(g))}} \tag{6}$$

Eqn. 5 uses $b(g)$, which is an indicator function for game g being a big win by either team, and β, a scaling parameter for the ratings when used in the big-win term. Combining Eqns. 4–6, one gets the log-likelihood function below.

$$\begin{aligned}\ell_R(\mathbf{\Omega}|\boldsymbol{D})=&\sum_{g=1}^{G}c(g)([\ln 0.5-d(g)]\mathbf{1}_{(w(g)=l(g))}+\ln(1-0.5e^{-d(g)})\mathbf{1}_{(w(g)=h(g))})\\&+2\sum_{g=1}^{G}b(g)c(g)([\ln 0.5-\beta d(g)]\mathbf{1}_{(w(g)=l(g))}+\ln(1-0.5e^{-\beta d(g)})\mathbf{1}_{(w(g)=h(g))})\end{aligned} \tag{7}$$

I also constrained the ratings $s(t)$ such that $-6\le s(t)\le 6$ for all t, so even the best team would have a 0.25% probability of losing to an average team and the worst team would have a 0.25% probability of defeating an average team. The strongest example of where this applies is Wichita State in 2013–2014. They entered the NCAA Tournament with a 34-0 record, so maximizing likelihood would give them a strength of ∞. There is a difference, however, between a team having only wins in the data and a team being unbeatable, so I set a threshold to ensure a reasonable minimum probability. I used separate ratings for each team in Division I, plus another for the combined non-Division I teams they played. Additional details about optimizing the log-likelihood in Eqn. 7 can be found in Appendix A.

6 Tournament prediction

For the NCAA Tournament, I maximize the dual-proportion model likelihood from Eqn. 2, using tournament strengths $s_T(t)$ from Eqn. 8. Even though the regular-season ratings $s_{RS}(t)$ were designed to generalize well to the NCAA Tournament, there is still valuable information not included in regular-season wins and losses. In Eqn. 8, x_2, ..., x_N are the other predictors and $\{\beta_n\}$ are coefficients found in regression. No intercept is needed because I only use differences of s_T.

$$s_T(t)=\beta_1 s_{RS}(t)+\sum_{n=2}^{N}\beta_n x_n(t) \tag{8}$$

The details of regression for the coefficients $\{\beta_n\}$ are in Appendix B. I maximized likelihood to find the coefficients, but did not use likelihood to evaluate the model. In a very random application such as the NCAA Tournament, having a high likelihood is achieved largely by a method being indecisive. If a method identifies which games are too close to call reliably and assigns each team near-50% probabilities of winning, that method can perform well relative to more decisive methods. If a method assigns 51% probability to one team and 49% to another in games such as those between #2 seeds and #3 seeds, then it will get relatively little credit for picking correctly but will suffer little damage for picking incorrectly. Considering the frequency of NCAA Tournament upsets, minimizing the downside is particularly important if the objective is high likelihood.

In bracket prediction, the games that are too close to call reliably must still be called, and one gets no credit for selecting the wrong team but assigning high uncertainty to the prediction. The appropriate score function, therefore, is the number of bracket points. I calculate this by what has become the standardized point system: 1, 2, 4, 8, 16, and 32 points per correct prediction in the Round of 64, Round of 32, Sweet 16, Elite 8, Final 4, and championship game, respectively. In practice, these may be multiplied by a constant, but it is the same point system.

Calculating the points is straightforward, but selecting teams is not a simple matter of choosing the team with the higher $s_T(t)$ at each stage. Bracket predictions select the winner only, so marginal probabilities of winning the round are more appropriate than probabilities conditional on reaching the round. These marginal probabilities must multiply conditional probabilities for the current and previous rounds. For example, if team t is eligible to play in games g_{64}, g_{32}, and g_{16} in the Round of 64, Round of 32, and Sweet 16, respectively, then Eqn. 9 shows the probability that team t is a Sweet 16 winner.

$$\mathrm{P}(w(g_{16})=t)=\mathrm{P}(w(g_{64})=t)\times\mathrm{P}(w(g_{32})=t\,|\,w(g_{64})=t)\times \mathrm{P}(w(g_{16}))=t\,|\,w(g_{32})=t) \quad (9)$$

Later rounds would involve additional conditional probabilities. Conditional probabilities after the Round of 64 must use a scale-mixture of possible opponents. For example, $P(w(g_{32})=t|w(g_{64})=t)$ in Eqn. 9 would mix both teams that team t could play in the Round of 32.

Thus, early-round opponents' strengths are reflected in bracket selections. For example, the final model from Eqn. 10 predicts that in the 2014 tournament, Tennessee had an 82.8% probability of beating Massachusetts and Duke had a 90.3% probability of beating Mercer in the Round of 64. It predicts that Tennessee had a 52.6% probability of beating Duke if it played Duke in the Round of 32, but selects Duke to win in the Round of 32 because it was more likely to be there and had an 81.8% probability of winning if it played Massachusetts.

The dual-proportion model has conditional probabilities in each game that have transitive relationships regarding the more likely winner. If Tennessee is the more likely winner in a Tennessee-Duke match and Duke is the more likely winner in a Duke-Mercer game, then Tennessee is the more likely winner in a Tennessee-Mercer game. If a model is matchup-specific in this regard, such that Team A likely to beat Team B, Team B is likely to beat Team C, and Team C is likely to beat Team A, then optimizing bracket selections requires a much more exhaustive algorithm than what I use. Like Kaplan and Garstka, I use a method that predicts the champion first and works backwards until predicting the Round of 64 winners. By bracket pool rules, it must select a team to win in a round if it already selected the team in a later round. For the remaining picks, Kaplan and Garstka attempted to choose a team based on the expected value of its bracket points in a subsection of the tournament ending with that game. Instead of a subsection of the tournament, I use the expected value of bracket points in that game only. This is a constant times the marginal probability, so I optimize by the marginal probability instead.

