Improving Bracket Prediction for Single-Elimination Basketball Tournaments via Data Analytics

Abstract

The NCAA men's basketball tournament highlights data analytics to everyday people as they look for help building their brackets. A k-Nearest Neighbors method was developed to improve the estimate of a team's true winning percentage and thus improve the prediction of individual tournament match's outcome. Several strategies were then developed to extend the single-game outcome prediction to the whole-bracket selection. Using data from year 2002-2013, the parameters in the k-Nearest Neighbors method were fine tuned. For the 2014 tournament, our method correctly predicted 42 games, which outperforms direct application of Pythegorean expectation (tabulated on kenpom.com) to the Log5 estimate. Compared to human pools, our method is better than 95% of the brackets submitted to the ESPN Tournament Challenge and the predictions from all credible college basketball analysts.

1. Introduction

Every March in the United States, work productivity dramatically drops as many people turn their attention to the National Collegiate Athletic Association (NCAA) Division 1 (D-I) men's basketball tournament, known colloquially as "March Madness™." This single-elimination tournament consists of a beginning line-up of 64 teams with four subsets (often termed regions) of 16 teams. Within each region, the teams are ranked (or "seeded") from 1 to 16 so that the highly favored teams do not have to compete against each other in early rounds. The tournament seeding and overall bracket are announced on "Selection Sunday", with the first of the 63 games scheduled to start on the following Thursday around noon.[1]

A popular tradition that accompanies this event is tournament pools. In tournament pools, groups of people submit tournament brackets filled in with their predictions of which team they think will win each game. As most pools have

[1] It should be noted that while a recent change in 2011 increased the size of tournament to 68 team (67 games), office pools usually do not include predicting the outcome of the "First Four" games. This delayed realization of the 64-team pool does, however, reduce the time to pick the one perfect bracket from the 9,223,372,036,854,775,808 possible brackets.

Preprint submitted to Elsevier *March 21, 2016*

prizes for the winners, more and more people are turning to data analytics to help them build the perfect tournament bracket. This has never been as true as it was in 2014 when Warren Buffet announced a $1,000,000,000 (USD) prize for anyone who correctly predicted all 63 games' outcomes.

Fig. 1 shows the distribution on the number of games correctly predicted for those 9.22×10^{18} or 2^{63} brackets. Note that for an office pool's brackets to follow this distribution, all the games for all of the brackets would need to be predicted using a fair coin. Because there are some games where the odds are clearly in favor of one team, this distribution does not accurately represent the distribution of most office pools. Fig. 1 also shows the real-life distributions based on the 2013 (with average number of games correctly predicted being about $34 - 35$ games and standard deviation being slightly more than 4; $\mu = 34.76$, $\sigma = 4.10$) and 2014 ($\mu = 35.54$, $\sigma = 4.60$) ESPN Tournament Challenge™ pools.

The main contributions of this paper are twofold, which are predicting the outcomes of single matches and of the entire bracket. First, a k-Nearest Neighbors algorithm was developed to enhance the performance of the Log5 method, a commonly used single-game outcome prediction method. The developed clustering algorithm is used to quantify the win-loss relationship between any potential match in the tournament based on how each team involved in the match played against opposing teams of *similar style* during the regular season. This alternative method is expected to successfully replace the Pythagorean expectation, a current method in college basketball analytics, which considers a team's winning percentage over the "average" NCAA D-I team. Second, three strategies were considered when extending the single-game prediction to the bracket selection, focusing either on each individual game's outcome in a round by round fashion, or the winners of particular spots in the bracket (i.e. overall champion), or all winners throughout some portion of the bracket. Using data from the 2002-2013 NCAA tournaments, the parameters in the k-Nearest Neighbors algorithm were tuned specifically for an upcoming year's NCAA tournament. For the 2014 tournament, the algorithm was able to achieve a better bracket prediction than 95% of the brackets submitted to the 2014 ESPN Tournament Challenge™, and 20 percentage points higher in the tournament ranking than the one using the current Pythagorean expectation based method.

The remainder of this paper is organized as follows. In Section 2, background is given on tempo-free statistics and current methods for predicting single-game outcomes. Section 3 presents the details of the k-Nearest Neighbors algorithm. This is followed by the extension of single-game prediction to bracket selection in Section 4. Next, various bracket selection methods are compared on the 2014 NCAA tournament in Section 5. Finally, concluding remarks are provided and future research is outlined in Section 6.

