Method to their March Madness: Insights from Mining a Novel Large-Scale Dataset of Pool Brackets

Mason Wright
Computer Science & Engineering
CSE, University of Michigan
masondw@umich.edu

Jenna Wiens
Computer Science & Engineering
CSE, University of Michigan
wiensj@umich.edu

ABSTRACT

Each March, the NCAA Men's Basketball Tournament attracts tens of millions of viewers, a billion dollars in revenue, and millions of pool brackets from fans attempting to predict the tournament outcome. Previous studies have examined March Madness pools, but no prior work gives an in-depth exploratory analysis of a large-scale bracket dataset.

We present a novel dataset of over 200,000 brackets collected from an online pool platform, evenly split between the 2015 and 2016 tournaments. An exploratory analysis of the data reveals insights about the strategies of online pool entrants, which range from rational to absurd. The pool bracket distribution is shown to have surprising quirks that are stable from year to year, such as a large number of *all-upsets* brackets, and a tendency to over-back the top-ranked team.

This exploratory work is a first step toward a generative model that can be used to predict the empirical bracket distribution, given tournament features such as estimated pairwise win probabilities.

1. INTRODUCTION

March Madness has become a tremendously popular institution since the tournament expanded to 64 teams in 1985. In recent years, fans' interest in the tournament has led them to create over 10 million online pool brackets per year.

A pool bracket selects a winner for each of 63 games in the single-elimination tournament, such that a team picked to win any game must also have been picked to win its earlier games. The score of a bracket, under 2^{r-1} scoring (where r is the round), is 1 point per winner picked correctly in the first round, 2 points each in the second round, and so on through round $r = 6$. Under a typical pool format, each of N entrants pays 1 betting unit to enter a bracket, and all k brackets that tie for the highest score split the pot equally, earning $\frac{N}{k}$ betting units each [8]. A rational player should thus propose a bracket with a high expected score, but that is unlikely to be similar to many opponents' brackets.

The empirical distribution of pool brackets can reveal fascinating insights about pool entrants, their strategies, and their beliefs about win probabilities. Prior studies have analyzed real-world data from March Madness pools, either from a handful of office pools [8, 9] or from a large online platform [5]. These papers do not, however, provide a detailed exploratory study of a large bracket dataset.

We focus on a novel dataset, containing over 200,000 brackets from the 2015 and 2016 tournaments. The dataset is an anonymized list of pool brackets for each year, and is available upon request for research purposes. These two years differ in that 2015 had a heavy favorite (Kentucky) that was picked by 50.7% of all sampled brackets, while 2016 had more parity at the top, with the favorite (Kansas) being picked by only 25.7%. 2015 also had fewer upsets and saw higher scores in general, with a mean of 85 points per bracket, versus only 69 points in 2016.

We provide an exploratory analysis of the dataset and uncover trends that persist across years, some of which do not appear in prior literature. In particular, we note the surprising frequency of canonical brackets, such as *all-upsets* brackets and *pick-the-seeds* brackets. We observe a bimodal distribution of scores, which we show would occur even with uniformly randomly picked brackets, under 2^{r-1} scoring. We show that brackets follow a power-law-like distribution. In relation to prior work, our results support the idea that picking the champion correctly is essential to winning a large pool [8]; that top picks tend to be over-backed, with relative "bargains" common among 2- and 3-seeds [8]; and that Duke is now over-backed, rather than under-backed as before [9].

Our analysis is a first step toward a generative model for deriving entry brackets. In future work, we plan to use our dataset to derive approximately optimal entries for future NCAA tournament pools. First, we will train a model for the distribution of entry brackets, conditional on features of the tournament, including pairwise win probabilities and the ID of each team. Then, we will algorithmically search for a bracket with high expected score in a pool of a given size, assuming other entrants pick brackets according to the learned distribution.

2. RELATED WORK

Prior works have studied March Madness brackets, and a few have collected real brackets from office pools or large online platforms, but none provides an in-depth exploratory analysis of a large-scale bracket dataset.