Each game's marginal probability is composed of the probability of reaching the game and the conditional probability of winning the game. If one includes the past rounds' expected points, one exaggerates the importance of the probability of reaching the game. The team with the highest probability of reaching the game can be chosen as the winner in the previous round, anyway. Suppose that Teams A, B, C, and D have strengths of 6, 5.8, 5.3, and 0, respectively, using the dual-proportion model. Further suppose that the winner between Teams A and B will play the winner between Teams C and D in the Round of 32. Team A would be selected to win the Round of 32, if using that round only, because its 0.889 expected bracket points surpass Team C's 0.540. Including the Round of 64 would give Team C 1.538 expected bracket points, more than Team A's 1.479. However, the three picks combined would produce only 2.128 expected bracket points, as opposed to 2.477 if selecting Team A twice and Team C once.

Choosing based on one round only already incorporates early-round games because marginal probability incorporates early-round conditional probabilities. What about the future rounds, though? Suppose that we change Team C's strength to 5.7, so it is now the choice for the Round of 32, but Team A has higher conditional probabilities of defeating teams after the Round of 32. This is why I start with the latest round and move to the earliest. Teams A and C would be among eight teams considered for the Sweet 16 selection, and Team A's higher conditional probability would help it to be selected. If neither team is the Sweet 16 selection, then their conditional probabilities after the Round of 32 do not matter, because the points will not be collected, anyway.

My selection method can fail in the following case. Team C would have been selected over Team A for a round, but Team A is selected to win a later round, and the total points are higher with Team C winning the early round and either Team C or a Team E winning the later round. Modify the earlier example so Teams A through D are eligible to play the winner of Teams E through H in the Sweet 16, and that Teams A through H have strengths 6, 5.95, 5.7, −6, 5.7, 5.7, 5.7, and 5.7. Team A would be chosen over Team C because it has 0.208 expected bracket points in the Sweet 16, compared to 0.190 for Team A, which is the second highest. Team C actually has the higher possible points in the Round of 32 and Sweet 16 combined (0.949, as compared to 0.868 for Team A).

If one wants to ensure optimized points, an alternative method could still pick winners based on one round only. However, it would evaluate all games rather than setting a late-round winner to be a winner in all earlier rounds. If a discrepancy exists between an earlier-round game and a later-round game, the combined points for relevant games could be evaluated for potential combinations of predicted winners. One combination selects the predicted later-round winner to win all previous games. The other combinations all select the predicted early-round winner to win the early-round game. These must test both this team and the predicted opponents as winners of later-round games, though. The additional computation is not substantial, but the logic is complicated because these round-based discrepancies can interact with each other. In practice, the steep increase in points by round prevents

reversals in the order of teams' winning probabilities. So, the simplified method may not need corrections for actual tournament layouts. For example, a cross-validated fit is performed with six seasons of data in Section 7. In this, there were no differences between the simplified method of selecting teams and the method that evaluated all discrepancies across multiple rounds.

$$s_T(t)=0.330381s_{RS}(t)+0.027246w_{10}(t) \\ +0.131491\ln(n_3(t)+1)+0.274043a(t) \quad (10)$$

My final model uses the s_T in Eqn. 10. There, $w_{10}(t)$ is the proportion of wins that a team had in games against the teams with the 10 highest ratings $s_{RS}(t)$. For teams that played no games against the top-ten teams, I set the proportion to 0%. The bracket selections were the same with this proportion set to 50% instead, so the assumption does not appear to matter. The rationale behind this predictor is that games against the best teams would generalize better to the NCAA Tournament, which contains the best ten teams in the country, among others. The variable $n_3(t)$ is the number of appearances that a team has made in the past three NCAA Tournaments. The underlying hypothesis is that the players and the fans for programs that appear regularly in the NCAA Tournament are more accustomed to the travel and fanfare of the tournament, and are more ready for it.

The predictor $a(t)$ is an indicator function for playing basketball and football in conferences that are Bowl Championship Series (BCS) automatic-qualifying (AQ) football conferences. In the data, all teams that played football in a BCS AQ conference also played basketball in that conference, so it sufficed to use the football information. This model uses data through the 2013–2014 season, the final season of the BCS, but $a(t)$ could still be created using what were the last BCS AQ conferences, while they exist. One can use a substitute definition of a football "power conference" if conferences undergo substantial change. The rationale was that power conference teams have experience playing for large audiences, and have large fan bases that travel well. The predictor also provides an extra strength-of-schedule adjustment. When the BCS existed, teams in AQ conferences were guaranteed a place in the highest paying bowls (the BCS bowls) if they were their conferences' champions. Some teams in the Football Bowl Subdivision had no conference, and get $a(t)=0$ because independent programs have different dynamics. This includes Notre Dame, which was not in an AQ "conference," but had criteria for inclusion in BCS bowls. The rationale behind using football definitions for basketball predictions is that teams that play both basketball and football in major conferences tend to be more major programs for athletics in general than those that do not sponsor football or do so on a lower level. An example of a team set to $a(t)=0$ for this reason is Villanova, which played basketball in the Big East when it was an AQ conference, but played football in the Colonial Athletic Association.

7 Model comparison

To test the dual-proportion model, I performed six-fold cross validation with the NCAA Tournament data from the 2009 tournament through the 2014 tournament. In each case, I used 1 year of the tournament as the validation set and trained the regression coefficients on the remaining 5 years. The model in Eqn. 10 averaged 103.17 bracket points per year if using the same set for training and validation, and showed a very minor drop-off to 102.5 in cross-validation. I compared the dual-proportion model to models from literature and the popular Pick the Seeds strategy.

