Rating/Ranking Teams Through the (Spanning) Trees

John A. Trono*

**Saint Michael's College, Colchester, Vermont 05439, USA + email address: jtrono@smcvt.edu*

Abstract. The games in a college football season can be used to create a graph, allocating one vertex per team, and assigning each game's score differential as the weight for a directed edge connecting those two teams. Several different approaches will be described herein that attempt to determine the relative strengths of all participating teams. The performance of these approaches will be evaluated by the post season predictions (deduced from the ratings) for the teams invited to compete in a sanctioned bowl game. By restricting the margin of victory to a single point, these techniques will also be evaluated with respect to the ranking systems that contribute to the Bowl Championship Series (BCS) rankings.

1. Introduction

Computers are very useful tools when it comes to exploring ideas that require a significant amount of arithmetic. Inventors of mathematical models, that evaluate how well an athletic team has performed in a given season, rely on this device's speed and accuracy to produce the specified results. In fact, many popular strategies for rating (and/or ranking) a large set of teams would be infeasible to calculate by hand, whereas a computer can generate the results almost instantly.

Many algorithms for assigning ratings to a collection of teams necessarily repeat their underlying calculations until the assigned ratings converge to one final set of values. A specific tolerance value, for how close consecutive ratings for each team must be before a program has recognized that convergence has been achieved, can be an input to this class of algorithms. A smaller tolerance produces a more accurate result, but typically causes more iterations to be completed by the computer (which may delay the final results substantially).

Other algorithms are nondeterministic, using pseudo random number generators during their execution. This style of computation must be run to completion thousands - or even millions - of times so that the average of all these samples can accurately approximate the results of such a probabilistic technique. Once again, devoting more time to producing more random samples will allow for any anomalous result that is generated to be overwhelmed by those that are not, which in turn should help the program produce results which are almost exactly the same as what the underlying model would yield if the average was calculated after an infinite number of iterations (which is obviously not possible).

2. Margin of Victory

Most rating systems will use all available pair-wise comparison data to produce an evaluation of how strong each team is because those specific outcomes provide a measure that everyone would agree to. By definition, one team that has defeated another has proven that during that specific contest, the winning team was 'better' because their team's entire effort was quantifiably larger that day, according to the rules that govern how final outcomes are determined. However, some final scores may not be indicative of the relative disparity between the abilities of the winner and the loser, especially when the difference between the two team's scores in that game is significant. Several authors have suggested limiting how high a score differential to allow in the dataset when calculating any such system's ratings. (Typically, for football, these limits range between twenty and thirty points.)

When teams that are fairly well matched compete, their best players will play the majority of the game - unless they are injured, suspended, or ejected from the game - and the result will be as good a measure for that particular contest as possible. However, once one team begins to dominate another, coaches will eventually begin to substitute in other, less talented players to: give them some game-time experience; lessen the likelihood of an injury to those key players who are removed from the game; and to not run up an even larger score difference as a sign of good sportsmanship. Once the second and third string players are in the game, they may continue the onslaught since they will be putting forth their best efforts to demonstrate their abilities to their coaches, teammates (and fans). In such mismatches, the losing team may still believe that they

can produce a victory, and their starters may narrow the score differential quite a bit, especially as the game draws to a close. Games like this provide some information, but may also confuse systems that only see the final score. For instance, Nebraska was leading Oklahoma State 49-14 in 1988, with less than one third of the game left to be played. This 35 point margin was almost reduced by a factor of two as Oklahoma State scored a couple of 'meaningless' touchdowns in the last few minutes of the game, to shrink the margin to 21 points. (The final score was 63-42.) If it could somehow be determined that team A is truly 40 points 'better' than team B, and the final score differential is much larger - or smaller - than this quantity, it is likely that these non-characteristic scores will adversely affect results produced by an objective rating system.

