Application of the Bradley-Terry Rating System to USAU 2015 Regular Season College Results

Wally Kwong

September 1, 2015

Abstract

The Bradley-Terry rating system was applied to USA Ultimate 2015 college results. A numerical solution for the ratings was found using Newton's method. Each point between two teams is treated as a separate discrete event. Analysis is done to extrapolate expected game results and expected game scores based upon the ratings. Game rating differentials are time weighted via exponential decay. Games occurring in the first week having weight of 0.5, while games occurring in the most recent week have weight of 1.0. Exceptions are made to ignore results between teams with highly differing ratings. In order to calculate team ratings, a numerical method is used. Teams are initially assigned a game rating of 1000, and ratings are iteratively updated via a numerical method until a set convergence is obtained.

1 Background

1.1 Current USAU Algorithm

Results from the USAU college regular season are used to determine the allocation of bids to postseason events. In its current form, USAU uses a time weighted average of game rating differentials ΔR as determined by the formula[1]

$$\Delta R = 125 + 475\frac{\sin\left(\min\left(1, \frac{1-r}{0.5}\right) \cdot 0.4\pi\right)}{\sin\left(0.4\pi\right)}$$

where the value of r is determined from the final score by

$$r = \frac{\text{losing score}}{\text{winning score} - 1}$$

[1] http://play.usaultimate.org/teams/events/rankings/#algorithm

1.2 Bradley-Terry Model

The Bradley-Terry model assumes that given two individuals i and j with respective ratings of R_i and R_j, then the probability of observing the pairwise comparison $i > j$ is

$$\Pr\left(i > j\right) = \frac{R_i}{R_i + R_j} \tag{1}$$

Due to the nature of this assumption, the model implies that relations are transitive.

This model is currently applied to NCAA college hockey results via the unofficial KRACH ratings[2][3]. College hockey suffers similar issues

[2] http://www.mscs.dal.ca/∼butler/krachexp.htm

[3] http://www.uscho.com/rankings/krach/d-i-men/

to ultimate in its lack of immediate connectivity and in its selection of teams to postseason events.

The Bradley-Terry model has issues when dealing with undefeated or winless teams. An undefeated team will have an infinite rating, causing an issue for convergence when using numerical methods. A similar issue arises with a winless team obtaining a 0 rating, thus creating a second order effect with their opponents. This is addressed in the KRACH by assigning each team a fictional draw (i.e. an additional win of 0.5) against a fictional team of "average" rating.

2 Methods

Data was collected from the USAU website for all sanctioned tournaments during the 2015 college regular season. This comprised of 78 tournaments and 3501 games in the men's division, and 65 tournaments and 2032 games in the women's division. Data on some tournaments was edited in order to ignore non-rostered teams playing at sanctioned events (e.g. high school and alumni teams). Games that were ignored by USA Ultimate for various reasons (e.g. academic ineligibility etc) were not removed from this data set. This data should not be considered as clean as any official listings provided by USA Ultimate.

From the season results, actual matchups and win results can be observed. The expected wins is then

$$W_i = \sum_j N_{ij} \frac{R_i}{R_i + R_j}$$

where W_i is the expected number of wins for team i and N_{ij} is the number of times teams i and j have played each other.

Newton's method can be used to provide a numerical process to obtain true ratings. The first order correction in this case is then

$$\begin{aligned}
\frac{\partial W_i}{\partial R_i} &= \frac{\partial}{\partial R_i} \sum_j N_{ij} \frac{R_i}{R_i + R_j} \\
&= \sum_j N_{ij} \frac{\partial}{\partial R_i} \frac{R_i}{R_i + R_j} \\
&= \sum_j N_{ij} \left(\frac{1}{R_i + R_j} - \frac{R_i}{(R_i + R_j)^2} \right) \\
&= \sum_j N_{ij} \frac{R_j}{(R_i + R_j)^2}
\end{aligned}$$

which can then be used in the approximate correction

$$\Delta R_i = \frac{\Delta W_i}{\partial W_i / \partial R_i}$$

where ΔR_i is the differential to be applied to the estimated team rating, and ΔW_i is the difference in observed and estimated wins.

Teams were given an initial rating of 1.0 and 5 tied games against a fictional opponent of rating 1.0. Game were not time weighted for exponential decay. Games recorded as ties or double forfeits were ignored. Games recorded as a forfeit or recorded as a win without corresponding scores were arbitrarily treated as a 15-6 score. A lower bound of 0.001 was set as the minimum team rating to avoid issues due to overcorrection. Newton's method was repeatedly applied until a maximum correction in the team rating of $\epsilon < 10^{-7}$ was obtained.

