Matrix Completion Based Model V2.0: Predicting the Winning Probabilities of March Madness Matches

Hao Ji[1], Erich O'Saben[2], Rohit Lambi[1], and Yaohang Li[1]
[1]Department of Computer Science, Old Dominion University, Norfolk, Virginia 23529
[2]Department of Computer Science, George Mason University, Fairfax, Virginia 22030
hji@cs.odu.edu, eosaben@masonlive.gmu.edu, rlambi@cs.odu.edu, yaohang@cs.odu.edu

ABSTRACT

Observing the uncertainty property of matchup scores, in this paper, we present a new predictive model (V2.0) of coupling matrix completion process with a perturbation strategy to generate the winning probabilities for March Madness matches. We first perform the perturbation process to estimate the possible fluctuations in the outcome of regular season matches, where a set of perturbed score matrices is generated by taking into account the standard deviation of historical performance of each team. Then, matrix completion is carried out on each perturbed score matrix to estimate the potential spread in the outcome of a tournament game. Finally, the winning probability for each possible tournament game is evaluated based on the number of wins and losses in the completed matrices. We analyze the parameters, which are encountered in the perturbation process and matrix completion based on historical records of game scores, and identify appropriate values of these parameters to improve the prediction accuracy. The effectiveness of our predictive model is demonstrated in the Kaggle's March Machine Learning Mania competition 2016.

Keywords: March Madness Prediction, Matrix Completion, Perturbed Score Matrices, Winning Probabilities, March Machine Learning Mania.

1. INTRODUCTION

Each March, the National Collegiate Athletic Association (NCAA) conducts a popular college sporting event known as March Madness, a single-elimination tournament to select the national championship from 68 college basketball seeded teams [1]. One of the best parts of March Madness tournament is not only watching the great competitions, but also following the excitement of participating in a bracket challenge to predict the outcome of the tournament games. In 2016, tens of millions of bracket predictions have been created and submitted to bracket challenge contests organized by the companies and associations, for instances, NCAA [2], Kaggle [3], ESPN [4], Yahoo [5], and NBC [6].

The March Madness event has attracted the attention of researchers to apply data science to predict the winners. The early work by Colley [7] and Massey [8] predicted the match outcomes by solving systems of linear equations. Since then, there have been many developments in the field. For example, Smith and Schwertman [9] used a linear regression model and identified the nearly linear relationship between the tournament seeds and the game results. Ruiz and Perez-Cruz [10] adapted a classical soccer forecasting model to produce predictions for basketball games. Lopez and Matthews [11] designed a logistic regression model using team-based possession metrics, whose bracket won the Kaggle competition 2014. Gupta [12] developed a dual-proportion probability model with a team rating system to produce the bracket predictions. In comparison to existing models, we created a predictive model based on matrix completion approach to forecast the winning probabilities [13], which allowed us to successfully predict 49 out of 63 tournament games in March Madness 2015.

Even though our previous matrix-completion-based model worked well in March Madness prediction, a challenge we faced is how to overcome probability assignment issues that arose because of the uncertainty property of matchup scores. It is well known that the final scores played by the same two teams may vary significantly if the match were performed again, due to a fluctuation in the relative strength of the teams. Sometimes an upset happens [14], where a lower-seeded team beats a higher-seeded team. Therefore, a winning probability based on a range of the potential outcomes of a game would be able to yield more accurate prediction results, compared to a single instance of predicted scores.

In this paper, we present a new predictive model (V2.0) that combines matrix completion and a perturbation process to generate the winning probabilities of March Madness matches. First of all, we construct a set of perturbed score matrices to account for the possible fluctuations in the outcome of regular season matches. Then, we apply matrix completion to each perturbed score matrix to estimate the range of the potential outcomes of a tournament game. As a result, the predicted winning probability of each possible tournament game is calculated from the corresponding entries in the completed matrices.

The rest of the paper is organized as follows. Section 2 describes the proposed predictive model. The results are shown in section 3. Finally, Section 4 concludes the paper.

2. METHODS

The predictive model V2.0 incorporates three primary components: (1) perturbation process, which generates a set of perturbed score matrices. (2) matrix completion, which completes the perturbed score matrices. (3) probability adjustment, where the predicted winning probabilities are derived from the completed score matrices.