In light of this discussion, many proponents feel that the systems which rank teams, to determine those teams that have achieved the most outstanding seasons, should be used in certain situations instead of rating systems that may be more accurate for predicting subsequent contest results than for overall, 'season evaluation'. Many ranking systems tend to focus solely upon wins and losses, and thereby ignore each game's margin of victory (MOV) by essentially adjusting said score differentials to be at most a single point. The Bowl Championship Series (BCS) modified its approach in this fashion in 2002, hoping that it would provide a more reliable strategy for determining which two teams had earned the privilege of playing in the final game of the season, where the winner would be declared the National Collegiate Athletic Association (NCAA) National Champion (in football).

Prior to this tactical change, some experts conjectured that every team had an incentive to run up a large MOV (especially against weaker teams) in an effort to influence their own BCS ranking. This belief comes from the fact that achieving a large MOV would make a larger impact on a team's rating (which ultimately drives even most ranking techniques) than a close game over a quality opponent, as each game is usually weighted equally in the computer rating systems. And, since the rankings determined by the six officially recognized, computer-based systems contribute to the final BCS rating for each team, it was deemed a worthwhile endeavor to remove the possibility for any level of motivation to be present which would promote the aforementioned unsportsmanlike behavior. The other portion of the final BCS rating is derived from the two sets of human experts whose votes comprise the Harris sportswriters' poll and USA Today/ESPN coaches' poll. The BCS formula is used to select the two best Division I college football teams each year, after every game has been played, where: the highest and lowest computer system ranking per team are ignored, and the other four system's ranks are essentially a four person poll. Each of those three components add in a percentage to the final BCS rating, where 1.0 represents every voter - or system - designating that that team was ranked first; each of these voting total percentages are reduced by a factor of 3, and then all three components are summed to produce each team's final BCS rating, which ranges from 0 to 1.

By enforcing a MOV upper bound of one point, a significant amount of information is being ignored. This type of information loss has been shown by Berry (2003) to reduce a rating system's ability to deduce a team's 'strength', thereby reducing that system's ability to accurately forecast future outcomes. However, the original idea behind what comes next was motivated by the desire to focus on (and use) games whose scores are reasonably close, thereby (hopefully) ignoring all games with large differentials.

3. Using Graphs and Trees

Many athletic conferences or sports leagues utilize round robin scheduling, where every team plays a home and away game against every other member of that group. (Because some sport's seasons are too short, they can't accommodate the full round robin schedule. Therefore, teams alternate who they host each year and who they travel to, so that every team still plays all the other teams - just once each year.) If the teams are visualized as nodes in a graph, and edges connect teams that have played each other, then this round robin format produces a completely connected graph.

Most sports complement such inter-conference scheduling with games where teams from different conferences compete, and the resultant graph representations would contain dense clusters of connectivity (i.e. the conference schedules) with a typically much smaller number of connections joining some conference clusters to others. These edges are usually fewer in number, though more edges may connect 'rival conferences'. If the edges are directed, and the non-negative score differential of the game is the assigned weight for that edge, that begins with the losing team and ends at the victor, then it may be possible to deduce a team's relative strength

using the path from team A to team C. For instance, if team B defeats A by 10 points, and B also lost to C by 7 points, then the sum of the weights from A to C would be 17, and perhaps C is 17 points 'better' than A. Unfortunately, such graphs will invariably contain 'inconsistencies', i.e. cycles in the graph which violate the transitive law, e.g. B defeats A by 10, and loses by 7 to C, but we then find out that A has outscored C by 20! Without these inconsistencies, we would have a partial ordering imposed within the graph, and relative team strengths could be deduced from all the information contained in such an acyclic graph.

Since we can expect such graphs to contain cycles after a reasonable number of games have been played, the first idea was to create a minimum spanning tree (MST) from the graph representing all the scores for any particular season. By definition, a tree is a directed acyclic graph, and a MST is a connected graph where the total of all weights assigned to the included edges is the smallest possible sum. A greedy algorithm can quickly find a MST, but it may not be unique, though its sum would be. Because many game differentials are the same, the game scores were sorted in a special manner, before the greedy technique was employed, to guarantee a unique MST.