Much of the prior work has focused on identifying which teams are most likely to be under- or over-backed based on empirical distributions. Metrick noted that the lowest-seeded (best) teams tend to be over-backed to win the championship, relative to Nash equilibrium frequencies. Metrick's conclusions are based on a set of 24 office pools with 1500 total entries in the year 1993 [8]. Niemi similarly found that the top few favorite teams are over-backed to become champion, meaning entrants could have increased their expected return on investment by picking a lower-ranked (worse) team as champion, based on 3 years of data (2003–2005) from a single office's pool that had over 100 entrants per year [9]. In contrast, McCrea noted that entrants in the ESPN Tournament Challenge tend to pick too many high-seeded (worse) teams to achieve upset wins in early rounds, based on aggregate statistics over 3 million entries in 2004–2005 [7].

One of the largest prior studies scraped 500,000 brackets from the ESPN Tournament Challenge in 2004, and a smaller number in 2005 [5]. Clair and Letscher proposed a method to find brackets with high expected return, in a pool against opponents from the observed bracket distribution. Although they used similar data to our dataset, they used the data only to train a model of opponents' bracket selections, and did not present a detailed analysis of the data. Our future work will aim to improve on the strategic bracket search process of Clair and Letscher, by relaxing their assumption that a bracket's performance relative to that of a random opposing bracket is distributed as the difference of two normal distributions.

Our analysis considers a large dataset we assembled, comprising over 200,000 brackets from two consecutive years, thus comparable in size only to that of Clair and Letscher. (The work of McCrea considered 3 million brackets, but used only aggregate statistics published by ESPN, such as how often each team was picked to reach a given round, not raw bracket data.) We validate previous observations, and we also derive novel insights that may not have been detectable with less data.

3. METHODS

Here we present the methods used for data collection and analysis. We discuss the size of dataset, how it was obtained, and the extent to which it is representative of the full set of entry brackets from the online source. We also present the probability model we use to estimate the expected score of each bracket, over possible tournament outcomes.

3.1 Data Collection

We mined data from the ESPN Tournament Challenge, a large online platform for March Madness pool entries. We collected about 102,000 completed brackets from 2015 and 112,000 from 2016. We checked each bracket to ensure team names were correct and game outcomes were legal.

The full ESPN Tournament Challenge 2016 had 13.7 million entries, of which roughly 10.6 million were completed (the rest were missing at least one pick). The 2015 pool had 12.5 million entries, of which about 9.6 million were completed. (It is remarkable that over 3 million brackets were partially filled out and then abandoned in 2016!)

Therefore, our sample contains just over 1% of completed brackets from each of 2015 and 2016. While a small fraction of the total dataset, it appears that 100,000 brackets per year is sufficient to discover many statistically significant trends.

We have only a subset of the all bracket entries because we scraped brackets with a long delay between requests, which made data collection slow, and because 200,000 brackets seemed sufficient to provide a rich model of the full distribution. We collected brackets sequentially by ID number, beginning somewhere in the middle of the ID range. We see no evidence that this collection method biased the sample distribution, relative to the full set of brackets at ESPN. Moreover, as shown in Subsection 4.1, our sample coverage of high and low scores includes almost the full range of outcomes from each year's entry set.

3.2 Probability Model

In our analysis, we aim to estimate the *expected* score of a bracket, over possible tournament outcomes. Expected scores are useful, because the actual score of a bracket is sensitive to the realized tournament outcome, and thus is a noisy signal of the bracket's strategic quality.

To find a bracket's expected score, we need a probability model for actual tournament outcomes. We use the expert forecast from FiveThirtyEight for each year. For any team i of the 64 teams in the main draw, and any round $r \in \{1, \ldots, 6\}$, let $P(i \to r)$ be the probability that team i wins at least r tournament games. Prior to the tournament, FiveThirtyEight publishes estimates of $P(i \to r)$ for all teams and rounds [3]. This forecast is created by aggregating expert predictions and factoring in home advantage, among other features.

Note that this probability model has the benefit that it does not assume statistical independence of game outcomes. It allows for the possibility that even two first-round games, for example, may have correlated outcomes. As an aside, we would need a richer probability model to estimate the likelihood that one bracket will achieve a higher score than another, because that would require a probability distribution over all game outcomes jointly. For that purpose, it is common to assume games are Markovian, with fixed pairwise win probabilities between any two teams [4, 5, 6, 9].