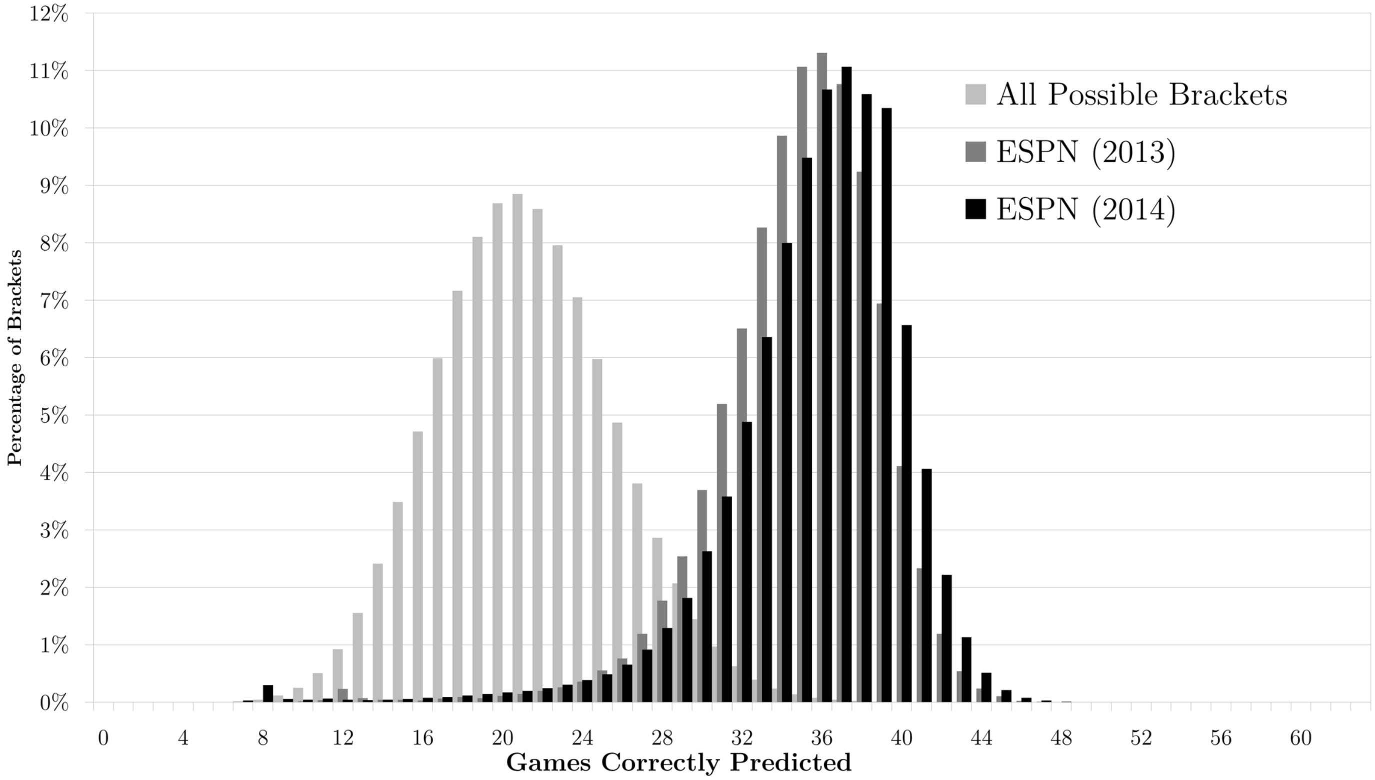

Figure 1: Distribution of the number of games correctly picked for all possible brackets along with the corresponding distributions in the 2013 and 2014 ESPN Tournament Challenge™

2. Background

2.1. Tempo-Free Statistics

Traditionally in basketball, team and player statistics are averaged on a "per-game" basis. As such, when looking at a stats page on the NCAA D-I men's basketball website, one will find such statistics categories as PPG (i.e., points per game), OPP PPG, (i.e., opponent's points per game), APG (i.e., assists per game), and RPG (i.e., rebounds per game). These statistics, however, are highly dependent on a team's tempo. The tempo of a basketball team is defined by the number of possessions a team has in a game. In NCAA D-I men's basketball, each team has 35 seconds[2] to shoot the ball and at least hit the rim; otherwise, the ball is turned over to the other team due to a so-called "shot clock violation." Some teams pass the ball frequently, using nearly all of the 35 seconds looking for the best open shot; whereas other teams try to make the first available shot they have. As basketball can be viewed as a turn-based game where every possession ends with points, a defensive rebound off of a missed shot, or a turnover before possession is transferred to the other team, the two teams will have nearly the same number of possessions in their game. Nevertheless, as the number of possessions in a single game may vary significantly from team to team and even between the same team in different games, there has been a shift by some people away from per-game statistics in exchange for "tempo-free" statistics, which are independent of a team's tempos (or number of possessions per game).

The notion of tempo-free statistics has been promoted by analysts like Ken Pomeroy, who runs a website that provides tabulated tempo-free statistics (kenpom.com). It is argued that tempo-free statistics offer some sense of how well a team would fare against the "average" team in some category of the game stats. This further implies that tempo-free statistics provide attributes for estimating the true winning percentage of a team against the "average" team. A true winning percentage is the percentage of games a team *should* win if they were to play sufficiently many games as to average out any amount of luck. Given that each D-I team has typically faced a good sample of D-I opponents throughout the course of the regular season, it is argued that each D-I team's tempo-free statistics can be used to construct a good estimate for the team's true winning percentage against the "average" D-I team. Subsequently, one can better compare any arbitrary pair of teams in their hypothetical head-to-head match-up, based on the two teams' true winning percentages.

As basketball consists of offense and defense, one should include both offense and defense related tempo-free statistics in the winning percentage estimation. The adjusted offensive efficiency, denoted by $AdjOE$, is an estimate of the number of points a team would score against a team with "average" D-I defense per 100

[2] The NCAA started to implement a 30-second shot clock in the 2015-16 season.

possessions. In other words, $AdjOE$ is the tempo-free statistics on scoring. In the 2014 tournament, Creighton, a No. 3 seed, had an $AdjOE$ at 125.7 points scored per 100 possessions, which is the highest (or best) among all teams in the tournament, while Coastal Carolina, a No. 16 seed, had the lowest $AdjOE$, at 97.3 points scored per 100 possessions.

A simple example using tempo-free statistics may be helpful. Suppose Team A and Team B are going to play each other. Looking over the team statistics, both teams average 80 points per game. They seem like a very even match-up; however, their tempos are 60 and 80 possessions per game, respectively. This is equivalent to saying that Team A would average 125 points per 100 possessions while Team B would average 100 points per 100 possessions. When the two teams played each other, they would have nearly the same amount of possessions, save for who starts and ends the game with the ball. As such, if either team's average points per possession hold, then tempo-free statistics of the two teams suggests that Team A would beat Team B because Team A is more effective at converting possessions to points than Team B, given all other things being equal.

While this does help make the case for using tempo-free statistics, there is more to basketball than just offense. The adjusted defensive efficiency, denoted by $AdjDE$, is an estimate of the number of points a team would allow against a team with "average" D-I offense per 100 possessions. $AdjDE$ is essentially the tempo-free statistics on points allowed. Arizona, a No. 1 seed in the 2014 tournament, had the best $AdjDE$ with only 86.9 points allowed per 100 possessions, while Eastern Kentucky, a No. 15 seed, had the worst adjusted defensive efficiency with 108.3 points allowed per 100 possessions. Following the earlier example, when assuming the defenses of the two teams were equal, it was clear that Team A should be predicted to win. Consider now, however, that Team A allowed on average 105 points per 100 possessions while Team B only allowed 95. It is no longer clear who would win as the superior defense of Team B may be enough to slow down the scoring from Team A. As such, the next section will discuss the existing methods of predicting single-game outcome.

2.2. Single-Game Outcome Prediction

2.2.1. "Chalk"

The first, and perhaps simplest, idea for predicting single-game outcome is picking "chalk". This is a term in sports betting that refers to picking the team that is favored to win based on the betting odds (Tracy, 2013). In the case of NCAA tournament, each team is assigned a seed number that reflects the opinion of experts on which team they would favor. Thus the "chalk" idea would always choose the higher seeded team to win for each head-to-head match-up. Consequently, the tournament bracket is uniquely specified according to the teams' seeding. This simple idea typically works well in small office pools.

2.2.2. The Log5 Estimate

Let $P(A \succ B)$ be the probability that Team A would beat Team B given their respective true winning percentages against the "average" D-I team, $\bar{p}_A$ and $\bar{p}_B$. The Log5 estimate, proposed by Bill James for use in baseball, is used to estimate the single-game outcome in terms of the winning probability of either team (James, 1981). With Log5, $P(A \succ B)$ is estimated as:

$$P(A \succ B) = \frac{\frac{\bar{p}_A}{1-\bar{p}_A}}{\frac{\bar{p}_A}{1-\bar{p}_A} + \frac{\bar{p}_B}{1-\bar{p}_B}} = \frac{\bar{p}_A(1-\bar{p}_B)}{\bar{p}_A(1-\bar{p}_B) + \bar{p}_B(1-\bar{p}_A)}. \tag{1}$$

It is suspected that the name Log5 comes from the resemblance the formula has to the logit function and the fact that it assumes the winning probability of the "average" team in a league is 0.5. Note that this formula satisfies the following properties: $0 \leq P(A \succ B) \leq 1$; $P(A \succ A) = 0.5$; $P(B \succ A) = 1 - P(A \succ B)$. In addition, if $\bar{p}_A = 1$ or 0, then one has $P(A \succ B) = 1$ or 0 for any $\bar{p}_B \neq 0$. If $\bar{p}_B = 0.5$, then $P(A \succ B) = \bar{p}_A$. Finally, if $\bar{p}_A \geq \bar{p}_B$, then $P(A \succ B) \geq 0.5$. For information on why these properties intrinsic to the Log5 estimate make it ideal for the winning probability estimation, the reader is referred to Miller (2008) and Hammond et al. (2014).

