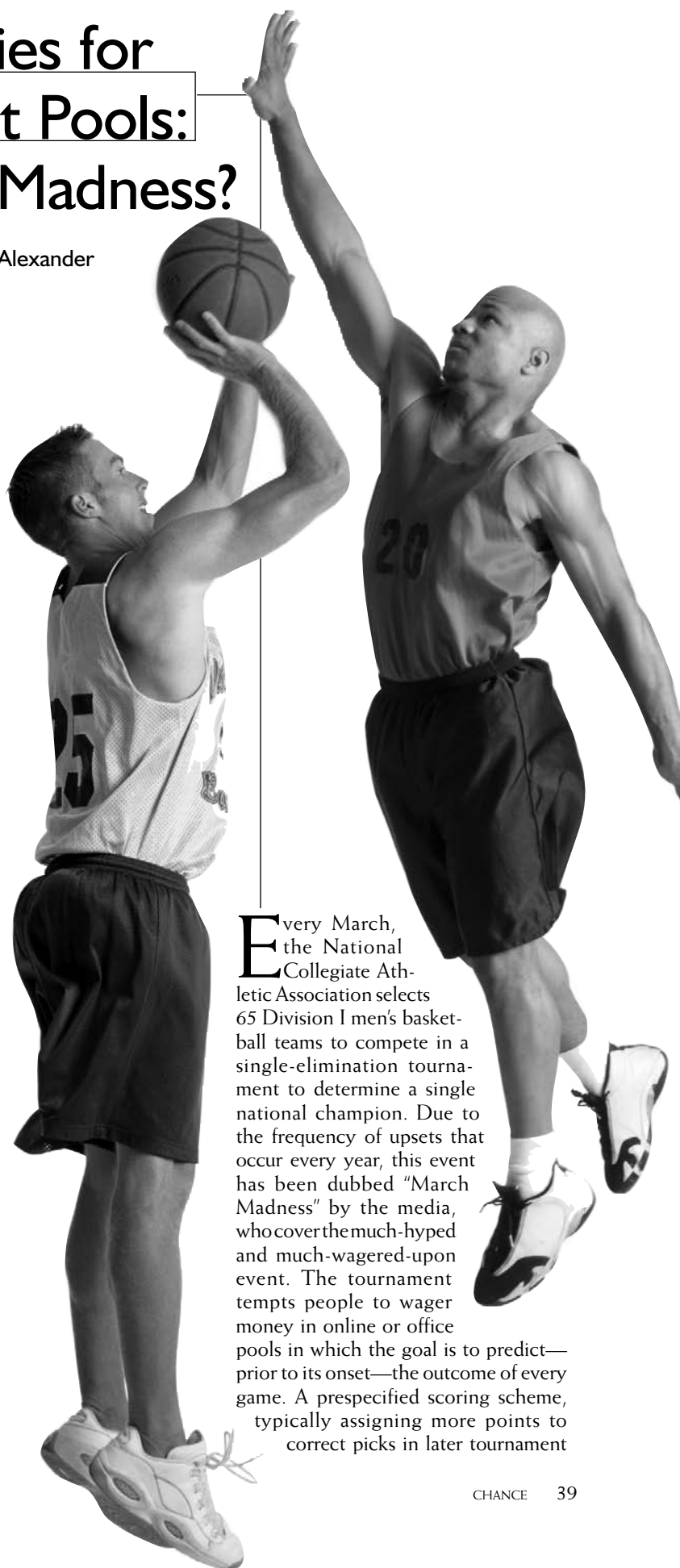


Contrarian Strategies for NCAA Tournament Pools: A Cure for March Madness?

Jarad B. Niemi, Bradley P. Carlin, and Jonathan M. Alexander



Every March, the National Collegiate Athletic Association selects 65 Division I men's basketball teams to compete in a single-elimination tournament to determine a single national champion. Due to the frequency of upsets that occur every year, this event has been dubbed "March Madness" by the media, who cover the much-hyped and much-wagered-upon event. The tournament tempts people to wager money in online or office pools in which the goal is to predict—prior to its onset—the outcome of every game. A prespecified scoring scheme, typically assigning more points to correct picks in later tournament

Table 1—Assumed Win Probabilities, Idealized Four-Team Tournament

	A	B	C	D
A	-	.57	.70	.78
B	.43	-	.64	.73
C	.30	.36	-	.60
D	.22	.27	.40	-

The (x; y) entry is the probability that team x beats team y.

Table 2—Assumed Opponents' Choices, Idealized Four-Team Tournament

Team	Round	
	1	2
A	.90	.75
B	.75	.20
C	.25	.05
D	.10	.00

The (x; y) entry is the proportion of opponent sheets that have team x winning in round y.

rounds, is used to score each entry sheet. The players with the highest-scoring sheets win predetermined shares of the total money wagered. In most states, such pools are considered legal, provided the "poolmaster" does not accept payment of any kind—including his entry fee.