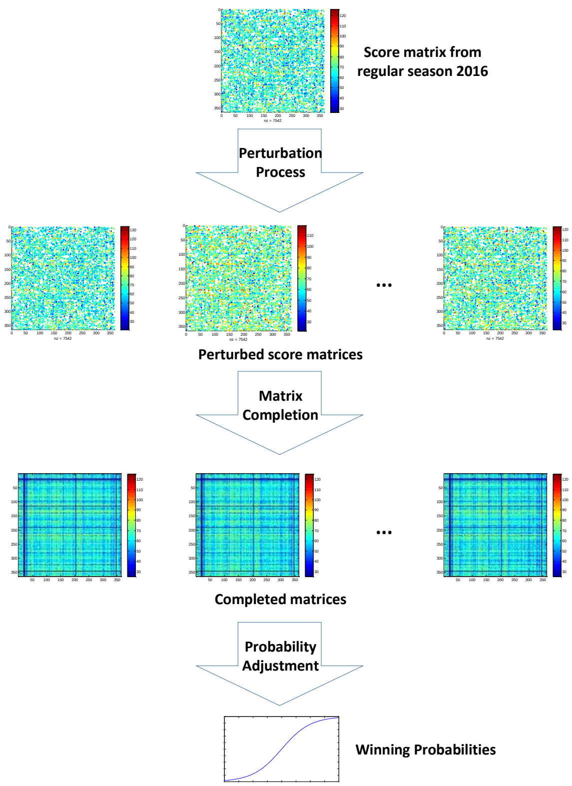


Figure 1. Procedure of the predictive model V2.0

Figure 1 presents the procedure of the proposed predictive model V2.0. Compared to our last year's predictive model (V1.0) [13], the model V2.0 relies only on the score information of regular season matches to predict the winning probabilities, instead of considering relevant game details, such as assists, turnovers, and teams' rank. Moreover, the perturbation process is introduced to estimate the possible fluctuations on matchup scores. The perturbation process can improve the prediction accuracy on the potential upsets in tournament, which will be demonstrated in Section 3.

We participated in the bracket challenge contest "the March Machine Learning Mania 2016", which is hosted by Kaggle.com. Each participant group can submit at most two bracket predictions containing the winning probabilities of every possible matchup in the tournament.

The Log Loss function below is employed to evaluate each submission,

$$Logloss = -\frac{1}{n}\sum_{i=1}^{n}(y_i \log(p_i) + (1 - y_i)\log(1 - p_i))$$

where n is the number of games, p_i is the winning probability of team 1 to win over team 2, and y_i equals 1 if team 1 wins over team 2 and 0 otherwise. The bracket prediction with a smaller value of $LogLoss$ achieves better prediction accuracy.

2.1 Perturbation Process

We formulate score records from regular season 2016 into matrix format, where 364 teams are placed in rows and columns and each nonzero entry stores a matchup score. Figure 2 shows the colormap of the score matrix 2016. It can be seen that the score matrix is an incomplete matrix, where most of the entries are unknown.

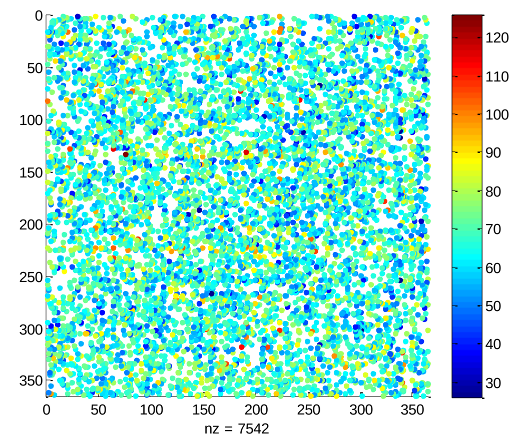


Figure 2. The Colormap of the score matrix 2016

In basketball games, it is common that teams experience random fluctuations in their performance. Therefore, there exists such a bounded range of scores on every entry in the score matrix, and any value in the range is likely to happen in real competitions. To this end, we generate a set of independently perturbed score matrices to sample and estimate the possible fluctuations in the outcome of regular season matches.

Let M denote the incomplete score matrix and Ω be a set of the indices of the matchup scores from the regular season. Each perturbed score matrix as a sample is created by adding a Gaussian random perturbation to each nonzero score entry in the score matrix, such that

$$M_{ij} = M_{ij} + t_i \ \text{ for } (i, j) \in \Omega$$

where t_i is a random variable for team i which follows the normal distribution $N(0, (s\sigma_i)^2)$.