By linearity of expectation, the expected score $s(\cdot)$ of a bracket x equals the sum over all rounds and games, of the probability that the picked team x_g will win game g, weighted by the points for that round r.

$$\mathrm{E}(s(x)) = \sum_{r=1}^{6} \sum_{g \in r} P(x_g \to r) \times 2^{r-1}.$$

This follows the standard 2^{r-1} scoring procedure, in which later rounds are weighted more heavily.

Using the dynamic programming method of Kaplan and Garstka [6], we can derive the maximum- and minimum-expected-score brackets, for a given year's win probabilities.

4. RESULTS

In this section we present summary statistics on our dataset, as well as the findings of our exploratory analyses. We show that the sample data comprise scores covering nearly the

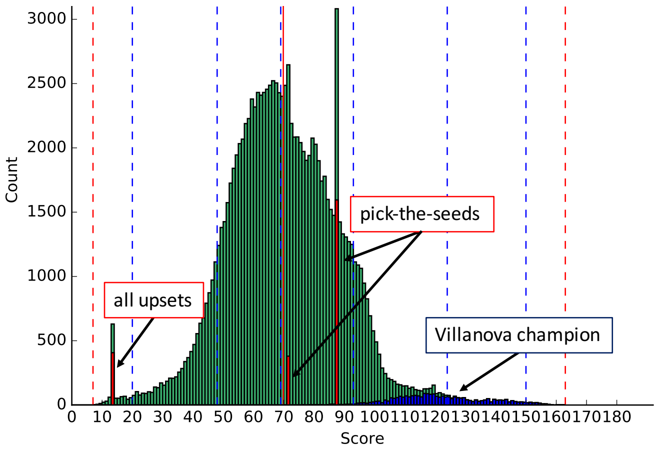

Figure 1: Histogram of actual scores for brackets sampled from 2016. Red bars show counts for all-upsets and pick-the-seeds brackets; note that some distinct brackets produce the same score. Blue bars show brackets that correctly picked Villanova to win.

full range of outcomes in the source data on ESPN, based on leader boards. We go on to analyze the distribution of observed scores, the frequency distribution of brackets, the distribution of champion picks relative to odds of winning, and the distribution of expected scores over possible tournament outcomes.

4.1 Sample Coverage

According to the ESPN Tournament Challenge leader boards, the overall top score was 173 in 2016, 183 in 2015 [1, 2]. Our sample had a top score of 163 in 2016, which tied for 99th overall (99.999th percentile); the 2015 sample had a top score of 178, which tied for 25th overall (99.9997th percentile). These high sampled scores suggest that our sample size is adequate to cover nearly the full range of outcomes in the overall pool.

4.2 Empirical Score Distribution

Figure 1 shows actual scores for our 2016 bracket sample. The sharp peaks for scores of 13, 71, and 87 correspond to all-upsets brackets, which all score 13 points for 13 first-round upsets, and pick-the-seeds brackets, which could score 71 or 87 points in 2016. In 2016, 407 sample entries (0.4%) were all-upsets, even though these brackets have no hope of winning any pool. This serves as a reminder that not all pool entrants take the contest seriously.

By all-upsets, we mean any bracket where the higher-seeded (worse) team wins each game. By pick-the-seeds, we mean any bracket where the lower-seeded (better) team wins each game. There are 8 distinct all-upsets brackets and 8 pick-the-seeds brackets per tournament, differing only in which teams win in the final two rounds. This is because all Final Four and championship teams will be 16-seeds in an all-upsets bracket, or 1-seeds in a pick-the-seeds bracket.

Prior work suggests that one must pick the champion correctly to have a good chance of winning a large pool [8], and this idea is supported by Figure 1. The blue bars show the score of every bracket that picked Villanova correctly as the champion. No sampled bracket scored above 134 without picking Villanova to win; the highest observed score was 163, almost 20 points higher.