To create the desired MST, all edges are first sorted into ascending order by their weight. To break any ties that occur during this sorting procedure, edges whose two connected team's winning percentages are closest will appear earlier in this ordered list, than those whose percentages are more disparate. (If the percentages are also identical, then the team's points scored and allowed averages are used to break this secondary tie; the closer the differences in the two team's averages, the earlier in the list they will appear - for teams with the same winning percentages, on these identically weighted edges.) These two tiebreaking rules were chosen because it seemed preferable to give priority to the edges where the teams were apparently more similar, according to their records (and averages), than those identically weighted edges that were possibly more 'anomalous'.

After the ordered list of edges has been created, the greedy algorithm would simply continue to build a MST from the list by only adding those edges that do not create a cycle in the graph being constructed. With N teams, $N-1$ edges must be added to create a connected, acyclic graph. Some edges that are included in the MST may contain relatively large weights, when compared to most of the other edges in the MST, but they should typically still be smaller than the largest weighted edges in the full graph. (The selection of those last few edges is heavily constrained, because of all the previously added edges, as the algorithm attempts to finish connecting all the nodes within the MST without creating a cycle.)

Once the MST has been constructed, we can traverse the edges in the tree one at a time, to determine every team's rating. Assuming transitivity, if the sum of the weights on the path from team D to team F is X, then we can deduce that F is X points better than D. The information deduced in this manner can be chained so that each team's rating is some amount that is above or below another team's rating. Finally, these 'relative ratings' can be sorted, and a ranking will have been created from the MST. (Unfortunately, one or more identical ratings will probably exist, given the nature of such an integer-based strategy, so those teams will receive the same rank.)

Applying this technique to the 2010 NCAA football season (before bowl games) produced Mississippi as the top rated team at +11, with UNLV as the worst at -15. This came about because Auburn defeated Kentucky by 3, and Mississippi defeated Kentucky by 7, and Kentucky had already been assigned a rating of +4. As it turns out, Mississippi had a record of 4 wins and 8 losses, where Auburn was 12-0, and six other teams were also assigned a rating of +7 (besides Auburn), with Minnesota (3-9) and Northwestern (7-5) being two of them, and there were only five teams whose rating was larger than +7.

As one can readily observe, this strategy did not successfully evaluate the teams very well at all. Several other MSTs were created by first marking those edges in the 'original' MST as used, and then generating another 'MST'. For 2010, this strategy could be performed three times before all of one team's edges were included in one of the four 'MSTs' generated. All three subsequent rankings generated in this manner were as disappointing as the ranking produced by the original MST, and so was the ranking generated from averaging all four ratings (for each team).

4. Adding Non-determinism

Because the average rating computed as just described seemed to be 'better', the logical next step was to move away from producing a unique MST to the generation of many random spanning trees (ST); perhaps averaging

the deduced rating from a large number of STs would approximate the 'true rating' for each team more accurately. Therefore, the collection of game scores would be randomly shuffled, and then a ST would be created. To ensure that no particular game was included in too many more STs than any other one, the games are sorted into ascending order according to the number of times each game/edge has been previously incorporated into a ST, after the creation of every fourth ST. (Ties are broken in the same two ways as previously described.) Next, it must be determined how many iterations are needed to produce an accurate set of ratings.