Table 1 summarizes the number of bracket points earned by each method. Full Model refers to the dual-proportion model using the predictors from Eqn. 10. Ratings Only is the dual-proportion model using the regular-season ratings as the only predictor. Sagarin, Vegas Odds, and Bradley-Terry are the three models from Kaplan and Garstka (2001). West refers to the model from West (2006). Details of these models' calculations are in Appendix C.

The dual-proportion model outperforms other methods for an average tournament, but not for every tournament. In some cases, the difference is substantial, such as how the four-predictor model outperformed the Bradley-Terry model by more than the value of every game in the Round of 64 combined. The model is even competitive at achieving high log-likelihood, even though it this was not a goal in creating the model. This is probably attributable to how the regression coefficients for the tournament strength parameters are set to maximize log-likelihood in the training set. Table 2 shows the log-likelihoods of each model, which are from cross-validation, except for Pick the Seeds. I used the same probability of the higher seed winning each game, which was 62.93% because it maximized log-likelihood for the full 6-year data set.

8 Competitive strategy

Section 6 presented a model for finding tournament selections based on a higher expected value of bracket points.

Table 1 Bracket points by method in six-fold cross-validation.

Year	Full model	Ratings only	Sagarin	West	Pick the seeds	Vegas odds	Bradley-Terry
2009	132	126	140	126	106	138	94
2010	62	63	81	78	78	61	57
2011	63	62	46	46	57	50.5	43
2012	150	154	135	124	88	66	96
2013	130	112	116	82	79	63	61
2014	78	65	60	64	68	49	57
Mean	102.5	97	96.33	86.67	79	71.25	68

The expected value may not be the appropriate goal, however, because one's goal when entering a bracket is to have the maximum points in the pool, not to have a high number compared to one's own theoretical possibilities. Looking only at the possible points in one's own bracket also overlooks that the competing brackets are not randomly selected from potential brackets. They are made by humans, whose brackets include subjective evaluations of teams, secondary motivations such as choosing teams for which one wants to cheer, and strategies about how to compete with other humans.

A bracket based on expectations alone may be too conservative to win a competition about extreme values. So, in hopes of drawing general conclusions, I tested this bracket alongside 19 general strategies about how to modify a bracket of high-expected-value predictions. I performed Monte Carlo simulations for 2013 and 2014 NCAA Tournaments comparing each of the 20 brackets below to competing brackets from a probability distribution of ESPN users' bracket selections.

1. All Favorites: I selected the highest-probability winners of each round.
2. New Champion #1: The champion from All Favorites is replaced with the 2nd-highest-probability champion.
3. New Champion #2: same as #2, but with 3rd-highest-probability champion
4. New Champion #3: same as #2, but with 4th-highest-probability champion
5. New Champion #4: same as #2, but with 5th-highest-probability champion
6. New Runner-Up #1: The champion from All Favorites is kept the same, but the other team in the championship game is replaced with the 2nd-highest-probability team from the same half of the tournament.
7. New Runner-Up #2: same as #6, but with 3rd-highest-probability runner-up
8. New Runner-Up #3: same as #6, but with 4th-highest-probability runner-up
9. New Runner-Up #4: same as #6, but with 5th-highest-probability runner-up
10. New Final 4 #1: One team from the All Favorites Final 4 is replaced, such that the new four-team set's combined probability of making the Final 4 is 2nd highest.
11. New Final 4 #2: same as #10, but with 3rd-highest-probability set of teams
12. New Round of 32 #1: The four highest-probability upsets replace Round of 32 winners from All Favorites.
13. New Round of 32 #2: same as #12, but with two upsets instead of four
14. New Round of 64 #1: same as #12, but upsets in the Round of 64 instead of Round of 32
15. New Round of 64 #2: same as #14, but with two upsets instead of four
16. Upsets If 48% Probability: This modifies All Favorites starting with the Round of 64. All upsets with higher probability than 48% are chosen.

Table 2 Log-likelihood by method in six-fold cross-validation.

Year	Sagarin	Full model	Ratings only	West	Vegas odds	Pick the seeds	Bradley-Terry
2009	−31.61	−32.43	−32.01	−32.06	−29.40	−39.76	−42.13
2010	−35.02	−37.02	−37.60	−37.15	−43.42	−42.11	−42.06
2011	−39.24	−37.93	−37.48	−37.69	−41.63	−41.88	−42.27
2012	−35.39	−32.98	−33.28	−41.18	−38.35	−40.52	−41.04
2013	−31.60	−37.62	−37.41	−37.16	−40.24	−42.11	−41.88
2014	−35.31	−38.30	−39.16	−35.11	−43.87	−42.94	−41.89
Total	−208.16	−216.28	−216.94	−220.34	−236.91	−249.32	−251.27

17. Upsets If 46% Probability: same as #16, but with a 46% threshold
18. Upsets If 44% Probability: same as #16, but with a 44% threshold
19. Upsets If 42% Probability: same as #16, but with a 42% threshold
20. Upsets If 40% Probability: same as #16, but with a 40% threshold

For strategies #2 to #13, the winner of a game was replaced by a lower-probability substitute team. All previous games needed to make the substitute team reach the game were changed. Strategies #12 to #20 involve upsets, which mean replacing the team from All Favorites if it is present. If it was eliminated in a previous round, the higher-probability team of those present is considered the favorite. The other team winning is considered the upset.