Although one might be able to demonstrate that using the Log5 estimate is the best choice in predicting the single-game outcome for the head-to-head match-up between Teams A and B, this method heavily relies on the accurate estimation of each team's true winning percentage. However, this percentage number is purely theoretical and not readily known. Currently, it is commonly estimated with the Pythagorean expectation in college basketball analytics.

The Pythagorean expectation ($Pyth$) is a descriptive statistic designed to combine offensive and defensive metrics, or the adjusted offensive and defensive efficiencies for a D-I team in this case. It is considered an estimate of a team's true winning percentage against an "average" D-I team, i.e., $\bar{p}_A$ for Team A as denoted earlier. For the Pythagorean expectation tabulated on kenpom.com, the formula is

$$Pyth = \frac{AdjOE^{11.5}}{AdjOE^{11.5} + AdjDE^{11.5}}, \tag{2}$$

where the exponent of 11.5 was fitted by Ken Pomeroy using data from 2002 – 2013 seasons. Note that the Pythagorean expectation may vary for different sports by the exponent, but its value is always bounded between 0 and 1. Since the components (i.e., $AdjOE$ and $AdjDE$) in the Pythagorean expectation for each team are independent of the potential tournament match of the team, we speculate the potential improvement on this by considering specific tournament matches. In the next section, we thus propose a replacement of Pythagorean expectation in the Log5 estimate with the hope of better predicting single-game outcomes.

3. A k-Nearest Neighbors Method

k-Nearest Neighbors (kNN) is a clustering analysis method used to determine the membership of each entity based on the attributes of the entity relative to those of its k neighbors (Dudani, 1976). In this paper, the attribute used is the win/loss outcome of a game to the considered team against some opponent in the tournament, with neighbors being opponents that were faced by the team during the regular season. For example, both teams A and B will play in the tournament. To estimate the chance that each team may win or lose in the potential match-up, we consider how teams A and B have fared in the regular season against teams *similar* to teams B and A, respectively. To substantiate the estimation, a distance-based similarity metric is introduced for the k-Nearest Neighbors algorithm. The developed method is expected to improve the prediction of each possible tournament game's outcome, and in turn, improve the prediction of the perfect tournament bracket.

Let $\theta_{A,B}$ denote the win/loss outcome of a game where $\theta_{A,B} = 1$ if team A beats team B and 0 otherwise. Let Ω_A denote the set of opponents faced by team A during the regular season. Based on some similarity metric, let $\Omega^k_{A,B}$ denote the set containing the k opponents most similar to team B that team A played against in the regular season. Then two similarity-weighted measures associated with teams A and B, $w_{A,B}$ and $l_{A,B}$, are defined to take into account the win and loss outcomes from team A playing against its regular season opponents from set $\Omega^k_{A,B}$, i.e. $w_{A,B} \equiv \sum_{T \in \Omega^k_{A,B}} d_{B,T}\, \theta_{A,T}$ and $l_{A,B} \equiv \sum_{T \in \Omega^k_{A,B}} d_{B,T}\, (1 - \theta_{A,T})$, where $d_{B,T}$ is a similarity index between team B and team T. Similarly, the two measures are defined for team B based on its playing against some of its regular opponents, i.e., $w_{B,A}$ and $l_{B,A}$. Then $p_{A,B}$, an alternative winning percentage considering not only team A but also team B, a particular opponent of team A, is estimated as:

$$p_{A,B} = \frac{w_{A,B} + l_{B,A}}{w_{A,B} + l_{B,A} + w_{B,A} + l_{A,B}}. \tag{3}$$

One can then replace $\bar{p}_A$ and $\bar{p}_B$ with $p_{A,B}$ and $p_{B,A}$ in the Log5 estimate. In other words, the estimated winning percentage against the "average" D-I team is replaced by the estimate of a more specific winning percentage of team A against team B, two tournament teams, through synthesizing the realized outcomes of teams A and B against teams similar to them in the regular season.

To help illustrate the difference created by the kNN method, we use the first-round game in the 2014 tournament between fifth-seeded Cincinnati and twelfth-seeded Harvard as an example. We simply set $k = 10$ in this example, and let $d_{B,T}$ be defined as in Equation (4). Cincinnati finished the regular season with a win-loss record of 27-6 ($Pyth = .8701$) and Harvard finished with a record of 25-4 ($Pyth = .8402$). With the Log5 estimate using Pythagorean expectation, one would give Cincinnati a 56% chance of beating Harvard. However, when examining the 10 most similar opponents in the regular season for each team (i.e., $k = 10$), we found that Cincinnati's record against the 10 teams that they had played and were most similar to Harvard is only 5-5. This is compared to Harvard's record of 8-2. Using our method above, Cincinnati's expected winning percentage against teams similar to Harvard is calculated as $p_{c,h} = .3760$ while that same number for Harvard is $p_{h,c} = .6240$. Using these numbers in the Log5 estimate would give Harvard a 73% chance of beating Cincinnati. When the game was played, Harvard did in fact upset Cincinnati.

Our kNN method provides flexibility for better empirically based learning given available historic data. More specifically, the method allows the specification of a similarity metric to determine which teams are regarded similar and the specification of k on the number of similar teams to examine. In the subsections that follow, historical data from the past tournaments were used to make the specifications with the objective of maximizing the number of correctly predicted single games in the past tournaments. For computational reasons, we first specified the value of k in the kNN method with the similarity metric fixed to some intuitively promising one. We then fixed k to its selected value and identified an appropriate similarity metric with promising prediction power.

3.1. Specifying an appropriate k value

To specify the number of neighbors, we fixed the similarity metric as follows. For each pair of tournament teams A and B, we specified its similarity metric with the normalized difference between the two teams' Pythagorean expectations. It is presented as:

$$d_{A,B} = 1 - \frac{|Pyth_B - Pyth_A|}{Pyth_{\max} - Pyth_{\min}}, \tag{4}$$

where $Pyth_{\max}$ and $Pyth_{\min}$ are the largest and smallest Pythagorean expectations among all the tournament teams.