The question at hand is: How to make picks in order to maximize profit? Many strategies exist for filling in a pool sheet, based on everything from the teams' AP rankings to the colors of their uniforms. One sensible mathematical approach might be to try to maximize your sheet's expected score and assume this will, in turn, maximize profit. This approach can, indeed, be profitable when the scoring scheme is complex, particularly when it awards a large proportion of the total points for correctly predicting upsets. Tom Adams' web site, *www.poolologic.com*, provides a Java-based implementation of this approach for a variety of pool scoring systems. This web site also can produce the highest expected score sheet subject to the constraint that the champion is a particular team. In a tournament, a pool sheet consists

of picks for the winner of every game, where these winners can only come from the winners of the previous round. We define the probability of a sheet as the product of the win probabilities for each game chosen on that sheet. Since the probability of an individual game outcome turns out to be well-approximated using a normal approximation, which we describe below, this calculation is tedious but straightforward.

Most office pool scoring schemes are relatively simple, awarding a set number of points for each correct pick in a given tournament round. In such cases, the sheet that maximizes expected score will typically predict few upsets, and thus have too much in common with other bettors' sheets to be profitable. What's more, pool participants tend to "overbet" heavily favored teams—a fact it seems a bettor could use to competitive advantage.

Betting strategies that attempt to choose reasonable entries while simultaneously seeking to avoid the most popular team choices are sometimes referred to as contrarian strategies. To quantify the amount a given sheet has in common with the other sheets entered in a pool, we define the similarity of a sheet as the sum of the similarity of each game chosen by that sheet to the picks for that game on every other sheet. To compute the similarity of a single game, we multiply the points available for that game by the proportion of people in the pool who chose that team to win that game. After summing the individual similarities, we normalize the statistic by dividing by the maximum possible points available. This provides a statistic that ranges from 0 to 1; it is 0 when the sheet has no picks in common with any other sheet and 1

when the sheet has exactly the same picks as every other sheet in the pool.

To illustrate the concepts of probability and similarity, consider a four-team tournament where team A plays team D and team B plays team C in the first round, and the winners of these games play each other in a second-round game for the championship. We adopt a simple scoring scheme where we award one point for each correct first-round pick and two points for a correct second-round (championship) pick. In order to calculate the probability of a particular set of picks, we need the probability that each team will beat each other team in the pool. Suppose we adopt the probabilities shown in Table 1 (we illustrate how these can be calculated from published betting lines or team computer ratings below). In order to calculate similarity, we instead need to know how all the other players in our pool have made their picks. The necessary summary of these picks is shown in Table 2, where the entries are the proportions of opponent sheets that have the team in that row winning their game in the round indicated by the column.

So, for example, we see that 0.20 (20%) of our opponents have team B winning the championship. Now consider two sheets entered into this pool. The first sheet picks A and B in the first round, followed by A winning the championship. The second sheet picks A and C in the first round, followed by C winning the championship. The first sheet appears to be a 'safe' strategy, as A and B are better than their first-round opponents and A is better than B. The second sheet appears to be more of a long-shot, as it predicts C will defeat two superior teams. Using our definitions above,

Table 3—Exploratory Data Analysis of Chicago Office Pool Sheets

Year Participants	2003 113	2004 138	2005 167
Champions Bet by Seed			
1	86 (76%)	61 (44%)	137 (82%)
2	14 (12%)	58 (42%)	18 (11%)
3	4 (4%)	10 (7%)	5 (3%)
4	7 (6%)	3 (2%)	4 (2%)

Seed is the rank of the team in one of the four regions of the country.

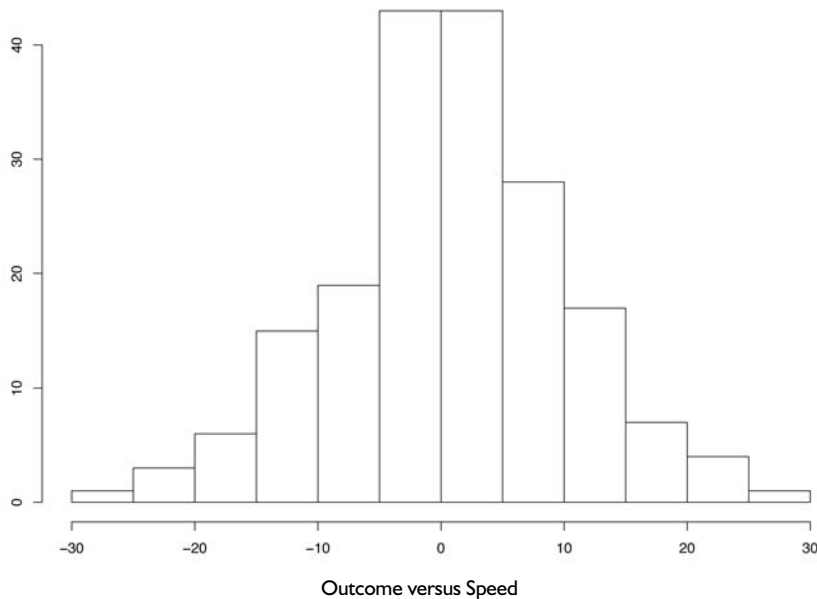
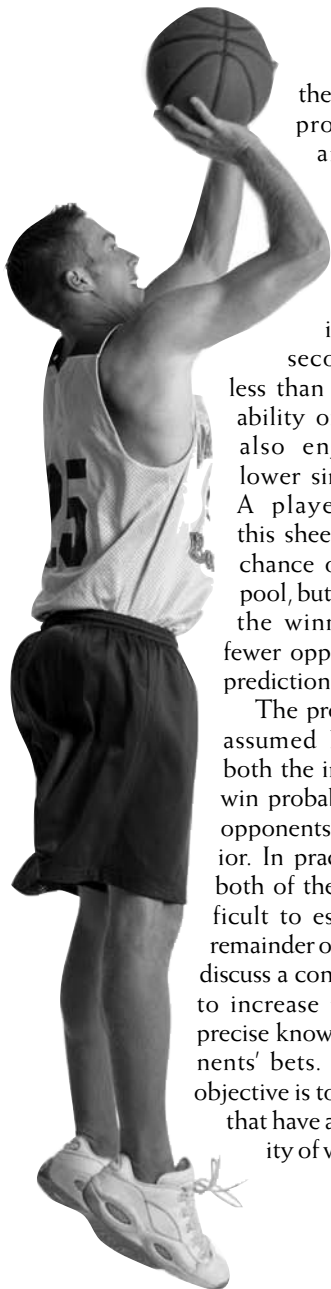