The tournament results were set with probabilities from the dual-proportion model using Eqn. 10. I also performed simulations that used the ESPN distribution for the strategies and tournament results, as well as simulations that used the dual-proportion model for the strategies and an average of the dual-proportion model and ESPN probability matrices to simulate tournament results. These two have been omitted due to space concerns, because they had less useful results. I simulated different numbers of competing brackets to test if larger pool sizes would require riskier strategies. For each year and pool size, I generated 10,000 sets of competitors. For each competitor set, I generated 100 sets of tournament results, creating 1,000,000 simulated tournaments per pool size per year. Ordinarily, pools break ties with a separate question. This was not included, so I recorded two proportions of pools won. One treats ties as a loss, and the other treats ties as a win, with the actual proportion somewhere between the two.

Table 3 shows the results for the All Favorites strategy and for the three strategies that fared best. For ranking strategies, I used averages of the proportions with and without ties included. I also treated strategies as the same if their predictions for the year were all the same. Strategy #1, which is completely based on the expected value, wins a very high proportion of the pools. In 2014, it won 25.45%–26.22% of pools against 40 competitors. This is much higher than 2.44%, the probability if every bracket were equally likely to win. This is not surprising because it knows the assumed probability distribution and the competitors do not. What is surprising is how at least one modification of All Favorites was more effective than All Favorites itself in every simulation. By contrast, the All Favorites strategy was the best in half the simulations completely based on the ESPN distribution, even though it was competing with very similar brackets.

The three best 2013 strategies (#3, #6, and #17) shared a common trait. Compared to All Favorites, they all replaced Kansas in the Final 4 and championship game

Table 3 Best-performing strategies, by number, in Monte Carlo simulation.

Year	Size	Win%, favorites	Best	Win%, best	2nd best	Win%, 2nd best	3rd best	Win%, 3rd best
2013	11	26.52% (28.49%)	6	27.70% (29.47%)	1	26.52% (28.49%)	15	26.48% (28.42%)
	21	17.45% (19.01%)	6	19.87% (21.32%)	17	18.87% (20.20%)	7	17.81% (19.06%)
	41	11.40% (12.59%)	6	13.90% (15.07%)	3	13.60% (14.36%)	17	13.30% (14.43%)
	101	6.73% (7.53%)	3	9.69% (10.35%)	7	8.36% (9.15%)	6	8.16% (9.04%)
	1001	2.12% (2.48%)	3	3.08% (3.48%)	17	2.59% (2.97%)	2	2.32% (2.68%)
2014	11	36.32% (37.31%)	2	39.28% (41.03%)	11	37.23% (38.24%)	7	36.74% (37.51%)
	21	30.57% (31.39%)	10	31.30% (32.15%)	7	31.30% (32.05%)	1	30.57% (31.39%)
	41	25.45% (26.22%)	10	25.77% (26.59%)	1	25.45% (26.22%)	16	25.40% (26.20%)
	101	18.64% (19.55%)	10	18.73% (19.64%)	1	18.64% (19.55%)	16	18.61% (19.51%)
	1001	3.52% (4.15%)	10	3.79% (4.42%)	1	3.52% (4.15%)	16	3.47% (4.09%)

Table 4 Selected probabilities of winning by round, for dual-proportion model.

Year	Team	Round			
		16	8	4	2
2013	Florida	40.9%	23.2%	15.2%	8.0%
	Louisville	54.9%	38.8%	27.8%	19.6%
	Michigan	28.4%	15.4%	9.2%	4.3%
2014	Florida	66.3%	52.0%	39.9%	25.2%
	Michigan State	42.4%	28.2%	12.6%	5.6%
	Virginia	41.7%	26.6%	11.1%	4.6%
	Wichita State	64.4%	52.5%	41.6%	28.9%

with Florida. The best for small pools was #6. The best for large pools was #3, which took the additional risk of switching the champion from Louisville to Florida. Tables 4 and 5 show that choosing Florida to reach the championship has three qualities of interest. It has high probability, it overlaps with many competitors' choices at some point, and it allows the strategy bracket to distinguish itself from the competition at some stage. The only other 2013 strategy (#7) that outperformed All Favorites chooses Michigan rather than Kansas to win in the Sweet 16, Elite 8, and Final 4. This choice has the same three characteristics.

The need to have some overlap is a finding that conflicts with many popular assertions about strategy, and also with those by Metrick (1996). The reason Metrick never reaches this conclusion is because he only inspected one round of the tournament. This partial-overlap strategy works because it is outscoring different groups of competitors during different rounds. The 2013 Strategy #6 attempts to outscore those with unconventional predictions in the championship game by predicting Louisville. It will give itself a chance to outscore very conventional brackets in the Sweet 16 and Elite 8, if Florida can upset better-seeded teams. It can also outscore a large minority of competitors who picked Florida to surpass its seeding-based expectations, if Florida wins in the Final 4.

Table 5 Selected probabilities of winning by round, for competitors.

Year	Team	Round			
		16	8	4	2
2013	Florida	42.0%	20.4%	6.9%	3.1%
	Louisville	81.8%	52.5%	39.6%	21.9%
	Michigan	25.6%	13.3%	4.9%	2.7%
2014	Florida	84.9%	61.9%	41.1%	27.1%
	Michigan State	58.9%	47.1%	22.3%	14.6%
	Virginia	33.5%	22.0%	6.9%	3.8%
	Wichita State	29.2%	18.4%	12.1%	5.9%

The most successful strategies in 2013 are much less successful in 2014, relative to the other strategies. The exception is #7, which picks Virginia to win in the Final 4, replacing Michigan State as a Sweet 16 and Elite 8 winner and Florida as a Final 4 winner. The most successful strategy, #10, places first in all pools except the size-11 pools, where it places second. Strategy #10's changes from #1 are also about picking Virginia, but it stops with Virginia winning in the Elite 8. The big distinction between the years is that for 2013, the dual-proportion model and the ESPN distribution agreed on the best champion, Louisville. For 2014, Eqn. 10 and the ESPN distribution gave similar probabilities for Florida as the champion, but it was the top choice for the ESPN distribution only. Eqn. 10 preferred Wichita State, which it gave much higher winning probabilities after the Round of 64 than the ESPN distribution did. Of the eight teams that could have won the Sweet 16 game in Wichita State could have played, Louisville was the most popular choice with ESPN users (57.9% selected it). Selecting Wichita State even to make the championship game already provided a strategy with a distinguishing characteristic from competing brackets. This also explains the success of Strategy #1 and of #16, which was the same as #1 except that it picked Iowa State rather than Villanova as the victor of one game in the Sweet 16.