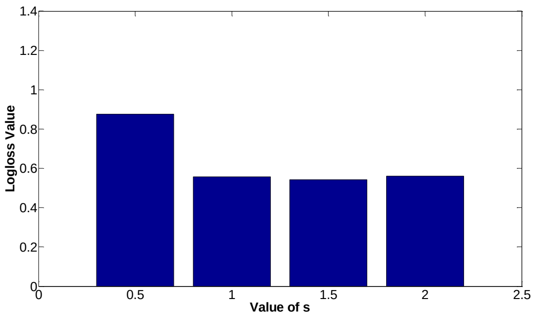


Figure 3. The average Logloss of the predictions for years 2012-2015

The standard derivation $s\sigma_i$ is specified as a multiple of the standard derivation σ_i of the game scores of team i played in the past. The scalar s is a positive number which is tuned to obtain a smallest average Logloss value of predictions. Our analysis found that $s = 1.0$ and $s = 1.5$ obtain the smallest Logloss values on historical records for 2012 to 2015, as shown in Figure 3.

2.2 Matrix Completion

Matrix completion is the process of recovering the unknown entries of an incomplete matrix based on a small set of observed samples [15, 16, 17]. In our model, we apply the Singular Value Thresholding (SVT) algorithm [18], one of the popular matrix completion approaches, to complete each perturbed score matrix. In theory, the SVT algorithm seeks a low-rank matrix X that minimizes the following Lagrange dual function,

$$\tau\|X\|_* + \frac{1}{2}\|\mathcal{P}_\Omega(X) - \mathcal{P}_\Omega(M)\|_F^2$$

where $\mathcal{P}_\Omega$ is the projection operation and τ is a Lagrange multiplier trading off between the nuclear and Frobenius norm.

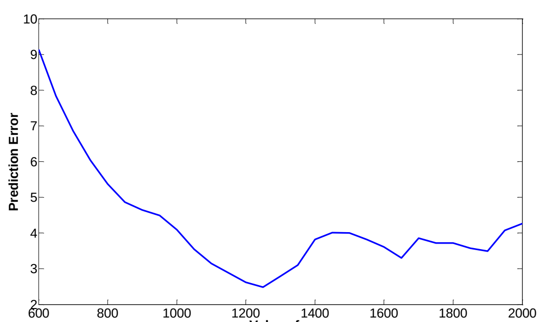


Figure 4. The prediction error for years 2012-2015

In general, parameter τ is specified to be a factor of $\sqrt{mn}$, such that $\tau = \omega\sqrt{mn}$, where m and n denote the dimension of the incomplete matrix and ω is a positive number. In order to figure out a satisfactory ω value, we use MSE (mean squared error) to measure the prediction error between the predicted scores from the completed matrices and the actual tournament scores from 2012-2015. Since the MSE for each year may differ significantly, to determine the optimal ω we chose the value that performs the best over all years. This was calculated by:

$$\omega^* = \underset{600 \le \omega \le 2000}{\arg\min} \left(\sum_{y=2012}^{2015} \left(e_{y\omega} - \min_{600 \le t \le 2000} \{ e_{yt} \} \right) \right)$$

where y is the tournament year and $e_{y\omega}$ is the MSE value for a completed matrix in a given year at a ω between 600 and 2000. Figure 4 shows prediction errors on the tournament games 2012-2015 at different ω. As a result, $\omega^* = 1250$ becomes an obvious choice for our model which achieves the smallest prediction error.

2.3 Probability Adjustments

By counting the number of wins by the teams from a set of completed matrices, in the model V2.0, we use the following equation to generate a winning probability $p_{team1,team2}$ of a game that team 1 beats team 2,

$$p_{team1,team2} = \frac{nwins_{team1}}{nwins_{team1} + nwins_{team2}}$$

where $nwins_{team1}$ and $nwins_{team2}$ denote the number of wins by team 1 and team 2, respectively. Additionally, based on the tournament statistics [1] that no team with seed 16 has ever won a team with seed 1, we apply the following rule

$$p_i = \begin{cases} 1 & if\ Seed_{team1} = 1\ \ and\ \ Seed_{team2} = 16 \\ 0 & if\ Seed_{team1} = 16\ \ and\ \ Seed_{team2} = 1 \end{cases}$$

to lower the $LogLoss$ value of our predictive model.