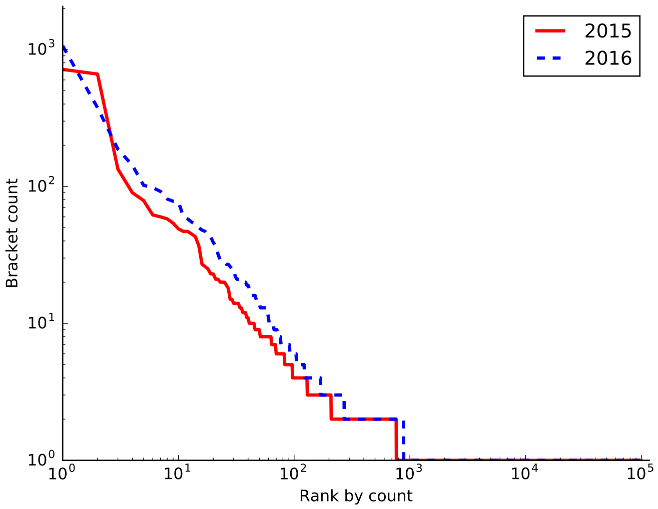

Figure 2: Frequency of each bracket from 2015 and 2016, sorted along the x-axis. Each axis uses a log scale.

4.3 Occurrence Frequency Distribution

Figure 2 shows the frequency with which each distinct bracket occurred in 2015 and 2016. The linear trend of the graph indicates a power-law distribution of occurrence frequencies, where lower-ranked brackets are geometrically less common. The most common bracket in each year was a pick-the-seeds bracket, which makes up 0.7% of the sample in 2015, 0.9% in 2016. Closer inspection shows that 8 of the 10 most popular brackets from 2016 were either pick-the-seeds or all-upsets. A similar trend appeared in 2015.

The bracket distribution has high entropy in each year we sampled, with 98,446 (95.6%) unique brackets in 2015, and 106,365 (94.9%) in 2016. If we had collected more sample data, we would likely have observed more copies of some of these brackets, but additional unique brackets would surely have appeared as well. The empirical entropy was 16.4 bits in 2015 (16.5 bits in 2016), which is nearly the maximum possible for a dataset of this size—the maximum possible would be 16.7 bits in 2015 (16.8 bits in 2016), or $\log_2$ of the data size.

4.4 Bimodal Score Distribution

Figure 1 shows a long tail of high scores for the brackets that correctly picked Villanova to win. As shown in Figure 3, even a uniform random bracket distribution tends to produce a bimodal score distribution. We show scores for 1 million brackets, where each bracket is generated by uniformly randomly picking the winner for each game. The extreme reward for picking the champion under 2^{r-1} scoring leads to a minor peak near 90, caused by the $\frac{1}{64}$ of brackets that happen to pick the champion. For a similar reason, the high reward of the champion tends to produce many

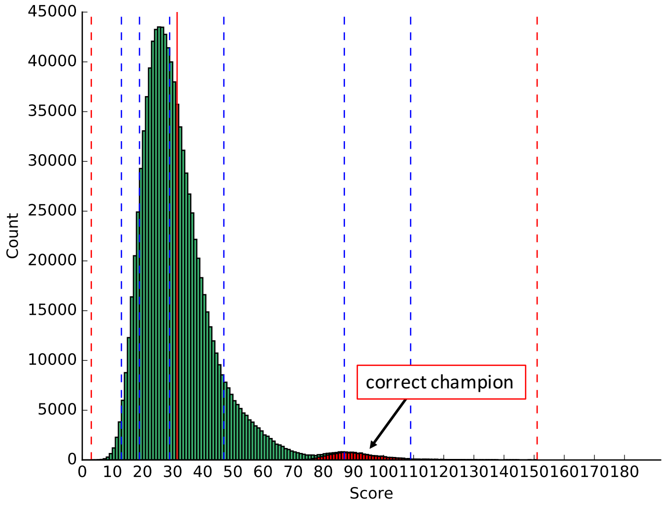

Figure 3: Scores of 1 million uniformly randomly generated brackets, under 2^{r-1} scoring. Red bars show brackets that correctly predict the winner.

high-scoring bracket "outliers" in a graph like Figure 1.

4.5 Champion Pick Distribution

In Figure 4, we plot the fraction of sample brackets that picked each team as champion, along with the difference from the estimated odds each team would win. In accordance with prior work, we see that the top-ranked team (Kansas) is over-backed: 25.7% of entrants choose it to win, while expert predictions give it only a 19.1% chance. Notice that several 2- and 3-seeded teams are under-backed, as predicted by prior work. Also note that the weakest teams in the tournament, including but not limited to 16-seeds, are dramatically over-backed. Our samples had 0.4% and 0.5% of brackets pick a 16-seed to win the tournament in 2015 and 2016, which is far above the estimated likelihood of a 16-seed becoming champion.