3.2. Determining an appropriate similarity metric

After k was fixed, we next determined an appropriate similarity metric. We generalized the definition of the similarity metric in (4) to the following: $d_{A,B} = 1 - \sum_i b_i |\xi^i_A - \xi^i_B|$, where i is the index of some statistic used to compare teams when making win/loss outcome predictions, and ξ^i_T is the normalized (or scaled) statistics. These statistics categories could include the Pythagorean expectation introduced earlier, as well as many others, both tempo-free and per-game ones. In (4), we essentially let $b_i = 1$ only for the index i associated with the Pythagorean expectation and let $b_i = 0$ for all other categories. Thus, $\xi^{Pyth}_T = \frac{Pyth_T}{Pyth_{\max} - Pyth_{\min}}$.

To determine an appropriate similarity metric, we considered eight non-traditional statistics that do not appear in frequently visited sports statistics sites but attainable on kenpom.com. These statistics are *Effective field goal percentage* [3], *Turnover percentage* [4], *Offensive rebounding percentage* [5], *Free throw rate* [6]. In our analysis, we considered them in both offense and defense categories.

With these statistics plus the Pythagorean expectation, $AdjOE$, and $AdjDE$, we introduced a similarity metric with 11 b_i's. We essentially used b_i's to quantify relative importance of the statistics selected in judging the similarity between teams, and subsequently, to determine for each tournament team the similarity between the teams it might play in the tournament and the teams it already played in the regular season. We formalized an optimization problem to determine the optimal b_i's such that some proxy for the number of correctly predicted tournament games is maximized. Let I be the set of indices for the statistics selected, $\mathcal{M}$ be the set of all tournament teams, k^* be the appropriate k value determined earlier. We present the mathematical formulation of the optimization problem as follows:

$$maximize \qquad \sum_{A\in\mathcal{M}} \sum_{B\in\mathcal{M};B\neq A} \theta_{A,B}\ x_{A,B} \tag{5}$$

subject to

$$\sum_{i\in I} b_i = 1; \tag{6}$$

$$d_{B,T} = 1 - \sum_{i\in I} |\xi^i_B - \xi^i_T| b_i, \qquad \forall A,B\in\mathcal{M}; T\in\Omega^{k^*}_{A,B}; \tag{7}$$

$$x_{A,B} \leq 1 + \frac{1}{k^*}\left\{ \sum_{T\in\Omega^{k^*}_{A,B}} \theta_{B,T}\ d_{B,T} - \sum_{T\in\Omega^{k^*}_{B,A}} \theta_{A,T}\ d_{A,T} \right\}, \qquad \forall A,B\in\mathcal{M}; \tag{8}$$

$$x_{A,B} + x_{B,A} = 1, \qquad \forall A,B\in\mathcal{M}; \tag{9}$$

$$0\leq b_i\leq 1, x_{A,B}\in\{0,1\}, \qquad \forall i\in I, A,B\in\mathcal{M}. \tag{10}$$

In the above formulation, Greek symbols represent given model parameters and Roman symbols represent unknown decision variables. For each pair of tournament teams A and B, we consider $x_{A,B}$ to be a proxy of $P(A\succ B)$ for two

[3] Effective field goal percentage is similar to regular field goal percentage except that it gives 50% more credit for made three-pointers, i.e., $eFG\% = \frac{1.5FGM_3+FGM_2}{FGA}$, where FGM_3 and FGM_2 are numbers of three-pointer and two-pointers made, respectively, and FGA is the number of field goals attempted.

[4] Turnover percentage is a tempo-free measure of ball security, i.e., $TO\% = \frac{TO}{Possessions}$, where TO is the number of turnovers committed.

[5] Offensive rebounding percentage quantifies a team's ability to extend its play, i.e., $OR\% = \frac{OR}{OR+DR_{opp}}$, where OR indicates the number of offensive rebounds and DR_{opp} indicates the number of defensive rebounds grabbed by the opponent.

[6] Free throw rate quantifies a team's ability to get to the free-throw line, i.e., $FT\% = \frac{FTA}{FGA}$, where FTA and FGA indicate the number of free throws attempted and the number of field goals attempted, respectively.

reasons. First, $x_{A,B}$ only accounts for the win outcomes of the two teams (i.e., $w_{A,B}$ and $w_{B,A}$); see Constraints (8). Second, it simply compares $w_{A,B}$ and $w_{B,A}$ to determine which team to win (i.e., $x_{A,B} = 1$ or $x_{B,A} = 1$). The introduction of this proxy makes the optimization problem more computationally convenient because it is a mixed-integer linear program. With the specification on $x_{A,B}$ and $x_{B,A}$, it is easy to see that objective function (5) approximates the number of tournament games correctly predicted.

4. Bracket Selection

Up to this point, we have focused entirely on improving the prediction of single-game outcomes, which is the backbone of ensuring excellent bracket selection. In this section, we extend the single-game prediction to the bracket selection, which is rather the ultimate goal of this paper. After surveying the literature, we concluded that this area is much less developed compared to single-game predictive analytics. We propose three different strategies on the extension, namely the *round-by-round* strategy, the *overall champion* strategy, and the *most likely bracket* strategy.

The round-by-round strategy selects the single-game outcomes in each round at a time. As each game in a round is independent of the others, the game can be individually predicted based on various single-game prediction methods described earlier. This selection process starts from round 1 with 32 single games and form the 16 match-ups in the second round with the selected winners. This process is repeated until a complete bracket is obtained. In some sense, the round-by-round strategy is a kind of “chalk”, but differs with the classic one in choosing the winners for individual matches.

The overall champion strategy calculates each team’s probability of winning the entire tournament and selects the team with the highest chance of winning the tournament as the projected overall champion. Let us use a 4-team single-elimination bracket to illustrate the calculation. Suppose teams A, B, C, D form a 4-team bracket, teams A and B play and teams C and D play in the first round, and then the two winners play the second round to determine the eventual champion. Then the probability that team A wins it all is $P(A \succ B) \times (P(A \succ C|C \succ D) + P(A \succ D|D \succ C))$. Note that each individual game’s outcome can be again predicted by different methods described earlier. When the overall champion is decided with the highest such probability, the projected team would also win all the games en route to the championship. For the real NCAA tournament bracket, this process is repeated for the remaining spot in the finals and the 2 spots left in the semi-finals (i.e., the three other regional champions), and so on, until the entire bracket is filled.