An important conclusion from these results is that no strategy of the style used in this simulation is universally superior. This is most evident in the disparities between 2013 and 2014 results. To adjust from the expectation well, one needs to obtain general data reflective of competing brackets for the relevant tournament, such as from ESPN. One then needs year-specific adjustments using the assumed distributions for competitors and actual results. One should avoid widespread selections of upsets, which never did well in the simulation. One upset team appears to be the best approach. It should replace teams for one to three games in rounds after the Round of 32. It needs to follow the three elements that produced success in simulation. The chosen upset must be high-probability. It must allow co-opting a large minority of competing brackets' picks. The resulting bracket must not be so similar to competitors' brackets that it is not distinguished.

9 Conclusion

I have introduced a method for rating teams based on their games in a season prior to the NCAA Tournament, plus a

method including these ratings that predicts winners in an NCAA Tournament bracket. The remaining three predictors in the bracket prediction method were the proportion of wins against top-10 teams, the number of NCAA Tournament appearances in the past 3 years, and an indicator function for being in a BCS AQ football conference. This method outperformed the Pick the Seeds strategy, each of the three Kaplan and Garstka methods, and Brady West's logistic model in average bracket points from six-fold cross-validation for the 2008–2009 to 2013–2014 seasons. Monte Carlo simulation showed that a modification of my model's selections can be better at winning a pool against human competitors. That modification must select a high-probability upset for one to three late rounds, which overlaps a large minority of competitors at some stage but still allows the entire bracket to have moderate late-round differences from competitors' brackets.

Future direction with this work would include replacing the Newton-Raphson optimization from Appendix A with an optimization method such as that described for the regression coefficients in Appendix B. It would also be worthwhile to inspect whether improvement could be found by changing the order of parameters optimized for the ratings from that described in Appendix A. The Monte Carlo simulation could also be revisited once additional years' data are available to see if the modification to the expected-value-based bracket can be made algorithmic.

Appendix A: Implementation of likelihood maximization for ratings

In Section 5, Eqn. 7 showed a log-likelihood function $\ell_R(\boldsymbol{\Omega}|\boldsymbol{D})$ that could be optimized to produce regular-season team ratings for use in tournament prediction. In this, $\boldsymbol{D}$ was a general reference to the data and $\boldsymbol{\Omega}$ was a set of the parameters. There were strength parameters $s(t)$ for team numbers t, one home-court strength parameter s_{HC}, and a scaling parameter β which converted strength differences for use in the big-win term of the log-likelihood function.

Multiple optimization methods are possible, but I ran 500 cycles of optimizing one parameter at a time while holding all others constant. The order I used set the non-Division I rating, then the remaining ratings in alphabetical order, followed by s_{HC} and β. I initialized the parameters s_{HC} and β to 0 and 1, respectively. I then initialized each team strength using Eqn. 11 and constrained the values to be from −6 to 6. Below, $\hat{\pi}(t)$ is team t's proportion of games won.

$$s_0(t)=\begin{cases}\ln 2\hat{\pi}(t), & \hat{\pi}(t)\le 0.5\\ -\ln(2[1-\hat{\pi}(t)]), & \hat{\pi}(t)>0.5\end{cases} \tag{11}$$

For each parameter in each cycle, I ran a Newton-Raphson algorithm to find the spot where each partial derivative of the log-likelihood was equal to zero. This is a maximum, because the log-likelihood is convex, as shown in Eqns. 13, 15, and 17. I use the following first and second derivatives of the log-likelihood for β, which are calculated across the set of games $\mathcal{G}_{\text{BW}}=\{g: b(g)=1\}$.

$$\frac{\partial \ell_R}{\partial \beta}=2\sum_{g\in\mathcal{G}_{\text{BW}}} c(g)d(g)\left[\frac{1}{1-0.5e^{-\beta d(g)}}\mathbf{1}_{(w(g)=h(g))}-1\right] \tag{12}$$

$$\frac{\partial^2 \ell_R}{\partial \beta^2}=-\sum_{g\in\mathcal{G}_{\text{BW}}} \frac{c(g)d(g)^2e^{-\beta d(g)}}{[1-0.5e^{-\beta d(g)}]^2}\mathbf{1}_{(w(g)=h(g))} \tag{13}$$

The remaining log-likelihood derivatives needed are in Eqns. 14–17. They use the functions $\eta(t, g)$, which is 1 if $h(g)=t$ and −1 if $l(g)=t$, and $\nu(g)$, which is 1 if game g is on $h(g)$'s home court, −1 if it is on $l(g)$'s home court, and 0 if it is on a neutral court. The number of games in a season is G, and the sets $\mathcal{G}(t)$ and $\mathcal{G}_{\text{HC}}$ are all of team t's games and all games not played on a neutral court, respectively.