3. RESULTS

By generating and completing 1000 perturbed scores matrices, our predictive model V2.0 generates the winning percentages for 2278 potential tournament games in 2016. For simplicity, the resulting two brackets with $s = 1.0$ and $s = 1.5$ are shown in Appendix, respectively.

Figure 5 presents the actual result of the March Madness 2016 [19], where the games we predicted correctly are highlighted in red color and the ones we lost in green and blue colors. One can find that our predictive model V2.0 is able to predict accurately the win/lose outcome of 47 out of 63 tournament games. More importantly, based on the perturbation process, we successfully predicted 11 out of 20 upset games (55%), which outperforms our last year's result that only 4/12 upsets (33%) were forecasted [13]. However, the final $LogLoss$ score of our prediction (0.598446) is slightly greater than that of last year by 0.068899. This is due to the fact that more severe upsets occurred this year, which heavily penalize our prediction. As shown in green color in Figure 5, for example, No.1, No. 2, No. 3, and No. 4 seed teams unfortunately lost their tournament games, which largely increase our $LogLoss$ score by 0.0807.

4. CONCLUSION

The predictive model V2.0 for March Madness 2016 is presented. To take into account the uncertainties of matchup scores by the teams, the perturbation process and matrix completion on score matrices are carried out to estimate the potential spread in the outcome of a tournament game. The predicted winning probabilities are then evaluated from the corresponding entries in the completed perturbed score matrices.

The predictive model proposed in this paper takes into account only score records. To gain further improvement in the prediction accuracy, our future work will focus on building a comprehensive predictive model involving relevant game details, such as assists, turnovers, and teams' rank.

ACKNOWLEDGEMENTS

Hao Ji acknowledges support from ODU Modeling and Simulation Fellowship. This research was supported by the Turing High Performance Computing cluster at Old Dominion University.