The most likely bracket strategy calculates the likelihood that each bracket occurs and selects the bracket with the highest chance of occurrence. Let us again use a 4-team single-elimination bracket to illustrate the calculation. Suppose teams A and C win their respective first round matches, and when they play in the second round, team A wins. Then the probability that such a bracket occurs is $P(A \succ B) \times P(C \succ D) \times P(A \succ C|A \succ B, C \succ D)$. With the independence assumption, this expression is further written as $P(A \succ C, A \succ B, C \succ D)$. With a 64-team single-elimination tournament, if a computer were able to calculate the occurrence probability of 1,000,000,000 brackets every second, it would take the computer nearly 300 years to finish all the calculation. To help handle this very computationally intensive task, the entire bracket is divided into smaller brackets. In our implementation, we made the following specification on the division. We applied the most likely bracket strategy to four regional 16-team brackets. Then we applied the strategy to the final-four bracket with the four previously selected regional champions.

Since the most likely bracket strategy looks at the overall likelihood of a bracket occurring, it has the potential to make the prediction of some games rather at odds compared to only focusing on independent prediction for each of the games. In other words, the likelihood of a bracket where one underdog team does advance may be greater than that of a bracket where the team is not chosen to advance, which encourages to pick upsets. This is considered a distinction from the previous two strategies and the distinction can be profound in the following situation. For a early-round match that is close, it is possible that the team with the lower probability for winning creates an extremely lop-sided win later. With the most likely bracket strategy, we would make that team to advance, thus producing an overall more likely bracket. On the other hand, human intuition for this case would likely suggest it would be a poor choice to pick a team to advance through later rounds when it was predicted to lose an earlier round.

The major difficulty of the most likely bracket strategy is dealing with small probabilities resulted from multiple products. While it is easy to feel confident in predicting an outcome of a single match when the probability split is 90% likelihood of team A winning to 10% for team B, most people would hesitate if the split were closer to 51% to 49%. Suppose there are 63 independent games that each had that same 90/10 split. The probability of predicting all 63 games correctly would be $(.9)^{63}$ or .13% . Even such a rather idealized bracket, the probability of it occurring is smaller than that with which most people are comfortable.

5. Numerical Studies

For our numerical experiments, we extracted regular-season and tournament data of 2002 – 2014 basketball seasons, which are attainable from public sources such as (espn.com) and (kenpom.com). Note that the regular-season data from each year are aggregate data at the end of each regular season. We used the data from the 2002 – 2013 basketball seasons as the training data, similar to the model learning conducted by Ken Pomroy. We used the data of Year 2014 as the test data.

We followed the two-step training procedure described in Section 3. In the first step, we varied the value of k between 3 and 30. We defined the training performance measure to be the number of games correctly predicted using the k value in the k-Nearest Neighbors method and the definition of $p_{A,B}$ in the Log5 estimate (i.e., our method) minus the number of games correctly predicted using Pythagorean expectation directly in the Log5 estimate (i.e., benchmark method). We considered the aggregate performance for all the 2002 – 2013 NCAA D-I men's basketball tournaments. From our numerical studies, we specified the value of k to be 25. Note that this value turned out to yield the best performance with both the round-by-round strategy and the overall champion strategy.

In the second step, we fixed k and solved the optimization problem (5) – (10). For implementation convenience, we employed a rather brute-force explicit enumeration approach, which is ensured by our access to high performance computing. We initially explored all the publicly available statistics with a total of 16 of them. After realizing the relative importance of the 11 statistics mentioned earlier, we varied b_i's in a 11-dimensional unit hypercube with 0.01 apart along each dimension. With fixed values on b_i's, the kNN algorithm was performed and the number of correctly predicted games were tallied. Interestingly, assuming only the Pythagorean expectation in the similarity metric, i.e., $b_{Pyth} = 1$, and $b_i = 0$ for all other i's, turned out to be the best option found.

The trained model was then applied to the 2014 tournament's data. Table 1 reports the number of games correctly predicted by each method (bracket selection strategy + single-game outcome prediction method), as well as some prominent predictions made by convention or experts. The ESPN National Bracket compiles brackets submitted to the ESPN Tournament Challenge™ by using a team's selection to advance on each bracket entry as one vote for that team to advance on the compiled bracket. For example, in a first-round match, 65% of brackets submitted predicted No. 9 seeded Oklahoma State to defeat No. 8 seeded Gonzaga. As such, the National Bracket picked Oklahoma State to advance to the second round. Repeating this for all 63 games yields the ESPN National Bracket.

For the 2014 tournament, using the kNN method in the Log5 estimate for the single-game prediction and the overall champion strategy for the bracket selection, we were able to predict 42 games correctly, better than 95% of all brackets submitted to that year's ESPN Tournament Challenge™ and better than all the brackets provided by the ESPN analysts examined. The combination of the kNN method and each of the other two bracket selection strategies yielded slightly inferior prediction performance. But these two predictions still outperformed the existing brackets, with correctly predicting 41 and 40 games, respectively. Note that with either round-by-round or overall champion bracket selection strategy, the Pythagorean expectation based method yielded 38 games

Bracket	# of Games Correctly Predicted
Overall Champion + Log5 w/ kNN	42
Most Likely Bracket + Log 5 w/ kNN	41
Round-by-Round + Log5 w/ kNN	40
Overall Champion + Log5 w/ Pyth Exp	38
Round-by-Round + Log5 w/ Pyth Exp	38
Joe Lunardi (E)	40
Jay Bilas (E)	39
Dick Vitale (E)	39
Seth Greenberg (E)	37
The ESPN National Bracket	39
The “Chalk” Bracket	39

Table 1: Comparison of the games correctly predicted in various 2014 Tournament brackets. ESPN Basketball Analysts are marked with an (E)

correctly predicted, which is equivalent to 75 percentile of all the brackets submitted to the 2014 ESPN Tournament Challenge™ or 20 percentage points lower than the proposed kNN method.

We next compared the two single-game prediction methods, namely our proposed kNN method and the Pythagorean expectation based benchmark method. When comparing the 32 games in the first round with the opponents fixed *a priori*, the kNN method outperformed the benchmark method by correctly predicting 27 games versus 26 games. In addition, prediction errors in early rounds propagate through later rounds. Thus we looked specifically at the 63 games that actually occurred to compare the two methods. In this case, the kNN method again performed better than the benchmark method for correctly predicting 44 games versus 41 games.

Finally, we compared the brackets in more detail. We report the three brackets via our proposed approaches in Figures 2 – 4, and report the true 2014 tournament bracket in Figure 5. Our results suggested many similarities among the three predicted brackets as they used the identical method to make single-game predictions. For instance, they predicted identical "elite eight" teams. In addition, they all made correct predictions on several underdogs winning their respective matches, e.g., 12-seed S. F. Austin beating 5-seed VCU and 12-seed Harvard beating 5-seed Cincinnati. Nevertheless, the proposed approaches would not likely predict big underdogs going deep in the tournament. So in a year like 2014 when a 7th seed and a 8th seed met in the final, the prediction outcome was rather satisfactory. We expect our bracket selection strategies would work better in tournaments with early-round upsets but the seeding order would at end prevail.