In addition to considering the games as a discrete event, a single point of ultimate may be considered as a "mini-game" and a discrete event in its own right. In this treatment, a game with a score of 15-8 is considered as 23 mini-games, with the "winning" team obtaining 15 wins and 8 losses. This gives increased sample size, and

the ability to predict precise game scores via the binomial distribution. Teams were not given fictional results against an "average team" in this scenario, since no team had entirely shutout victories or losses.

3 Results and Analysis

Following the method described in the previous section, the rating for each team in the USA Ultimate 2015 men's and women's college regular season was calculated. Calculated ratings for men's teams using win-loss record are shown in Table 1. Results using point scored (i.e. treating each point as a "mini-game") are shown for men's in Table 2 and for women's in Table 3.

At first glance, generating ratings using win loss record does not pass the "eye test"[4]. Teams with an undefeated record (e.g. Tennessee-Chattanooga, Franciscan, and UCLA (B) in the men's division) have relatively inflated rankings, while Carleton has a ranking of $R = 3.351$ (not shown in Table 1 since Carleton is not in the top 30). Using the underlying assumption of the Bradley-Terry model (the expected win percentage as defined in Equation 1), this would imply should Carleton repeatedly play Tennessee-Chattanooga, then Carleton would only win 27.5% of the games. This result seems rather contrived.

The ratings generated from points scored (Table 2) seems to better pass the "eye test". A key difference here is that the ratings no longer correspond to the probability a team will win a particular game, but rather the probability a team will score on a given point. Calculating the probability of a team winning a game is then based upon the binomial distribution. For com-

[4]Subjective? Absolutely.

Team	Rating
Pittsburgh	34.782
Oregon	20.727
North Carolina-Wilmington	19.709
Florida State	14.044
Washington	12.607
Colorado	11.748
North Carolina	9.375
Central Florida	8.674
Tennessee-Chattanooga	8.189
California-Santa Barbara	7.549
Texas A&M	7.233
Maryland	7.196
Rice	7.113
Franciscan	6.913
Florida	6.118
Amherst	5.651
Georgia	5.406
Ohio	5.172
Purdue	5.106
Cincinnati	4.853
Wisconsin	4.641
Arizona State	4.545
Minnesota-Duluth	4.422
Brandeis	4.164
Lewis & Clark	4.008
Tulane	3.987
Iowa	3.985
Texas	3.916
Chico State	3.898
UCLA (B)	3.773

Table 1: Top 30 college men's teams using Bradley-Terry ratings applied to win-loss record

Team	Rating
Pittsburgh	3.350
North Carolina-Wilmington	3.282
Oregon	3.053
Colorado	2.951
North Carolina	2.922
Georgia	2.803
Texas A&M	2.795
Massachusetts	2.788
Florida	2.751
Florida State	2.719
Washington	2.689
Central Florida	2.685
Cincinnati	2.611
Wisconsin	2.517
Arizona State	2.478
Minnesota	2.434
Carleton College	2.404
Texas	2.326
British Columbia	2.317
Maryland	2.302
Stanford	2.176
California-Santa Barbara	2.151
Tufts	2.098
Michigan	2.011
Auburn	1.996
UCLA (B)	1.985
Iowa State	1.949
Harvard	1.945
Missouri	1.944
Tulane	1.861

Table 2: Top 30 college men's teams using Bradley-Terry ratings applied to points scored

Team	Rating
Oregon	5.520
British Columbia	5.194
Stanford	4.870
Virginia	4.141
Colorado	3.771
UCLA	3.695
Washington	3.558
Carleton College	3.475
Central Florida	3.219
Florida State	3.081
Whitman	3.045
Tufts	3.011
Dartmouth	2.903
Pittsburgh	2.848
Notre Dame	2.835
Victoria	2.812
Ohio State	2.800
Northeastern	2.755
Kansas	2.637
Western Washington	2.627
Colorado College	2.574
Georgia	2.403
Texas	2.369
Middlebury	2.342
Southern California	2.302
California	2.284
Vanderbilt	2.219
Minnesota	2.175
California-Davis	2.096
Wisconsin	2.046

Table 3: Top 30 college women's teams using Bradley-Terry ratings applied to points scored

parison purposes, Tennessee-Chattanooga has a rating of $R = 1.342$ using points scored. Using this model, Carleton would hypothetically win 94.7% of games against Tennessee-Chattanooga.