$$\frac{\partial \ell_R}{\partial s(t)}=\sum_{g\in\mathcal{G}(t)} c(g)\eta(t,g)\left[\left(\frac{0.5e^{-d(g)}}{1-0.5e^{-d(g)}}+\frac{\beta b(g)e^{-\beta d(g)}}{1-0.5e^{-\beta d(g)}}\right)\mathbf{1}_{(w(g)=h(g))}-(1+2\beta b(g))\mathbf{1}_{(w(g)=l(g))}\right] \tag{14}$$

$$\frac{\partial^2 \ell_R}{\partial s(t)^2}=-\sum_{g\in\mathcal{G}(t)} c(g)\left[\frac{0.5e^{-d(g)}}{(1-0.5e^{-d(g)})^2}+\frac{\beta^2 b(g)e^{-\beta d(g)}}{(1-0.5e^{-\beta d(g)})^2}\right]\mathbf{1}_{(w(g)=h(g))} \tag{15}$$

$$\frac{\partial \ell_R}{\partial s_{\text{HC}}}=\sum_{g=1}^{G} c(g)\nu(g)\left[\left(\frac{0.5e^{-d(g)}}{1-0.5e^{-d(g)}}+\frac{\beta b(g)e^{-\beta d(g)}}{1-0.5e^{-\beta d(g)}}\right)\mathbf{1}_{(w(g)=h(g))}-(1+2\beta b(g))\mathbf{1}_{(w(g)=l(g))}\right] \tag{16}$$

$$\frac{\partial^2 \ell_R}{\partial s_{\text{HC}}^2}=-\sum_{g\in\mathcal{G}_{\text{HC}}} c(g)\left[\frac{0.5e^{-d(g)}}{(1-0.5e^{-d(g)})^2}+\frac{\beta^2 b(g)e^{-\beta d(g)}}{(1-0.5e^{-\beta d(g)})^2}\right]\mathbf{1}_{(w(g)=h(g))} \tag{17}$$

Eqn. 18 shows how the Newton-Raphson algorithm updates an estimate θ_i for parameter θ at Newton-Raphson iteration i.

$$\theta_i = \theta_{i-1} - \frac{\left.\frac{\partial \ell_R}{\partial \theta}\right|_{\theta_{i-1}}}{\left.\frac{\partial^2 \ell_R}{\partial \theta^2}\right|_{\theta_{i-1}}} \quad (18)$$

The advantage of the Newton-Raphson algorithm is speed. Testing revealed that it was safe to assume convergence after only six iterations, so the run-time was tolerable despite this running for each parameter in a procedure with 500 cycles of 347–354 parameters per cycle. I implemented the algorithm in C++ and ran it on a laptop with a dual-core 2.10 GHz AMD A6 processor and 4 GB of RAM. The 2008–2009 through 2013–2014 seasons ran with a mean of 5 min and 35.5 s of run-time, with a maximum of 6 min (2013–2014).

The disadvantage is that Newton-Raphson only converges to the appropriate position if the derivative of the function for the root-finding (in this case, the log-likelihood derivative) and its derivative are both continuous. For β, this fits because different values of β do not change the indicator function. The $s(t)$ and s_{HC} parameters, however, can cause changes in which team is the higher-strength team and which is the lower strength team, producing discontinuities.

The Newton-Raphson algorithm can still be used, if within a neighborhood where the function for root-finding and its derivative are well-defined. In this application, a game is in this neighborhood the vast majority of the time. The initialization establishes a proper neighborhood for most cases, and continued iterations should move teams away from the discontinuity more often than toward the discontinuity. My hypothesis was that strengths which approach the discontinuity would end up somewhere close to the correct point, and that the next cycle could start in a continuous neighborhood and fix the estimate. If a strength diverged, it would eventually be stopped by the limits of −6 and 6, and could be fixed during the next few cycles. The log-likelihood maximization will not find the global maximum, but it already was limited because I optimize by each of more than 300 parameters individually. Bracket predictions are based on the relationships for pairs of teams, though, so the ratings can still be useful if they converge to areas where the tournament teams' ratings are appropriate relative to each other.

The algorithm converges in log-likelihood. A prototypical pattern was a quick spike, followed by a quick drop and then a climb to a plateau. The algorithm generally found some sort of convergence after 30 cycles, but sometimes had a climb within the plateau that made 500 useful. I used the ending values rather than those with the highest likelihood observed, because they are more stable.

Appendix B: Implementation of regression for tournament strengths

In Section 6, Eqn. 8 shows an equation to produce NCAA tournament strengths $s_T(t)$ for teams t. These are meant to be used within the dual-proportion model from Eqn. 3. Eqn. 8 is a linear combination, but least-squares cannot be used because $s_T(t)$ are used within the non-linear log-likelihood function from Eqn. 3. I could not use the Newton-Raphson algorithm from Appendix A for the β_n. The speed also was not necessary because I always optimized fewer than ten parameters. The discontinuity problems in the log-likelihood function were more severe for these regression coefficients than for the strength parameters. Changes to strengths have a nearly linear effect on log-likelihood in Eqn. 7. By contrast, the regression coefficients β_n are multiplied, amplifying any inaccuracy. Instead of using Newton-Raphson, I initialized each β_n to zero, then optimized the β_n with a brute-force search followed by a modified bisection algorithm. I used 100 cycles of the brute-force search, and 100 cycles of the modified bisection algorithm with 10 bisections per parameter per cycle. In each cycle, I optimized one β_n at a time.