Figure 1. Histogram of game outcomes (favorite's score minus underdog's score) minus the imputed point spreads (based on final Sagarin ratings), 2003–2005 NCAA tournament data



the first sheet has probability 0.28 and similarity 0.79, while the second has probability 0.13 and similarity 0.31. The second entry has less than half the probability of the first, but also enjoys a much lower similarity score. A player submitting this sheet has a smaller chance of winning the pool, but figures to share the winnings with far fewer opponents if these predictions do pan out.

The previous example assumed knowledge of both the individual game win probabilities and our opponents' betting behavior. In practice, either or both of these may be difficult to estimate. In the remainder of this article, we discuss a contrarian method to increase profit without precise knowledge of opponents' bets. This method's objective is to identify teams that have a high probability of winning, but are

likely to be "underbet," relative to other teams in the pool.

Available Data

Our main source of data is three years' worth of betting sheets and actual tournament results for an ongoing Chicago-based office pool, summarized in Table 3. "Champions bet by seed" indicates how many people chose each of the top four seeds to win the tournament, with the corresponding percentages in parentheses. For example, in 2003, 86 out of 113 sheets (76%) had either Kentucky, Arizona, Oklahoma, or Texas (the four #1 seeds that year) winning the championship. The pattern appears consistent, except for 2004, when 44% of the sheets had a #1 seed winning and 42% of the sheets had a #2 seed winning.

This was due to 22% of the sheets having Connecticut (a #2 seed) as their champion. In that year, Connecticut was widely regarded as the best team in its region and did end up winning the tournament. The scoring scheme for this office pool awards 1, 2, 4, 8, 16, and 32 points for correctly predicted victors in rounds 1–6, respectively. The score for each sheet is then the sum of the points earned for each correctly predicted game. The percentage of the total pot awarded is 45%, 22.5%, 15%, 10%, and 7.5% for the first-place through fifth-place finishers, respectively.

Simulating Return on Investment

A number of numerical methods can be used to analyze tournament data. Perhaps the simplest approach would be to enumerate all possible tournaments, determine probabilities for each, and then obtain expected winnings for each sheet as it competes with opponent sheets. Unfortunately, because there are 63 games (excluding the pretournament "play-in" game between the 64th and 65th teams selected), there are 263 possible tournament outcomes. For this reason, one of the most useful tools in analyzing tournaments is to simulate a large number of tournaments, using the resulting relative frequencies of the outcomes to reduce the computation but preserve realism. After simulating these tournaments, we can calculate the return on investment of each sheet.

If we think of the outcome of a basketball game as the number of points scored by the favorite minus the number scored by the underdog, it turns out a histogram of these outcomes looks like a normal distribution (bell curve) centered at the point spread, the amount by which the betting public expects the favorite to win. Figure 1 provides this histogram for our data—the $63 \times 3 = 189$ games in the 2003, 2004, and 2005 tournaments. After subtracting the point spread from each game outcome (favorite's score minus underdog's score), we do get an approximately normal distribution centered on 0. Thus, in the long run, the bettors seem to "get it right": half the time, the favorite wins by more than expected ("covers the spread") and half the time not. About 95% of games land within roughly 24 points (two standard deviations) of the spread (i.e., between -24 and 24 in Figure 1). This means there is a simple formula (using the area under the correct bell curve) for converting a point spread to the probability that the favorite will win the game.