Any model has the potential to give facetious results; evaluating its overall accuracy and usefulness is subjective. Using the ratings from points scored will still yield results such as UCLA (B) being placed in the top 30. For the remainder of this article though, the ratings based upon points scored will be used. The ratings based upon win loss record will not be further discussed.

3.1 Calculation of game win probability from points ratings

As stated before, the ratings now correspond with the probability of a team scoring on a particular point. The probability a team wins an uncapped game to 15 with no overtime is then obtained via the binomial distribution of scoring at least 15 points out of 28:

$$\Pr(i > j) = \sum_{n=15}^{28} \left(\binom{28}{n} r^n (1-r)^{28-n} \right)$$

where the value of r is defined by Equation 1. The probability that a game does go to overtime is

$$\Pr(\text{overtime}) = \binom{28}{14} r^{14} (1-r)^{14}$$

and the conditional probability that a team does win in overtime is similarly obtained by calculating the probability of obtaining the score prior to game point, and multiplying by an additional factor of r.

For example purposes, the calculated probability in the men's division of Pittsburgh beating a team in the top 10 is shown in Table 4.

Pittsburgh's Opponent	**Probability**
North Carolina-Wilmington	0.5229
Oregon	0.6024
Colorado	0.6386
North Carolina	0.6489
Georgia	0.6909
Texas A&M	0.6937
Massachusetts	0.6962
Florida	0.7091
Florida State	0.7201

Table 4: Probability of Pittsburgh defeating teams according to points scored ratings

3.2 Point spreads

In addition to determining the winner of a game, the likelihood of a specific score can also be calculated. The likelihood of team i defeating team j with a particular score $S_i > S_j$ with a winning score of $S_i = 15$ without going to overtime is

$$\Pr(S_i, S_j) = \binom{S_i + S_j - 1}{S_i - 1} r^{S_i} (1-r)^{S_j}$$

This differs from the standard binomial distribution, since the last point of the game must necessarily have been scored by the winning team. (A game ending in 15-12 must necessarily have been preceded by 14-12. 15-11 is an impossible preceding case, since the game would have ended at that point. The modification in the above equation accounts for this correction to the binomial distribution.)

Table 5 shows the probabilities of various scores occurring if Pittsburgh ($R = 3.350$) were to play North Carolina-Wilmington ($R = 3.282$). The teams have similar ratings, and Table 4 shows Pittsburgh only having a 52.29% edge over North Carolina-Wilmington. The probability of

Pittsburgh	UNC-W	Probability
15	Up to 8	0.1143
15	9	0.0517
15	10	0.0614
15	11	0.0691
15	12	0.0741
15	13	0.0762
16	14	0.0381
17	15	0.0191
17	16	0.0188
16	17	0.0185
15	17	0.0183
14	16	0.0365
13	15	0.0731
12	15	0.0697
11	15	0.0637
10	15	0.0555
9	15	0.0456
Up to 8	15	0.0963

Table 5: Probability of final score in an uncapped game between Pittsburgh ($R = 3.350$) and North Carolina-Wilmington ($R = 3.282$)

each outcome is not heavily skewed in one direction or another.

A slight increase in relative team strength does have a large impact on expected score. Taking Florida State ($R = 2.719$) to be Pittsburgh's opponent, Table 6 shows the probabilities of the resulting score. The most likely score is 15-11, and there is a 12.82% chance of the game reaching overtime.

At first glance, the probability of overtime ($P = .1282$) in this scenario occurring may seem high compared to the probabilities of 15-13 ($P = 0.0791$) and 13-15 ($P = 0.0521$). This is explainable with the same logic previously used

Pittsburgh	Florida State	Probability
15	Up to 6	0.0995
15	7	0.0567
15	8	0.0698
15	9	0.0800
15	10	0.0860
15	11	0.0875
15	12	0.0850
15	13	0.0791
16	14	0.0391
17	15	0.0193
17	16	0.0173
16	17	0.0141
15	17	0.0127
14	16	0.0257
13	15	0.0521
12	15	0.0454
11	15	0.0380
Up to 10	15	0.0926

Table 6: Probability of final score in an uncapped game between Pittsburgh ($R = 3.350$) and Florida State ($R = 2.719$)

to modify the binomial distribution – winning a game with a score of 15-13 requires that the previous score be 14-13, and the leading team scoring the winning goal. On the other hand, obtaining a 14-14 result has two potential sources – a 14-13 game, or a 13-14 game.