For the brute-force search, I determined a maximum and minimum allowable value of β_n to keep $\beta_n x_n$ from −9 to 9 for all x_n in the data. The choice of −9 and 9 was based on the range of −6 to 6 for regular-season ratings. This way, one team could have a perfect 6 rating and I could still test whether amplifying the predictor was useful. Within these extreme values of each β_n, there were 99 intermediate points creating 100 equally-spaced steps from the minimum to the maximum allowable β_n. I tested the likelihood for each of these 101 values for the coefficient, holding the other coefficients constant, and selected the value with the highest likelihood. The brute-force search tested the same potential values of a coefficient in each cycle, but multiple cycles are necessary to eliminate the effect of the order of predictors optimized.

Generally speaking, the bisection method is a root-finding method like Newton-Raphson is. Its most common application is for the derivative of a log-likelihood function where the log-likelihood function ℓ_T is convex, which is the case here, as shown in Eqn. 19. In these cases, the root for the log-likelihood derivative is the maximum likelihood estimate. For remaining

discussion of the bisection method, I will only address the application to a log-likelihood derivative because more general applications are not relevant to this paper. Below, $\{g\}$ are games numbered from 1 to G. The difference between strength parameters for teams in game g is $d(g)$. The functions $h(g)$, $l(g)$, and $w(g)$ yield the higher-strength team, lower-strength team, and winner of game g, respectively.

$$\frac{\partial^2 \ell_T}{\partial \beta_n^2} = -\sum_{g=1}^{G} \frac{0.5e^{-d(g)}[x_n(h(g)) - x_n(l(g))]^2}{(1-0.5e^{-d(g)})^2} \mathbf{1}_{(w(g)=h(g))} \quad (19)$$

The bisection method starts with an interval where at the start, the log-likelihood derivative is positive, and at the end, it is negative. If continuous, the log-likelihood derivative must pass through zero somewhere in the interval. One divides the full interval in two, and evaluates the log-likelihood derivative at the dividing point between the intervals. If the value is still positive, the maximum likelihood estimate must be in the upper interval. If the value has become negative, the maximum likelihood estimate must be in the lower interval. Either the upper or lower interval becomes the new interval which gets divided in two, and the procedure repeats.

The bisection method assumes a continuous function over the interval, and I achieved this by using small intervals where the brute-force search had placed the interval center in a neighborhood with a continuous log-likelihood derivative. As an added precaution, I replaced the derivative of the log-likelihood at the dividing point with the slope passing through the midpoints of the two subintervals. Because a decrease in log-likelihood from the lower midpoint to the upper midpoint implies a negative slope and an increase implies a positive slope, I forwent the slope calculation and instead compared the two log-likelihoods directly.

The bisection method can only adjust a coefficient slowly because of the small starting interval. In practice, the brute-force search alleviates the need for large adjustments. In every run I checked, the bisection method never moved any coefficients outside of the first bisection cycle's starting interval. The method always converged almost immediately. The dual-proportion model using Eqn. 10 reaches log-likelihood of –212.897 after four cycles, and finishes at –212.896 after 200 cycles. I used the laptop described in Appendix A to test the run-time for the dual-proportion model using Eqn. 10. The run-time for the combined six seasons was 16 s for finding the regression coefficients. After that, calculating the bracket picks and bracket points took less than a second. This used the simplified bracket pick methodology from Section 6.

Appendix C: Implementation of reference models

I needed to replicate the predictors from previous researchers' models for my model comparison in Section 7. Kaplan and Garstka's Las Vegas odds model (labeled Vegas Odds in Tables 1 and 2) required historical point spreads and totals that were not available through their original source. I was forced to use multiple sources. For the 2013 and 2014 tournaments, I averaged the numbers from the Las Vegas Hotel & Casino and the Mirage Hotel & Casino (DonBest 2014). For the 2010, 2011, and 2012 tournaments, I combined the information from many online casinos (OddsPortal 2014). For the 2009 tournament, I relied on information released by SportsBook.Com (1800-Sports 2014). Except for 2009, these numbers were from slightly before the Round of 64 began, so they should be even more market-based than those used by Carlin, Kaplan, and Garstka.

As noted in Section 2.2, Kaplan and Garstka had calculated probabilities of winning for their Las Vegas Odds model using Gaussian CDFs based on parameters λ_i for teams i. The same is true for their Sagarin rating model, which is labeled Sagarin in Tables 1 and 2. These λ_i were expected points for a team, so pre-tournament Sagarin ratings are used directly (West 2014). The Las Vegas Odds model deduces them, starting with half the point total from the team's Round of 64 game. It adjusts by half the point spread, adding this for the predicted winner and subtracting it for the predicted loser (Kaplan and Garstka 2001). Kaplan and Garstka also had a Bradley-Terry model (labeled Bradley-Terry in Tables 1 and 2) fit to all regular-season and conference tournament games in which an NCAA Tournament Round of 64 team or NIT team played against a Division I opponent. In this, 97 teams i had strength parameters s_i: 64 NCAA Tournament teams, 32 NIT teams, and a combined team for other opponents. The probabilities $p_{i,j}$ that team i defeats team j follow Eqn. 20.

$$p_{i,j} = \frac{s_i}{s_i + s_j}, \sum_i s_i = 1, s_i \geq 0 \ \ \forall i \quad (20)$$

For their three methods, Kaplan and Garstka selected predicted winners in the latest rounds first and the earliest rounds last. If a team is selected for a later round, it is selected for all earlier rounds as specified by an algorithm they labeled UNPACK. If a game did not have a predicted winner, they claimed to select the team which was the argument of the maximum for the expected points in the subsection of the tournament ending with that

game. Context suggests that they meant the team with the highest total expected points for the subtournament, if picked to win all games in that subtournament. This is not the argument of the maximum, though, because selecting one team can affect another team's contributions to the expected points for the combined teams.