Sadly, true point spreads will not be available for every possible match-up prior to the tournament. However, many computer ratings are designed so the difference between two teams' ratings is an estimate of the point spread for a game between these two teams at a neutral site. This is true of all the ratings used below, though they differ in their emphases. Vegas ratings are based on only point spreads and over/under betting lines in the first round of the

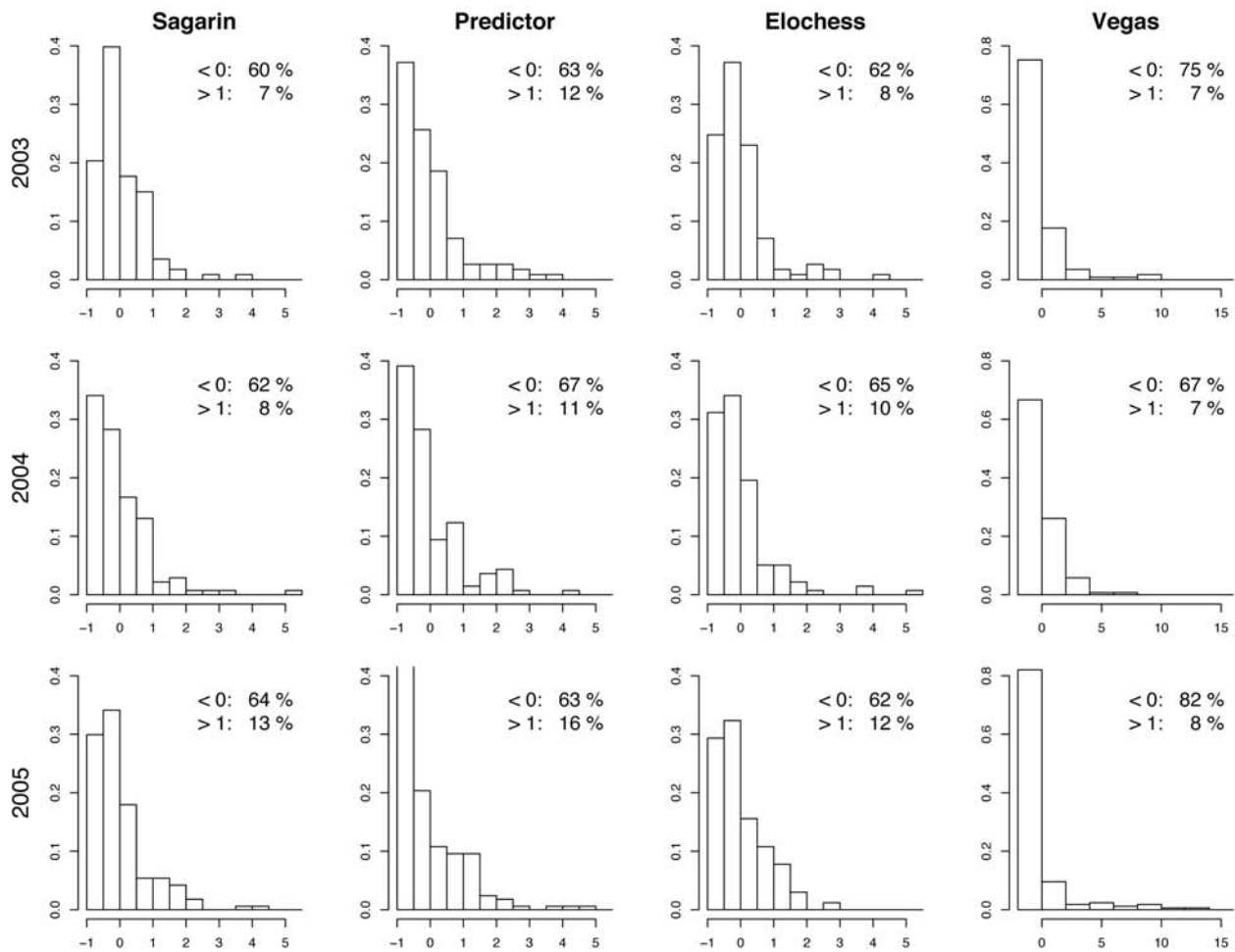


Figure 2. Histograms of simulated ROI for all pool sheets across years and rating systems

tournament. Elochess ratings are based on the win-loss results of all games in the regular season. Predictor ratings use the point differentials in all games in the regular season. Finally, Sagarin ratings are a combination of Elochess and Predictor.

To simulate a tournament, we start by picking one of the rating systems. Looking at a particular match-up between two teams, we calculate the favorite's rating minus the underdog's rating. Using a bell curve centered at this difference, we calculate the area under the curve and to the right of zero. This gives us the probability that the favorite will win that game. We then draw a uniform random number between zero and one, for example using `randn` in Microsoft Excel. If this number is less than the probability, our simulation says the favorite won that game, otherwise the underdog won it. We can