Given that each of the N teams could be connected to any of the other N-1 teams, it seemed like the program should create at least N^2 STs before beginning to examine if convergence has occurred. (Because each NCAA football team plays roughly a dozen games each season, this value was increased by a factor of 6, since games will appear in two team's schedules, and 12/2 = 6, and, 6 is the first perfect number!) Once the minimum number of STs have been generated, the order in which the teams appear, after their (averaged) rating is computed for all previous random samples, will be stored to compare against the team orderings that are calculated after subsequent random STs are produced. So, the next question to be answered is: how many, consecutive iterations should produce the exact same order before convergence is acknowledged? Ten iterations seemed too few, and one hundred seemed too many, especially if there are several hundred teams involved. Therefore, a compromise was reached, and 28 (the second perfect number!) was chosen. The algorithm will stop iterating if convergence has not been reached after N^3 random STs have been generated, and if that limit is reached, that final set of average ratings will be output. (About 55% of the time that the algorithm was invoked, for the data sets from 2002 to 2010, additional iterations were needed to determine if convergence had occurred; only 48.5 more iterations were used on average in those cases.)

5. Results using Variations on the Actual Margin of Victory

The weight assigned to each edge had five possible values in this study: the actual score differential, the $\sqrt{differential}$, $log_2(differential + 1)$, the minimum of the score differential and 17, and finally, zero or one. (These will now be referred to as: FULL, SQRT, LOG2, MIN17, and NO_MOV.) All these performed reasonably well (except NO_MOV) when predicting the 279 bowl games from 2002 to 2010, ranging from 170 correct for FULL down to 160 for MIN17 (and 154 for NO_MOV, where the differential was at most one point). These four totals (including 169 for SQRT and 163 for LOG2) are higher than most of the systems that have published their ratings at the end of every one of those seasons on Dr. Kenneth Massey's web site (http://www.masseyratings.com/cf/compare.htm). As a matter of fact, FULL had 25 correct in 2010, which is at least 2 more than every one of the 129 systems listed on Massey's site that year (except for one that correctly predicted 26). Bowl games are usually played at a neutral site, which allowed for the home field advantage prediction component to be ignored, since it is unclear how large that should be for each of the five ST variations described here.

There have been 40 inventors that have voluntarily contributed their system's ratings to the Massey comparison page each of the past nine years. Including the average ranking of all systems listed on that page, 20 of these 41 strategies correctly predicted between 153 and 157 winners from 2002 to 2010. (Only one system was lower than that range: 148.) The other 20 totaled from 158 to 174 correct, where only four had more than 167 predictions right.

A previous study by Trono (2010) also considered these public systems as well as several other strategies; the Power Rating system (PRS) using the actual score, as described in that study, was one of the most accurate systems. The PRS correctly predicted 169 winners (from 2002-2010), and so it appears that at least FULL and SQRT are quite competitive with the best known systems. Therefore, the conjecture put forth by Berry (2003) was once again confirmed by this experiment, since the variations that limited how much of the score differential could be used performed worse than when using the actual score, and the larger the score collapse, the lower the number of bowl games that were predicted correctly.

The spanning trees that were created using a MOV capped at 1 point (i.e. NO_MOV) did match closely the components that comprise the BCS system in eight of the nine past years. Excluding 2007 for now, the top ten teams listed by NO_MOV were in exactly the same positions, as 'voted' by the six sanctioned computer ranking systems, 31 of 80 times, and said positions were only off by one for 19 more teams. The average difference in placement was 1.825, and only thrice was that difference greater than five. With regards to the

final, overall BCS ranking, NO_MOV was the same 19 times, and it was off by one 16 more times, with an average difference roughly equal to 2.78; three times that difference was eight or larger. Six of these nine years, NO_MOV agreed with which team was designated #1 by the BCS formula, but the three other years (2005-07) NO_MOV ranked that team as #3, #3, and #5 respectively.

Those three years are also interesting because if NO_MOV was used to predict the winner of the BCS National Championship bowl game, it correctly did so all three times, where the average BCS computer ranking and the overall BCS rankings were wrong for all three games! Overall, NO_MOV correctly chose the winner of that championship game six of the last nine years, where the Las Vegas betting line - and the sportswriters' final poll - only did so five times, and the coaches' poll, like the two aforementioned BCS rankings, only had three out of those nine correct.