6. Conclusions and Future Research

In this paper, we present novel approaches for improving bracket selection of NCAA basketball playoff tournaments. We propose a k-Nearest Neighbors method to identify the similarities between regular-season opponents of each tournament team and its potential match-ups in the tournament. We use historic data to tune the number of neighboring opponents to consider and the weights on different basketball statistics to assign. We consider three strategies for extending the single-game outcome prediction to the bracket selection. Our kNN method outperforms the currently employed Pythogarean expectation based method. Our consideration on the bracket selection strategy essentially fills the vacancy in this area of study.

We acknowledge that we are still far from capable of building a perfect bracket. However, our effort should further promote the use of in-depth analytics in single-elimination tournament bracket selections. For example, with the setup of a user-friendly prediction aid website, we can expect everyday people to employ our prediction analytics in each future March to help build their own brackets in any form of NCAA basketball tournament pool competitions. For competitions where participants are allowed to submit multiple entries, the collection of our approaches is expected to give the users sufficient flexibility.

The key to successful building of a perfect bracket is excellent prediction on individual matches. While we consider this study a comprehensive one in that regard, we still could not possibly explore all the college basketball stats websites nor study the importance of each and every statistics category. We plan to collect more statistics for which data can be made available to our analysis and plan to consult expert basketball analysts to screening these statistics. By the same token, we plan to look for more historical data from earlier seasons. We suspect that one reason for the still to-be-improved performance is that we still do not have a large enough sample size on the historic data to combat the noticeable year-by-year variation in the NCAA tournaments.

Next, we would like to point out one specific major limitation in this study. That is, we currently do not have data collected during a regular season and additional information on those unexpected events such as key player's injury and suspension. For more sophisticated predictive analytics in the future, we certainly should find ways to record the statistics and calculate the metrics periodically throughout a regular season. We are almost certain that Ken Pomroy incorporated the temporal variation on each team's performance in the training of the Pythagorean expectation model as it was indicated on the website (kenpom.com) that he collected data, performed model training, and made periodic prediction throughout a regular season. We are confident that we will be able to further improve the single-game prediction once we have the same amount of information.

Meanwhile, this limitation may also explain the robustness of the "chalk" bracket. Compared to the kNN and benchmark methods that had years of good and bad performance, the "chalk" bracket consistently did well. We speculate this is because these methods do not account for the aforementioned information while it is intrinsically included in the expert rankings and thus reflected in the tournament seeding. This limitation could be addressed in future research, either through adding a temporal weight into the similarity metric or incorporating the seeding into the similarity metric.

In terms of bracket selection, we will first conduct a more careful study on the overall champion strategy. This strategy is much less computationally challenging than the bracket selection strategy, and meanwhile, it is more likely to preserve the continuity of good teams' winning, compared to the round-by-round strategy. One can easily see the generalization of the overall champion strategy by considering some portion of the bracket. Instead of selecting the overall champion, one can perform the calculation to select the team with the highest predicted probability of winning any portion of the bracket, e.g., the team advancing to *final four* from its region. Then the other single matches in the region can be predicted similarly as above. Finally, we treat the final-four bracket separately and make the predictions.

It is clear the three bracket selection strategies may fare differently for tournaments in different years that differ by the number of potential upsets and the significance of the parity in the tournament. With careful selection of the perspective for the tournament to be predicted, the prediction effect is expected to be further improved. In our future research, we will look for features that can give accurate indication on the behavior of each strategy.

Last but not least, it is of authors' great interest to apply the proposed methodology in this paper to other single-elimination tournaments, e.g., NBA/NHL playoff season and FIFA World Cup knock-out stage. With more days to perform the analytics and fewer teams to deal with, we expect to have better performances than NCAA tournament predictions. Of course, the challenge on predicting results of other sports will arise in deep quantitative understanding of those sports.

Acknowledgment

The authors would like to thank Mr. Ken Pomeroy for providing the historic data necessary to this research.

References

S. A. Dudani. The distance-weighted k-nearest-neighbor rule. *IEEE Transactions on Systems, Man, and Cybernetics*, SMC-6(4):325–327, Apr. 1976. ISSN 0018-9472. doi: 10.1109/TSMC.1976.5408784. URL http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=5408784.

C. N. B. Hammond, W. P. Johnson, and S. J. Miller. The James Function. *arXiv preprint arXiv:1312.7627v4*, pages 1–18, 2014.

B. James. *1981 Baseball Abstract*. Self-Published, Lawrence, KS, 1981.

S. J. Miller. A Justification of the log 5 Rule for Winning Percentages, 2008. URL http://web.williams.edu/Mathematics/sjmiller/publichtml/103/Log5WonLossPaper.pdf.

M. Tracy. What We Mean When We Say 'All Chalk', 2013. URL http://www.newrepublic.com/article/112740/all-chalk-jargon-march-madness-brackets.