Using the probability of individual score outcomes, the expected final point differential can be calculated via

$$\langle S_i - S_j \rangle = \sum_{S_i, S_j} (\Pr(S_i, S_j) \cdot (S_i - S_j))$$

For the case of Pittsburgh and North Carolina-Wilmington, the expected point differential is $\langle S_{Pitt} - S_{UNCW} \rangle = 0.2666$. For the case of Pittsburgh and Florida State, the expected point differential is $\langle S_{Pitt} - S_{FSU} \rangle = 2.6554$.

3.3 Strength of Schedule

A team's strength of schedule can be obtained from the ratings by[5]

$$\mathrm{SOS}_i = \frac{\sum_j N_{ij} R_j / (R_i + R_j)}{\sum_j N_{ij} / (R_i + R_j)}$$

If strict adherence to the premise that each point is a "mini-game" was applied, the value of N_{ij} would be the total points between two teams. However this seems slightly incorrect for this scenario, as a 15-6 game would be weighted differently than a 17-16 game. The teams really only played 1 game, despite any modeling assumptions made. For the purposes of strength of schedule calculation, N_{ij} was taken to be the actual number of games played between two teams. The results for both men's and women's strength of schedule are respectively shown in Tables 7 and 8.

[5] http://dbaker.50webs.com/method.html

Team	Strength of Schedule
Wisconsin	2.358
Pittsburgh	2.306
Central Florida	2.264
Texas	2.214
Florida	2.204
British Columbia	2.181
Carleton College	2.176
Texas A&M	2.152
Massachusetts	2.131
Auburn	2.119
North Carolina	2.116
Arizona State	2.084
Harvard	2.046
Georgia	1.994
Colorado	1.986
North Carolina-Wilmington	1.978
Florida State	1.970
Oregon	1.962
Stanford	1.938
Illinois	1.907

Table 7: Men's teams with highest strength of schedule (per game)

Team	Strength of Schedule
British Columbia	3.386
Victoria	3.216
Whitman	2.962
Stanford	2.919
Dartmouth	2.867
Washington	2.859
Carleton College	2.813
UCLA	2.723
Colorado	2.696
Northeastern	2.585
Oregon	2.526
Kansas	2.426
Tufts	2.416
Ohio State	2.411
Central Florida	2.345
Western Washington	2.319
Iowa State	2.176
California-San Diego	2.116
Wisconsin	2.089
Pittsburgh	2.073

Table 8: Women's teams with highest strength of schedule (per game)

4 Discussion

4.1 Accuracy of Model

Like any model, this treatment makes assumptions which are likely not borne in reality. The Bradley-Terry model assumes that every game (and in the modification used, every point) that each team plays equal. No considerations are made for team depth, offensive or defensive lines, or cap limits. If only considering the result win, trading points and using offensive / defensive lines will serve to reduce variance and decrease the probability of a potential upset. Capping a game and reducing its length will increase the probability of an upset.

The details drawn from the model are likely skewed as well. The probability of a blowout win is likely overstated. Teams may have a more open rotation in such a scenario in order to rest starting players. Team depth would become more of a factor, and explicitly accounting for team depth is not done in this model.

4.2 Bids to 2015 College Nationals

Using this model as the basis for determining bids to the 2015 USA Ultimate college championships would result in the allocations as shown in Table 9. On the men's side, the sole difference is the strength bid awarded to Maryland and the Atlantic Coast region under the current USAU algorithm is instead awarded to Minnesota and the North Central region. Of note is the fact that Cincinnati still earns a strength bid for the Ohio Valley region.

On the women's side, one of the strength bids awarded by the current USAU algorithm to the South Central region is shifted to the Northeast region by virtue of Northeastern's ratings. The question of which South Central team "lost" the

Region	Men's Bids	Women's Bids
Atlantic Coast	2	1
Great Lakes	1	1
Metro East	1	1
North Central	3	1
Northeast	1	3
Northwest	2	5
Ohio Valley	2	2
South Central	3	2
Southeast	4	2
Southwest	1	2

Table 9: Distribution of bids to 2015 college championships using the Bradley-Terry system

bid is up for debate; in the USAU rankings, Colorado College is ranked above Kansas, while the opposite is true if the Bradley-Terry system is used.

5 Acknowledgments

Stephen Wang was immensely helpful in discussions I had with him. He suggested that I treat each point as a "mini-game", and also steered me towards using numerical methods rather than trying to invert an enormous matrix.