As a note of caution, observe that in a small pool, it is rational for entrants to pick strong teams to win with probability greater than their actual odds of winning [8]. Only in very large pools should each team be picked according to its actual odds of winning. Some brackets in the Tournament Challenge may have been intended as entries in small pools.

Some authors have claimed that Duke tends to be under-backed due to a "Duke-hating factor" (which may stem from jealousy over the program's all-around excellence and class). The sample from 2015 and 2016, however, shows Duke slightly over-backed in each year, with 2.3% picking it to win in 2016 at 1.7% odds, and 9.8% picking it in 2015 at 5.8% odds.

4.6 Expected Score Distribution

We can compute the expected score of a bracket using expert estimates of $P(i \rightarrow r)$, as shown in Section 3.2. In the left panel of Figure 5, we plot each 2016 bracket's actual score against its expected score. Because 2016 was a year with high parity among teams, the maximum possible expected score was only 91.0. Note that most brackets cluster in a zone of high expected scores, with an actual score just below the expected, likely because the most-picked teams (Kansas and Michigan State) did not reach the Final Four.

In the right panel of Figure 5, we show only those brackets from 2016 that correctly picked Villanova to win the championship (2.7% of all brackets). These brackets are all far above the $y = x$ line, where actual score equals expected score, because Villanova had been given only a 6% chance of winning. All of the highest scorers are in the right panel, which shows how vital it is to pick the champion correctly.

5. CONCLUSION

We mined a novel dataset of over 200,000 March Madness brackets from 2015 and 2016. A surprising fraction of pool entrants pick canonical brackets, e.g., pick-the-seeds and all-upsets brackets. More than 94% of entries in our sample, however, were unique. Our results confirm that picking the champion correctly is essential to winning a large pool.

This analysis could be used to inform a generative model for deriving entry brackets, conditional on features of the tournament, including estimated pairwise win probabilities and the ID of each team. Such a model could then be applied to future tournament pools with the goal of maximizing one's bracket score.

6. REFERENCES

[1] ESPN Tournament Challenge 2015 leaderboard. `http://games.espn.go.com/tournament-challenge-bracket/2015/en/leaderboard`. Accessed: 2016-05-01.

[2] ESPN Tournament Challenge 2016 leaderboard. `http://games.espn.go.com/tournament-challenge-bracket/2016/en/leaderboard`. Accessed: 2016-05-01.

[3] FiveThirtyEight 2015 March Madness predictions. `https://github.com/fivethirtyeight/data/tree/master/march-madness-predictions-2015/mens`. Accessed: 2016-05-01.

[4] D. J. Breiter and B. P. Carlin. How to play office pools if you must. *Chance*, 10(1):5–11, 1997.

[5] B. Clair and D. Letscher. Optimal strategies for sports betting pools. *Operations Research*, 55(6):1163–1177, 2007.

[6] E. H. Kaplan and S. J. Garstka. March Madness and the office pool. *Management Science*, 47(3):369–382, 2001.

[7] S. M. McCrea and E. R. Hirt. Match Madness: Probability matching in prediction of the NCAA basketball tournament. *Journal of Applied Social Psychology*, 39(12):2809–2839, 2009.

[8] A. Metrick. March Madness? Strategic behavior in NCAA basketball tournament betting pools. *Journal of Economic Behavior & Organization*, 30(2):159–172, 1996.

[9] J. B. Niemi. *Identifying and Evaluating Contrarian Strategies for NCAA Tournament Pools*. PhD thesis, University of Minnesota Minneapolis, 2005.