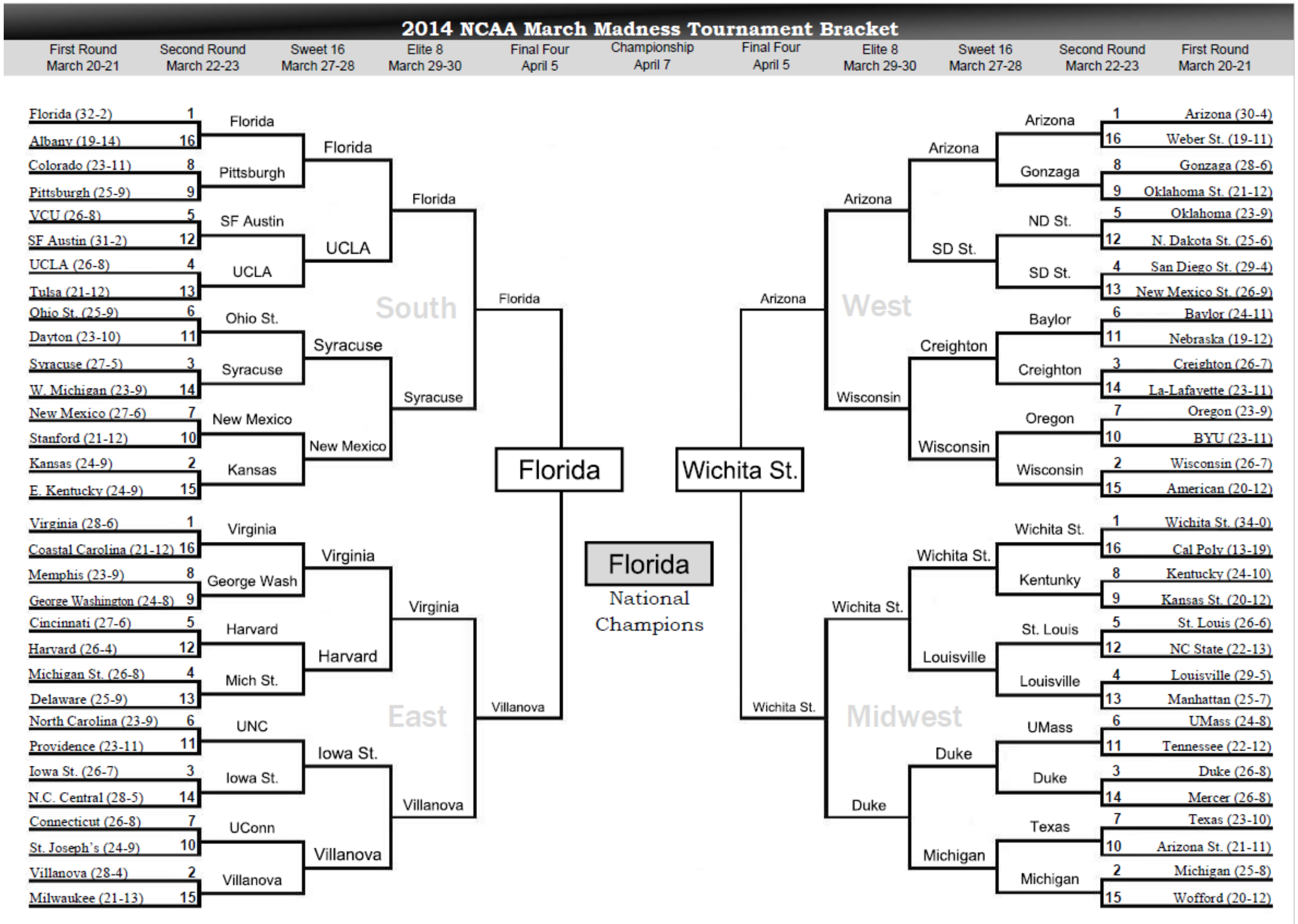

Figure 2: The 2014 tournament bracket selection via the Round-by-Round strategy ($k = 25$)

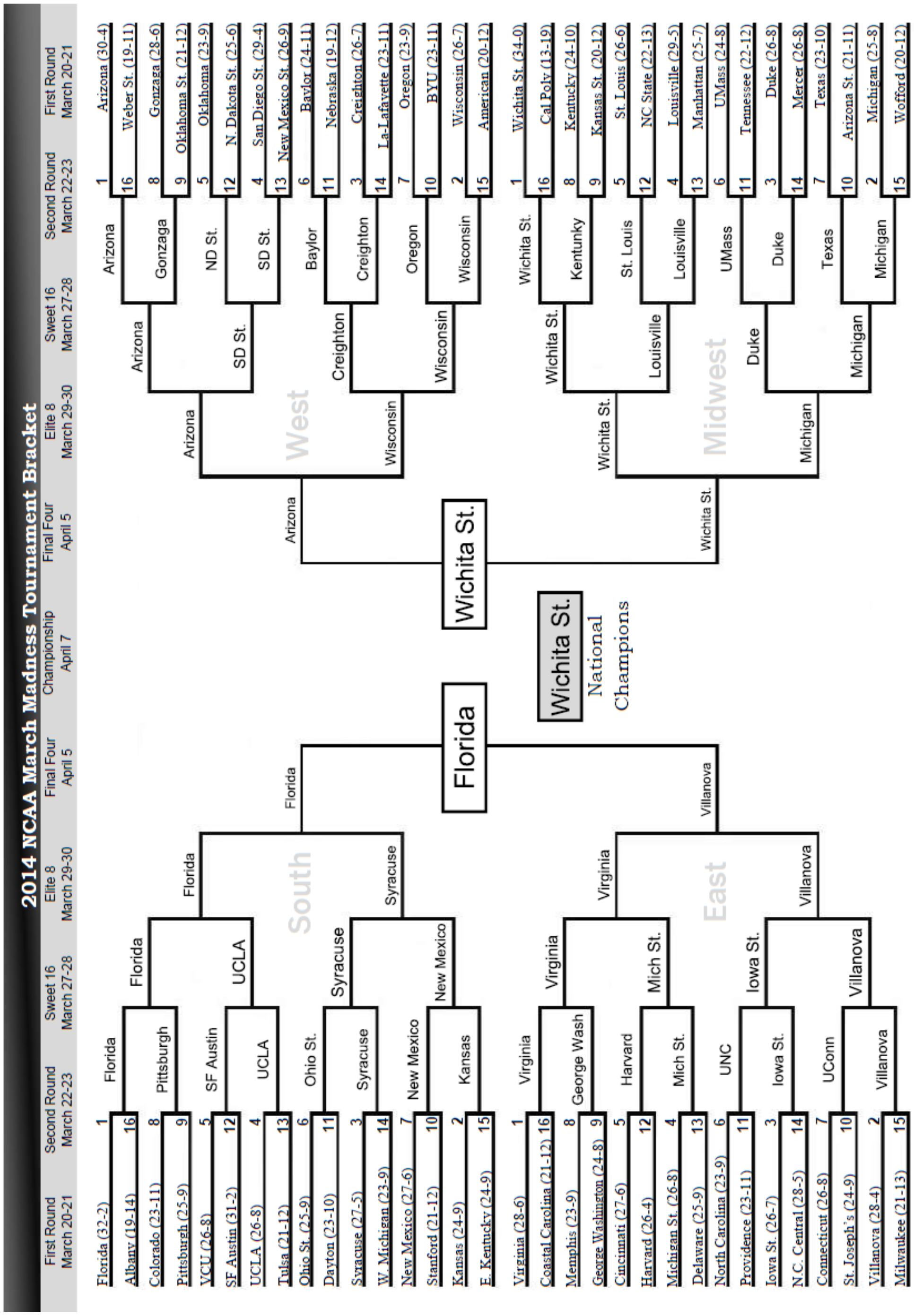

Figure 3: The 2014 tournament bracket selection via the Overall Champion strategy ($k = 25$)