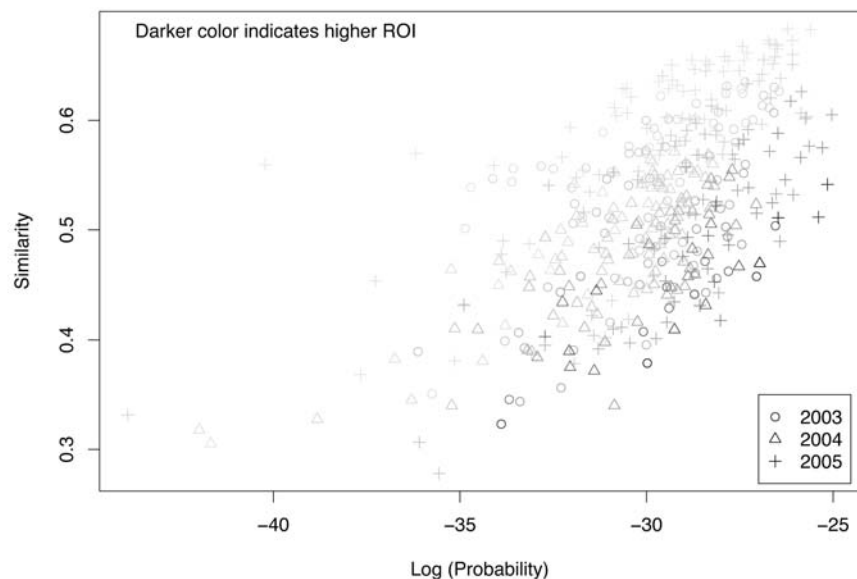


Figure 3. Scatterplot of similarity versus log(probability) with ROI indicated by shading, Predictor ratings

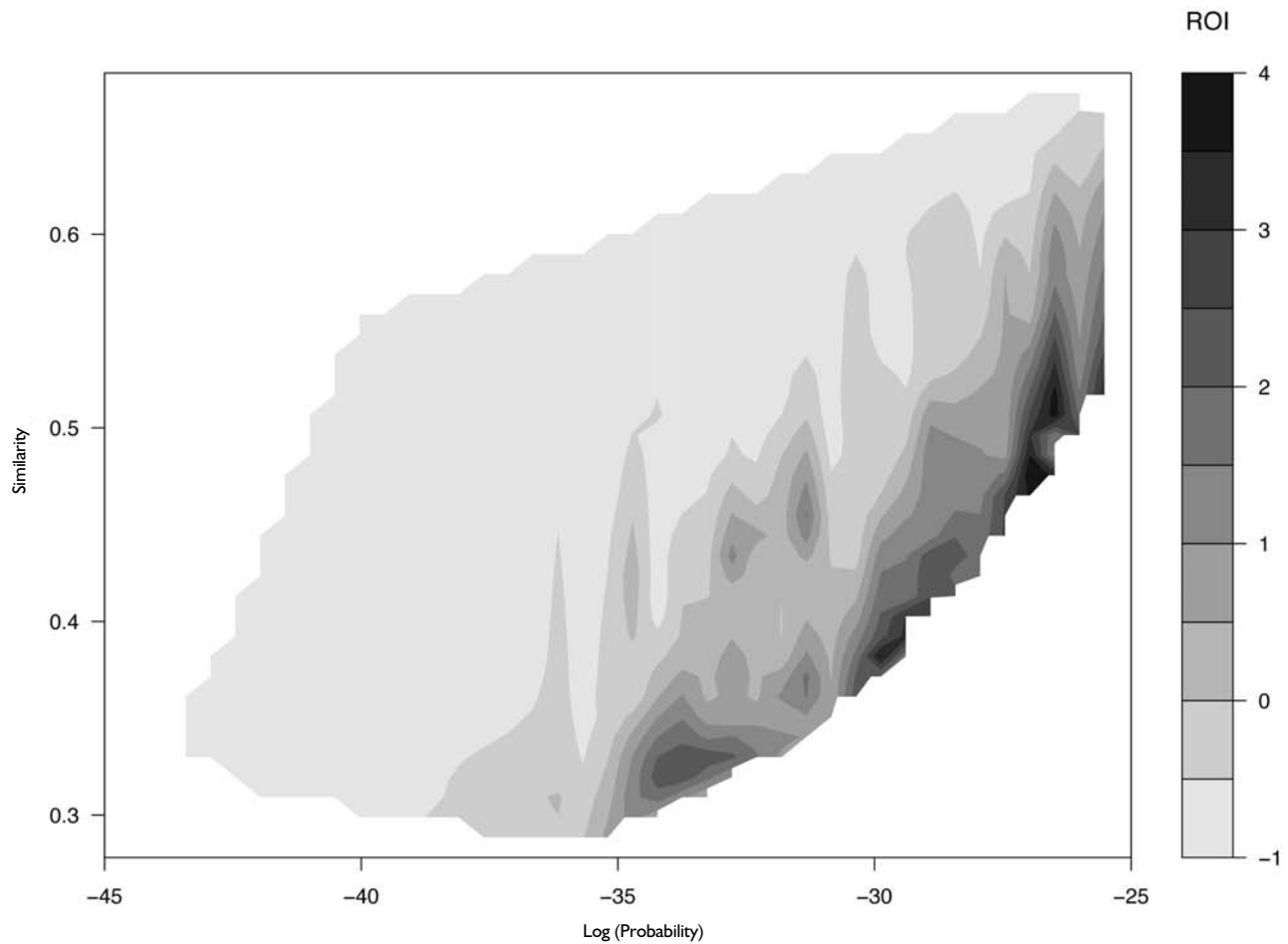


Figure 4. Filled contour plot of ROI by similarity and log(probability), Predictor ratings

repeat this process for all of the Round 1 games, followed by the resulting Round 2 match-ups and so on to simulate an entire 63-game tournament. Now, for each simulated tournament, all office pool sheets for that year can be scored, ranked, and awarded prizes as described previously. Repeating this process over many simulated tournaments, the return on investment (ROI) for each sheet may be calculated as the following:

$$\text{ROI} = \frac{\text{total won} - \text{total invested}}{\text{total invested}}$$