The model from West (2006) is labeled West in Tables 1 and 2. Rather than recalculate the probabilities, I used West's own work because it follows his intended implementation and incorporates his changes to included predictors. West has performed unpublished work over the years using ordinal logistic regression as in his published model. Beginning with the 2009 tournament, though, he changed predictors from those listed in Section 2. For the 2009 tournament, he added squared terms for each of the original four predictors. Only the squared term for the cumulative points scored minus points allowed was retained for subsequent tournaments. For 2010 and later tournaments, West removed the number of wins over top-30 Sagarin-rating teams as a predictor. In place of that, he added the number of assists per game and the ratio of assists to turnovers.

West provided the model's probabilities, as well as ratings, in a spreadsheet on his web site (West 2014). The regression coefficients are calibrated from previous years' data. So, this model uses the same six validation sets, but validation in time instead of six-fold cross-validation. I use the probabilities West released rather than the ratings, because the probabilities better suit the purpose of making bracket predictions. I use the same simplified procedure used for the dual-proportion model when translating probabilities to bracket predictions. For bracket points, I used the probabilities as-is. West's method is designed for potential use by the NCAA selection committee, not for bracket prediction after the tournament layout has been decided. So, it normalizes probabilities by round for the combined teams that reach the Round of 64, not by the specific matchups (West 2006). For the log-likelihoods in Table 2, the original probabilities are not proper, so I normalized them such that the two teams in each game had a combined 100% probability of winning.

The Schwertman et al. (1991, 1996), and Jacobson et al. (2011) models are not listed separately in Table 1 because their predicted brackets are all identical to those from Pick the Seeds. When a model considered multiple teams equally strong for a bracket selection, I calculated points from each option and averaged the points. This occurred most frequently with Pick the Seeds, which chooses four #1 seeds to make the Final 4 and then has no way to distinguish among them. It thus gets points if any #1 seed wins Final 4 games, but fewer points than a method that predicts the specific team.

References

"2011 NCAAB Tournament Odds 1st Round Match Ups." *Free Sports Picks*. 1800-Sports.Com, n.d. Web. May 29, 2014 (http://www.1800-sports.com/310-800.shtml).

"Basketball Odds Comparison, Basketball Betting Odds & Lines." *OddsPortal.Com*. OddsPortal.Com, n.d. Web. May 28, 2014 (http://www.oddsportal.com/basketball/).

Breiter, David J. and Bradley P. Carlin. 1996. "How to Play Office Pools If You Must." *Chance* 10(1):5–11.

Carlin, Bradley P. 1994. "Improved NCAA Basketball Tournament Modeling via Point Spread and Team Strength Information." *The American Statistician* 50:39–43.

"College Basketball Scores & History." *Sports-Reference.Com*. Sports Reference LLC, n.d. Web. April 30, 2014 (http://www.sports-reference.com/cbb/).

"ESPN – Tournament Challenge – Who Picked Whom." *ESPN*. ESPN, n.d. March 28, 2014 (http://games.espn.go.com/tournament-challenge-bracket/2014/en/whopickedwhom).

Geiling, Natasha. 2014. "When Did Filling Out a March Madness Bracket Become Popular?" *Smithsonian.com*. Smithsonian Magazine, March 20, 2014. Web. May 19, 2014 (http://www.smithsonianmag.com/history/when-did-filling-out-march-madness-bracket-become-popular-180950162).

Jacobson, Sheldon H., et al., 2011. "Seed Distributions for the NCAA Men's Basketball Tournament." *Omega* 39(6):719–24.

Kaplan, Edward H., and Stanley J. Garstka. 2001. "March Madness and the Office Pool." *Management Science* 47(3):369–82.

Koenker, Roger, and Gilbert W. Basset, Jr. 2010. "March Madness, Quantile Regression Bracketology, and the Hayek Hypothesis." *Journal of Business & Economic Statistics* 28:26–35.

Metrick, Andrew. 1996. "March Madness? Strategic Behavior in NCAA Basketball Tournament Betting Pools." *Journal of Economic Behavior & Organization* 30:159–72.

"NCAA – Men's College Basketball Teams, Scores, Stats, News, Standings, Rumors." *ESPN*. ESPN, n.d. Web. April 30, 2014 (http://espn.go.com/mens-college-basketball/).

"NCAA Basketball Betting Odds, NCAA Tournament Point Spreads & Money Lines." *DonBest*. Dodgeball Ventures, Inc., n.d. Web. May 18, 2014 (http://www.donbest.com/ncaab/odds/).

Quintong, James. 2014. "Tournament Challenge: Final Four Update." *College Basketball Nation Blog*. ESPN, April 5, 2014. Web. June 12, 2014 (http://espn.go.com/blog/collegebasketballnation/post/_/id/98226/tournament-challenge-final-four-update-2).

Sagarin, Jeff. "Final College Basketball 2013-2014 Through Results of 2014 April 7 Monday – NCAA Championship." *USA Today*. USA Today, April 8, 2014. Web. May 20, 2014.

Schwertman, Neil C., T. A. McCready, and L. Howard. 1991. "Probability Models for the NCAA Regional Basketball Tournaments." *The American Statistician* 45:35–8.

Schwertman, Neil C., Kathryn L. Schenk, and Brett C. Holbrook. 1996. "More Probability Models for the NCAA Regional Basketball Tournaments." *The American Statistician* 50(1):34–8.

West, Brady. 2006. "A Simple and Flexible Rating Method for Predicting Success in the NCAA Basketball Tournament." *Journal of Quantitative Analysis in Sports* 2(3):n.p.

West, Brady. "New Ratings 2014." *Brady West's Home Page*. University of Michigan, n.d. Web. May 31, 2014 (http://www-personal.umich.edu/bwest/new_ratings_2014.xls).