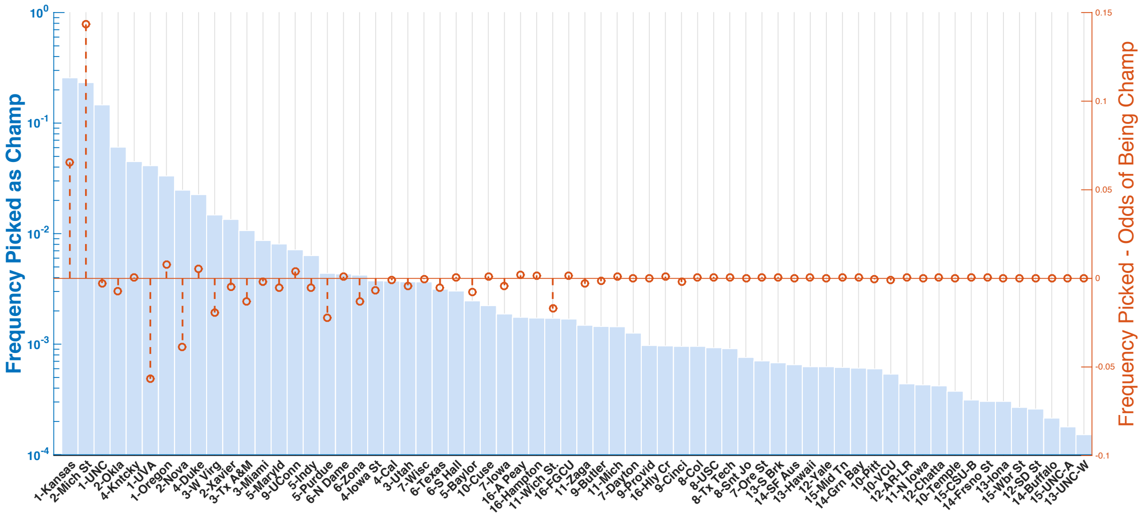

Figure 4: The 64 teams in the main draw of the 2016 tournament, ordered by number of sampled brackets that picked the team as champion (left axis, log scale). Circles show, for each team, the difference between the fraction of brackets that picked that team as champion and its odds of becoming champion (right axis).

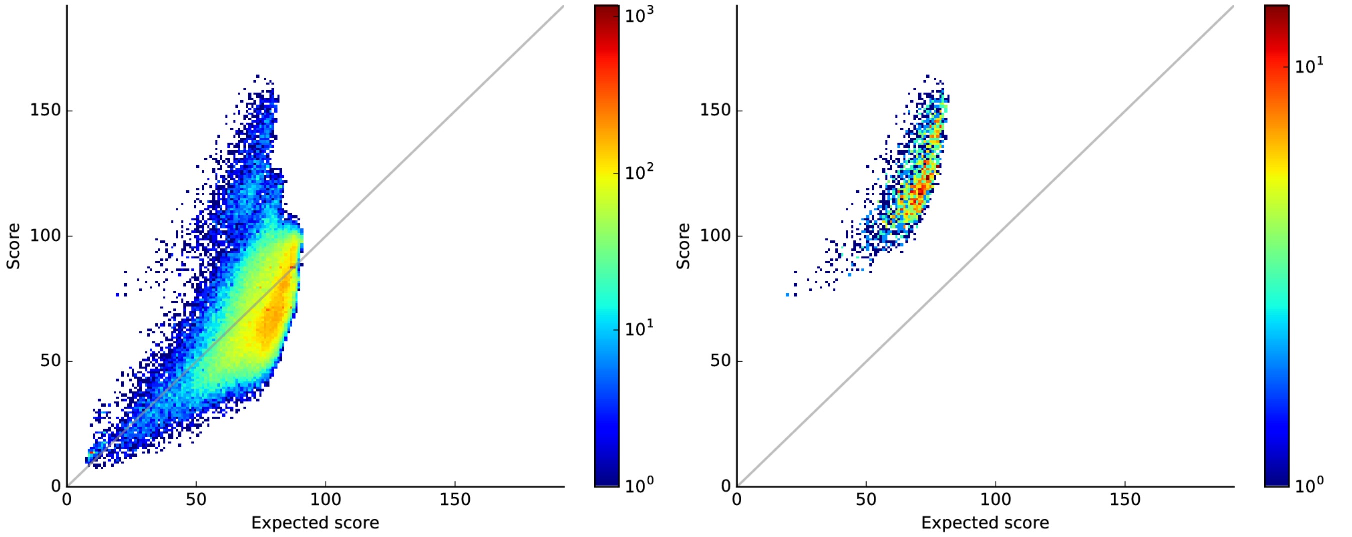

Figure 5: Histograms of expected score versus actual score, for 2016 brackets. At left, all sampled brackets are shown; at right, only those that correctly picked Villanova to win. Gray lines show where expected score equals actual score. (Note that colors at right have a different scale.)