This calculation is standardized, so each sheet costs \$1. A ROI of zero indicates a break-even strategy; whereas, a negative (positive) value indicates a losing (winning) strategy. We will calculate ROI for each actual sheet for each year and probability model, as well as certain 'optimal' sheets.

Table 4 — Simulated ROI for Various Sheet Selection Methods and True Probability Models

	2003				2004				2005			
Method	S	P	E	V	S	P	E	V	S	P	E	V
Maximum Expected Score	-0.2	2.5	0.8	14	3.2	3.1	3.4	2	0.3	3.2	2.3	14
Contrarian	3.7	2.5	3.5	14	4.9	3.1	5.4	7	3.6	3.2	2.7	16

Contrarian Motivation

Before developing a contrarian strategy, an important question is whether the idea has demonstrable merit. In this section, we show that most sheets in an office pool have low ROI and that maximizing point total methods also do not have

high ROI. We then turn to the problem of producing contrarian sheets with improved ROI.

We simulated 1,000 tournament outcomes for each of the four rating systems and each year of our data. We then evaluated how the 418 sheets from

our 2003–2005 Chicago office pools would have fared in these simulated tournaments. Thus, we are evaluating sheets not based on what actually happened, but on what was likely to have happened according to our rating systems. Histograms of these results can be seen in Figure 2. The rows correspond to years and the columns to different rating systems. The histograms then provide the proportion of sheets falling into each ROI category. For example, in 2003 using Predictor as the rating, about 37% of the players had a simulated ROI between -1 and -0.5 ; one player had a simulated ROI between 3.5 and 4 . Also shown on each histogram is the percentage of sheets having an ROI below zero, indicating a losing investment, and the percentage above one, a substantial (at least money-doubling) winning investment. Note that in all 12 cases, at least 60% of the strategies are losers in the long run, while the proportion that double one's money or better rarely exceeds 15%.

Figure 3 investigates the differences between those pool sheets consistently near the top and bottom of the simulated ROI distributions in Figure 2 by plotting similarity versus $\log(\text{probability})$ using the Predictor rating for the sheets in our data set. The plotting character indicates the sheet's year, while its shading indicates its simulated ROI (with darker shading corresponding to higher ROI). The figure suggests the first requirement for a high ROI is to have a relatively high probability, as there are few dark points with $\log(\text{probability})$ less than -30 . However, low similarity also appears to be a general characteristic of high ROI sheets. This relationship is further clarified by the filled contour plot in Figure 4, which indicates that given a sheet's $\log(\text{probability})$, low similarity tends to maximize ROI, and given a sheet's similarity, high probability tends to maximize ROI.

As mentioned above, previous thinking in this area has focused on identifying sheets that maximize score. To further illustrate that this method may not deliver a sheet with high expected ROI, we derived the maximizing sheet for each year. We then repeated our ROI simulation under all four rating systems. The expected score-maximizing sheet was entered into the simulated pools and its ROI performance

Table 5 — Actual (A) versus Expected (S, P, E, V) Champion Picks, 2003–2004

2003						
Team	A	S	P	E	V	U*
Kentucky	58	18	15	19	15	
Arizona	15	13	13	12	12	
Kansas	9	10	19	6	0	P
Oklahoma	8	6	5	6	4	
Illinois	5	2	3	2	0	
Texas	5	9	7	9	40	V
Syracuse	4	8	4	15	1	E
Florida	3	5	4	4	2	
Pittsburgh	2	11	11	9	2	S
Dayton	1	0	0	1	0	
Indiana	1	0	0	0	0	
Maryland	1	2	4	1	0	
Louisville	1	4	6	2	4	
Other	0	26	23	27	32	
2004						
Team	A	S	P	E	V	U
UConn	30	24	12	23	5	
Kentucky	27	6	8	6	29	
OK State	23	14	8	15	4	
Duke	19	24	31	18	24	P
Stanford	12	5	5	6	13	
Gonzaga	5	5	9	3	17	
Pitt	4	4	3	9	0	
Georgia Tech	3	9	10	9	2	
St. Joseph's	3	15	12	18	17	SEV
Texas	3	3	3	3	0	
Wisconsin	2	2	3	1	0	
Syracuse	2	1	0	2	0	
Michigan St.	1	0	0	0	0	
Wake Forest	1	2	4	2	7	
Cincinnati	1	1	5	1	2	
North Carolina	1	2	4	1	2	
Maryland	1	1	2	1	3	
Other	0	19	19	18	13	

*The most underappreciated team is indicated in column U.