REFERENCES

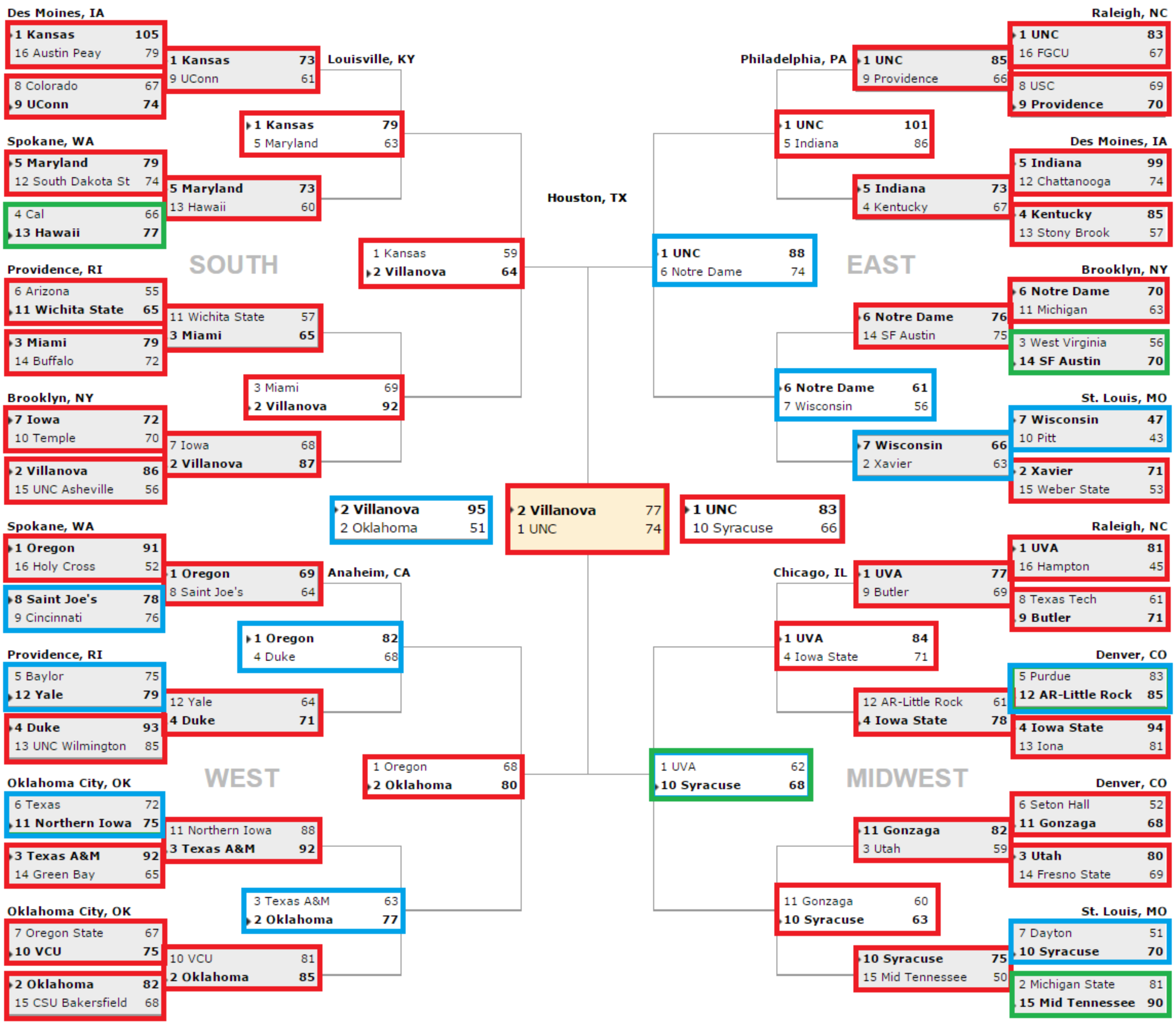


Figure 5. The actual result of March Madness 2016

[1] NCAA Men's Division I Basketball Championship. available at: http://en.wikipedia.org/wiki/NCA-A_Men%27s_Division_I_Basketball_Championship
[2] NCAA Bracket Challenge. available at: http://www.bracketchallenge.ncaa.com
[3] Kaggle March Machine Learning Mania 2016. available at: https://www.kaggle.com/c/march-machine-learning-mania-2016
[4] ESPN Tournament Challenge. available at: http://games.espn.go.com/tournament-challenge-bracket/2016/en/game
[5] Yahoo Bracket Challenge available at: https://tournament.fantasysports.yahoo.com/
[6] NBC Sports Bracket Madness. available at: http://madness.nbcsports.com/
[7] Colley W. N., 2002, "Colley's Bias Free College Football Ranking Method: The Colley matrix explained." available at: http://www.colleyrankings. com /method.html.
[8] Massey K., 1997, "Statistical models applied to the rating of sports teams." undergraduate honors thesis, Bluefield College.
[9] Smith, T., and Schwertman, N.C., 1999, "Can the NCAA basketball tournament seeding be used to predict margin of victory?" The american statistician, 53, no. 2: 94-98.
[10] Ruiz, F. J., and Perez-Cruz, F, 2015, "A generative model for predicting outcomes in college basketball." Journal of Quantitative Analysis in Sports 11 no. 1: 39-52.
[11] Lopez, Michael J., and Gregory Matthews, 2014, "Building an NCAA mens basketball predictive model and quantifying its success." arXiv preprint arXiv: 1412.0248.
[12] Gupta, A. A., 2015, "A new approach to bracket prediction in the NCAA Men's Basketball Tournament based on a dual-proportion likelihood." Journal of Quantitative Analysis in Sports. 11, no. 1: 53-67.
[13] Ji, H., O'Saben, E., Boudion, A., and Li, Y., 2015, "March Madness Prediction: A Matrix Completion Approach." In Proceedings of Modeling, Simulation, and Visualization Student Capstone Conference, Suffolk, VA.
[14] Bryan, K., Steinke, M., and Wilkins, N., 2006, "Upset Special: Are March Madness Upsets Predictable?" Available at SSRN 899702.
[15] Candès, Emmanuel J., and Benjamin Recht, 2009, "Exact matrix completion via convex optimization." Foundations of Computational mathematics 9, no. 6: 717-772.
[16] Recht, B., Fazel, M., and Parrilo, P.A., 2010, "Guaranteed minimum-rank solutions of linear matrix equations via nuclear norm minimization." SIAM review 52 no.3: 471-501.
[17] Ji, H., and Li, Y., 2015, "Monte Carlo Methods and Their Applications in Big Data Analysis." In Mathematical Problems in Data Science, pp. 125-139. Springer International Publishing.
[18] Cai, J.F., Candès, E.J. and Shen, Z., 2010, "A singular value thresholding algorithm for matrix completion." SIAM Journal on Optimization 20 no. 4: 1956-1982.
[19] NCAA Tournament Bracket – 2015. Available at : http://espn.go.com/mens-college-basketball/tournament/bracket

BIOGRAPHIES

Hao Ji is a Ph.D. student in the Department of Computer Science at Old Dominion University. He received the B.S. degree in Applied Mathematics and M.E. degree in Computer Science from Hefei University of Technology in 2007 and 2010, respectively. His research interest include High Performance Scientific Computing, Monte Carlo Methods, and Big Data Analysis.