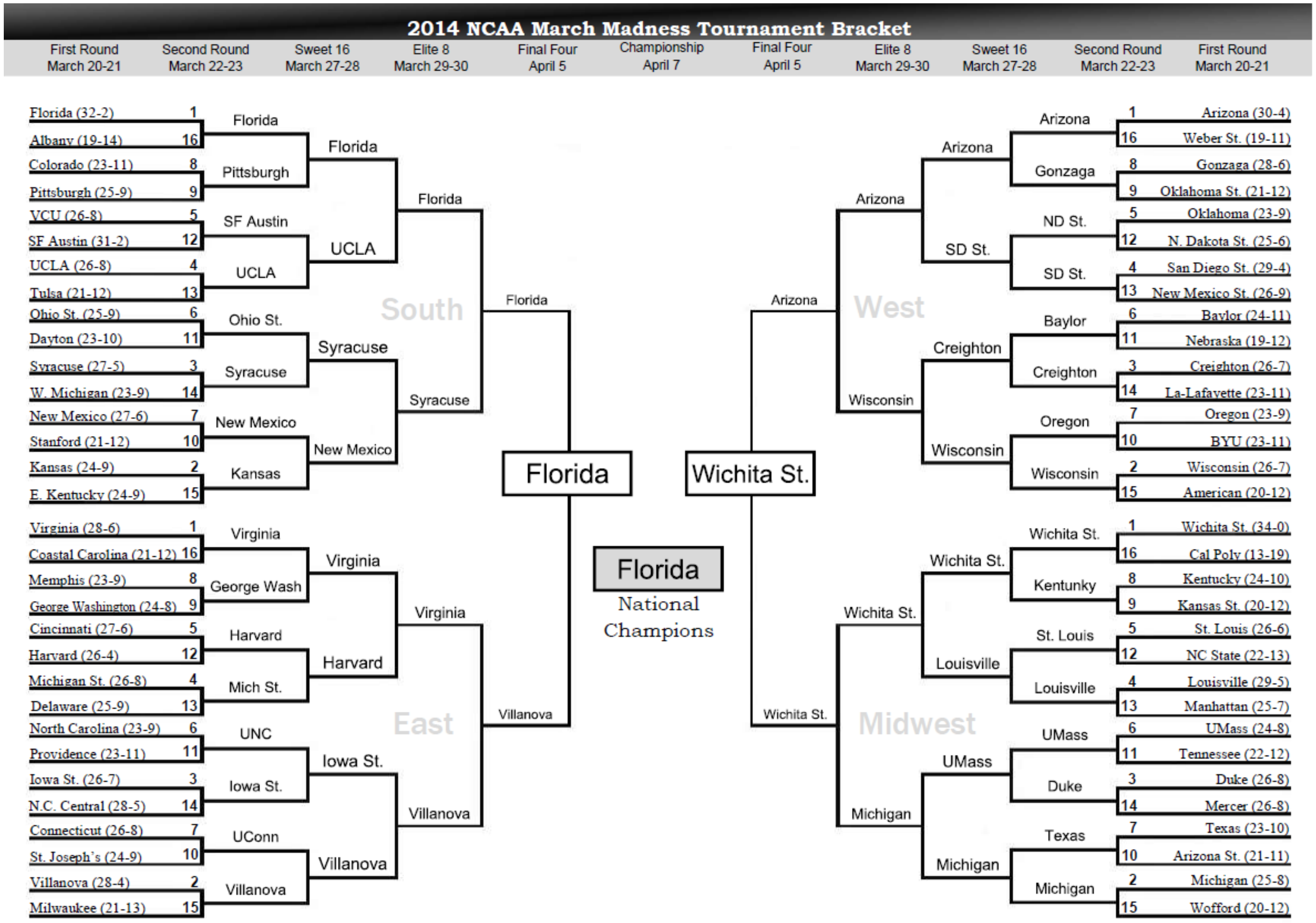

Figure 4: The 2014 tournament bracket via the Most Likely Bracket strategy ($k = 25$)

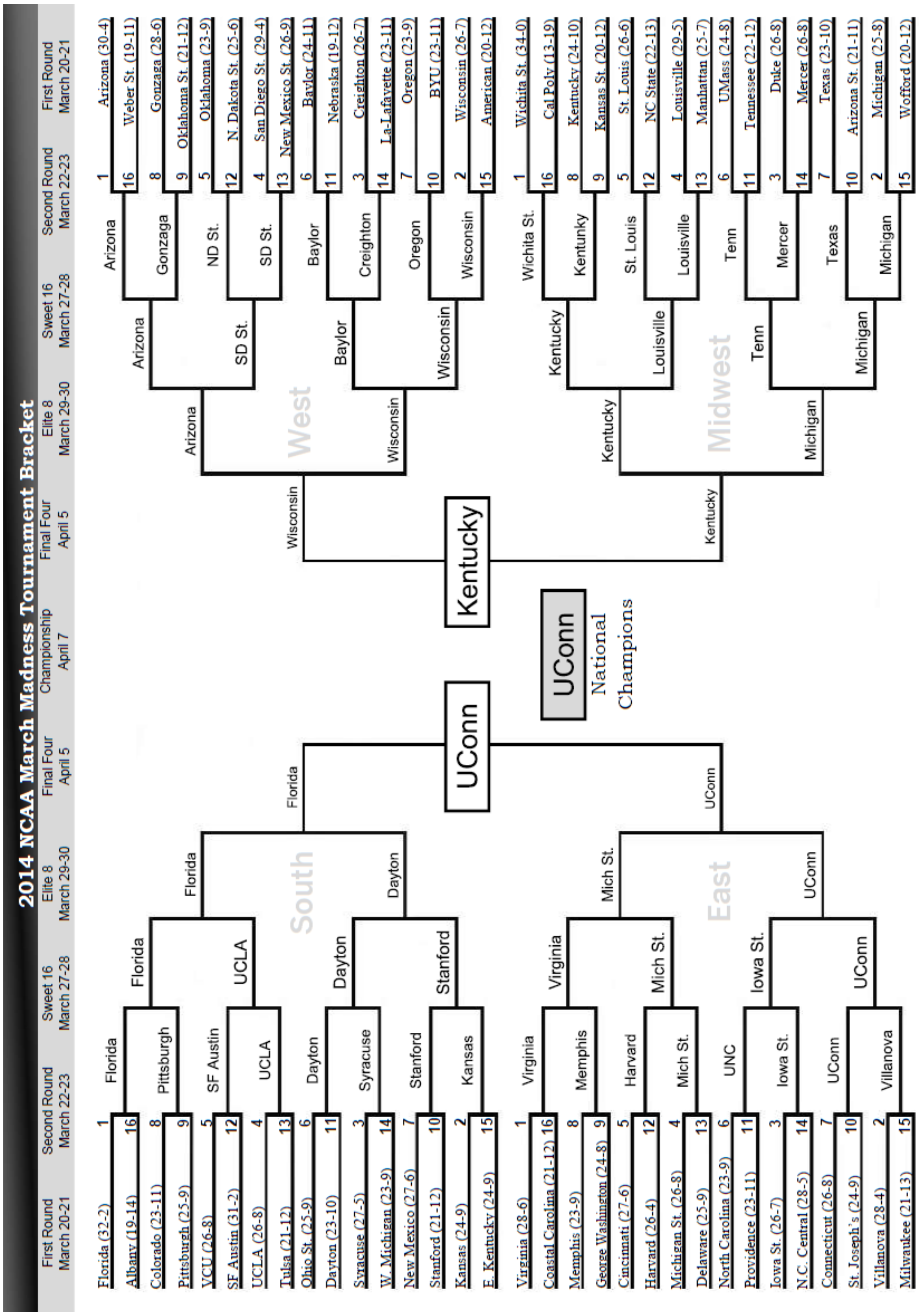

Figure 5: Actual Results for the 2014 Tournament