Table 5 (continued) — Actual (A) versus Expected (S, P, E, V) Champion Picks, 2005

2005						
Team	A	S	P	E	V	U
Illinois	83	31	15	67	18	
North Carolina	38	32	56	20	51	P
Duke	13	18	21	13	3	
Oklahoma St.	12	10	14	4	0	
Washington	3	11	7	9	37	EV
Wake Forest	3	14	10	7	28	S
Kentucky	3	7	4	8	1	
Gonzaga	2	1	0	2	0	
Florida	2	2	4	1	0	
Michigan St.	2	3	3	3	0	
Boston College	1	1	0	2	0	
Arizona	1	3	1	4	0	
Georgia Tech.	1	1	2	0	0	
Louisville	1	6	10	6	1	
Kansas	1	6	4	2	0	
Oklahoma	1	5	5	2	22	
Other	0	15	11	17	5	

*The most underappreciated team is indicated in column U.

evaluated in a competition with that year's actual sheets. The average ROI of the maximum expected score sheets are displayed in the first row of Table 4, where the column headings indicate the Sagarin (S), Predictor (P), Elochess (E), and Vegas (V) rating systems. We see many high ROI values, but also one negative and two moderate values. Since this method does not take opponents' bets into account, it is affected by how many opponents happen to have similar sheets. Moreover, its performance may degrade further over time, as more players discover www.poollogic.com and other sites that can perform these same calculations with just a few mouse clicks.

Contrarian Strategy

To increase our ROI, we will use information about how our 2003–2005 opponents bet to pick an underbet champion (i.e., the championship game's most underappreciated team) in each year, and then simply use a maximum score strategy to fill in the remainder of our

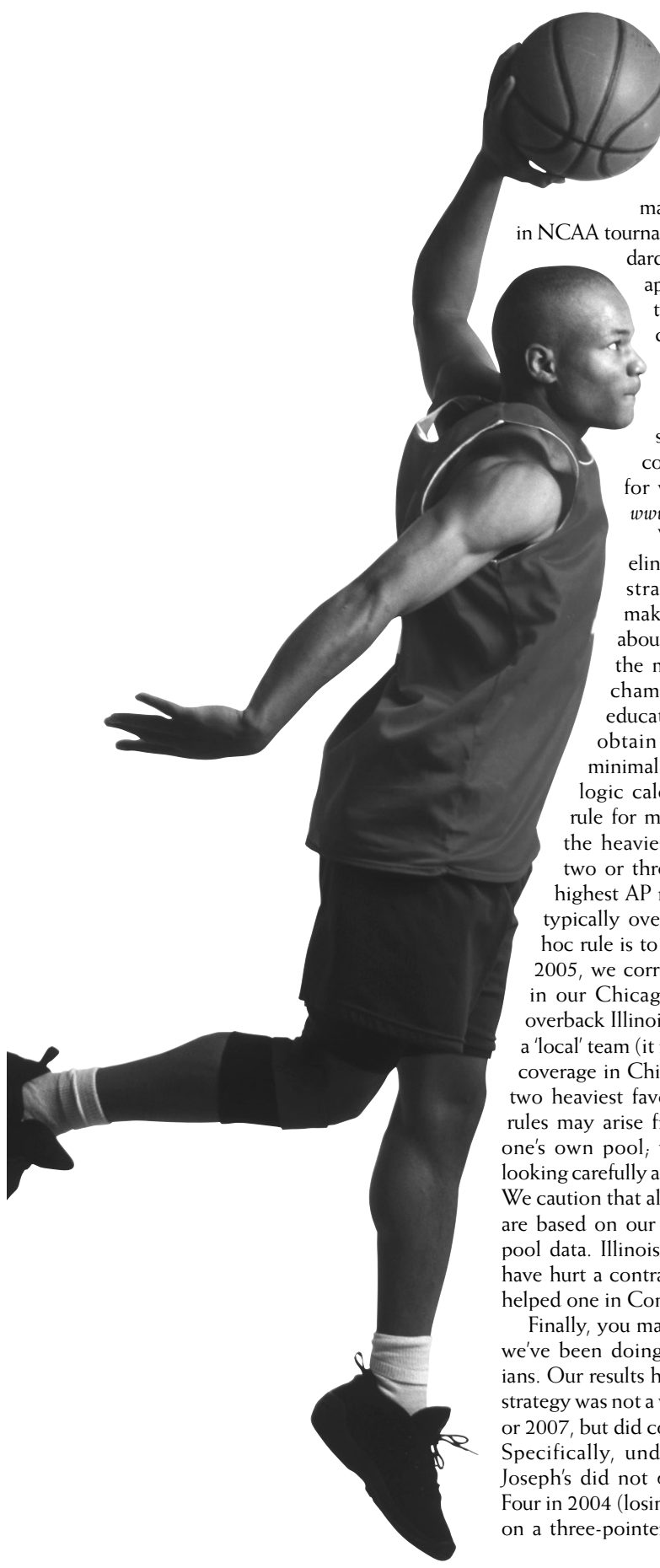
sheet. As before, our calculations may vary with the rating system we are using. If the resulting contrarian sheet does not perform well, this bodes ill for the practical setting, where opponent behavior can only be guessed.