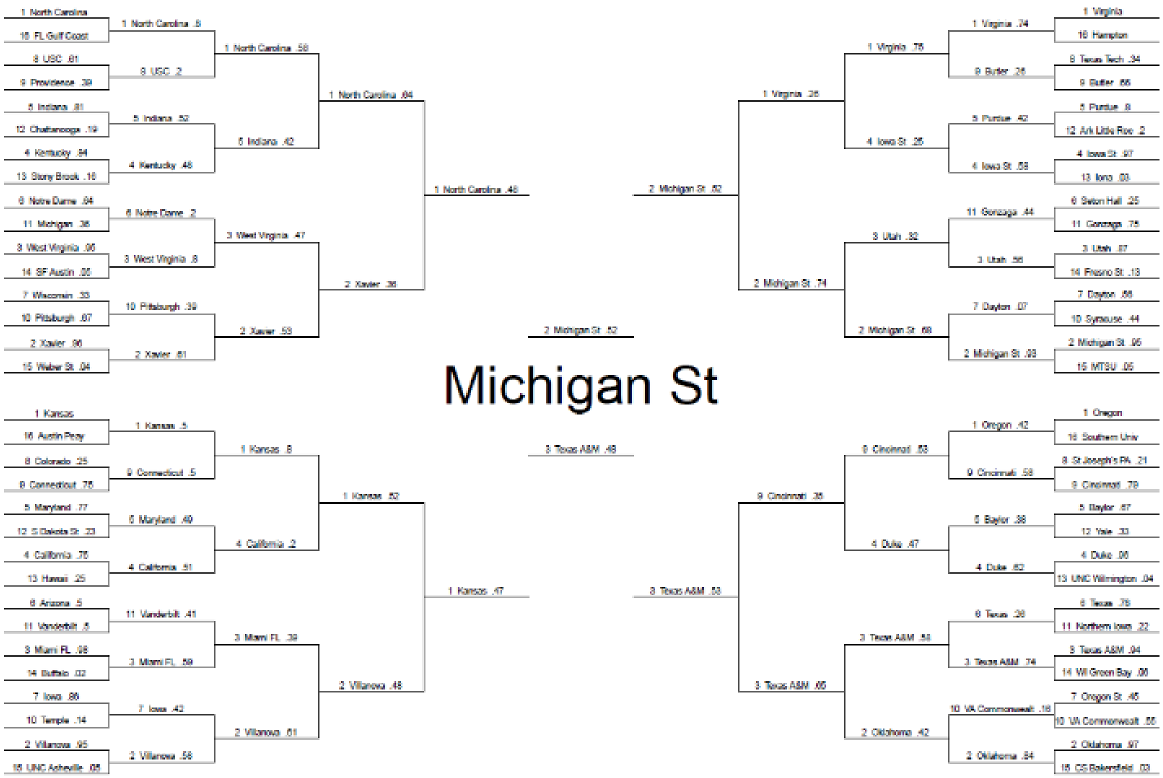
Erich O'Saben is a Ph.D. student in the Department of Computer Science at George Mason University. He received his B.S. in Computer Science from Old Dominion University. His primary interests are in artificial intelligence, machine learning, and predictive analytics.

Rohit Lambi is a M.S student in the Department of Computer Science at Old Dominion University with a B.S. degree in Information Technology from Pune University, India. He has worked for 3 years as a Software Engineer in Praxify Technologies, Inc. Also, as a M.S student he was a Graduate Research Assistant where he worked on different kinds of software development projects and a machine learning research project. His areas of interest include software development, predictive analytics and machine learning.

Yaohang Li is an Associate Professor in Computer Science at Old Dominion University. He received his B.S. in Computer Science from South China University of Technology in 1997 and M.S. and Ph.D. degrees from the Department of Computer Science, Florida State University in 2000 and 2003, respectively. After graduation, he worked as a research associate in the Computer Science and Mathematics Division at Oak Ridge National Laboratory, TN. His research interest is in Computational Biology, Monte Carlo Methods, and High Performance Computing.

APPENDIX

(1) The first bracket



(2) The second bracket