Table 5 contains information about the teams chosen to win the championship in each year. The column headed A gives the actual number of sheets that chose that team to win the championship. Subsequent columns give the number of sheets expected to pick that team as champion (i.e., the probability that team wins the championship times the number of people in the pool) under the S, P, E, and V ratings. The most underappreciated team in the championship (i.e., the team with the biggest difference between expected and actual championship picks) under the four probability models is indicated by the corresponding letter in the rightmost column labeled U.

From Table 5, we can see that, in general, the heaviest favorites have more people choosing them than the probability models expect. An exception to this rule arises from an apparent "Duke-hating factor," as even when Duke is a favorite, it tends not to be overbacked. However, Kentucky seems overappreciated in 2003 and 2004, and the extreme devotion to Illinois in 2005 is not too surprising in this Chicago-based pool.

Returning then to our quest for a high ROI sheet, we simulate ROI for a sheet taking the most underbet champion and then score-maximizing for all previous games subject to this constraint. The results are displayed in Table 4 in the row marked "contrarian." Comparing these results to the maximum expected score results, we can see that in four of 12 cases, the ROI is the same, and in the remaining eight cases, it is higher for the underbet champion sheet. Surprisingly, the average ROI under the Predictor model is the same in all three years using both maximum score and contrarian methods, as the underappreciated champion happens to also be the most probable champion in each year. However, in all but one of the remaining cases, the contrarian approach offers an often substantial improvement. These results confirm our intuition that if we can guess how our opponents select their champions, we may be able to improve our ROI by being contrarian.






Discussion

We have shown that a contrarian strategy improved simulated ROI over straight score-maximization strategies in NCAA tournament pools with standard scoring schemes. Our approach requires only that the user select a contrarian champion and fill in the rest of his sheet using maximization of expected score subject to this constraint, free software for which is available at www.poollogic.com.

Without further modeling of opponent betting strategy, one needs to make an educated guess about which team will be the most underbet in the championship. With this educated guess, one could obtain a good sheet with minimal effort using the poollogic calculator. One ad hoc rule for most pools is to avoid the heaviest favorites (say, the two or three #1 seeds with the highest AP rankings), as they are typically overbacked. Another ad hoc rule is to avoid local teams. In 2005, we correctly guessed bettors in our Chicago-based pool would overback Illinois because it was both a 'local' team (it received heavy media coverage in Chicago) and one of the two heaviest favorites. Other ad hoc rules may arise from experience with one's own pool; we will certainly be looking carefully at Duke in future years. We caution that all our empirical results are based on our 2003–2005 Chicago pool data. Illinois' 2005 success might have hurt a contrarian in Chicago, but helped one in Connecticut.

Finally, you may be wondering how we've been doing as real-life contrarians. Our results have been mixed. Our strategy was not a winner in 2004, 2005, or 2007, but did come through in 2006. Specifically, underbet champion St. Joseph's did not quite make the Final Four in 2004 (losing to Oklahoma State on a three-pointer at the buzzer), and

2005 and 2007 were certainly not years to be contrarian, with two heavy favorites (North Carolina and Illinois in 2005 and Florida and Ohio State in 2007) successfully arriving at the championship game in those years. In 2006, however, we would have won our pool if UCLA had won the championship game, as it was, we still finished in third place. Over those four years, our total return on investment is small but positive. We find this encouraging enough that we look forward to being contrarians again during the 2008 version of March Madness. 

Editor's note: Before placing a bet, make sure it is legal in the state in which you are placing it.

Further Reading

- Breiter, D.J. and Carlin, B.P. (1997). "How To Play Office Pools if You Must." *CHANCE*, 10: 324–345.
- Carlin, B.P. (1996), "Improved NCAA Basketball Tournament Modeling via Point Spread and Team Strength Information." *The American Statistician*, 50:39–43.
- Clair, B. and Letscher, D. (2005). "Optimal Strategies for Sports Betting Pools." Department of Mathematics, Saint Louis University.
- Kaplan, E.H. and Garstka, S.J. (2001). "March Madness and the Office Pool." *Management Science*, 47:369–382.
- Metrick, A. (1996). "March Madness? Strategic Behavior in NCAA Basketball Tournament Betting Pools." *Journal of Economic Behavior & Organization*, 96:159–172.
- Niemi, J.B. (2005). "Identifying and Evaluating Contrarian Strategies for NCAA Tournament Pools." Master's thesis, Division of Biostatistics, University of Minnesota.
- Schwertman, N.C.; McCreedy, T.A.; and Howard, L. (1991). "Probability Models for the NCAA Regional Basketball Tournaments." *The American Statistician*, 45:35–38.
- Schwertman, N.C.; Schenk, K.L.; and Holbrook, B.C. (1996). "More Probability Models for the NCAA Regional Basketball Tournaments." *The American Statistician*, 50:34–38.
- Stern, H. (1991). "On the Probability of Winning a Football Game," *The American Statistician*, 45:179–183.