Regarding 2007, it is hard to tell 'what when wrong', as two of NO_MOV's top ten teams only appeared between #20 and #25 in the polls, and three more weren't even ranked! South Florida was listed as NO_MOV's #2 team, and they were listed as #14 by the BCS computers, and Cincinnati was NO_MOV's #10 team and were ranked in a tie for #20 (with Connecticut), but South Carolina and California both had records of 6-6 and were listed as #40 and #49 on the Massey web site, and that is the average ranking assigned to them, given where all those submitted systems placed them. (Needless to say, with a year like 2007, it would be hard to convince anyone that NO_MOV does truly have a gift for ranking teams.)

However, the PRS thought three of those five teams were fairly strong, assigning South Florida the eighth highest rating, Oregon (8-4 and #9 according to NO_MOV) as the tenth best team and Cincinnati (9-3, and NO_MOV's #10 team) as #12. South Carolina ended up #25 and California #36 according to the PRS, but these teams did have some evidence to justify their finishing as #8 and #9 in NO_MOV's ranking. For instance, South Florida defeated West Virginia (10-2, and #12 in NO_MOV) and Auburn (8-4, and #16 in NO_MOV), Oregon had wins over two 10-2 teams (USC and Arizona State, the latter was #4 in NO_MOV) and a 9-3 Michigan team, while Cincinnati defeated South Florida and Oregon State (8-4, and victorious over Oregon) and 9-3 Connecticut (who also defeated South Florida). Meanwhile, California beat Oregon and Tennessee (9-4) while South Carolina defeated Georgia (10-2) and Kentucky, the latter defeated National Champion LSU (11-2 and #3 according to NO_MOV).

Given that the weights in these random STs are all 1, the edges mentioned above could place the winning team close to the top of such randomly constructed trees, where the highly rated teams would reside, and so these teams would be recognized as having earned relatively high ratings as well. (More seasons will have to be investigated to determine if 2007 is the only such 'anomalous year'; perhaps it might possible to generate all STs to determine each team's rating, instead of just taking a large random sample, but that has not yet been attempted.)

Several other graph-based systems have appeared in the literature - Callaghan *et al.* (2004), Park and Newman (2005) and Colley (2002) - but these systems were designed to rank teams, and to solve the BCS conundrum, not to determine the relative strengths of the teams. The descriptions for each system on Massey's site can be found by following the link to that system's home page (that is embedded within Massey's web page); many of these home pages refer to said systems not producing reliable results until a fair number of games have been played. Several of these explanations refer to the necessity of a signficant number of games having been completed for the 'required connectedness' to have occurred, so perhaps many of these systems may also rely upon graphical representations as their underlying structure as well when determining ratings.

6. Summary

The ratings for athletic teams can be estimated by treating the game results for that year as a graph, where a large number of spanning trees are randomly extracted from that graph, and the relative strengths for all teams are calculated as the average over all those random spanning trees. Several different functions were evaluated, regarding the weights assigned to each edge - corresponding to each game's final score - and the best performance observed was associated with the method which did not alter the score in any way. The prediction accuracy results reported here compare favorably to those systems described and evaluated in Trono (2010).

When margin of victory was removed, this strategy matched the components of the BCS formula reasonably

well, but it also correctly predicted the winner of the NCAA National Championship game three more times than the overall BCS formula, its computer average ranking component and the coaches' poll, over the past nine years, and, this strategy also had one more correct prediction than either the posted Las Vegas betting line, or the sportswriters' poll. (However, this strategy's overall prediction performance was worse than all other spanning tree variations in this time period - for all bowl games.)

References

Berry S. (2003) College football rankings: The BCS and the CLT. *Chance* 16, 46–49.

Callaghan T., Mucha P.J. and Porter M.A. (2004) The bowl championship series: A mathematical review. *Notices of the American Mathematical Society* 51, 887–893.

Colley W.N. (2002) Colley's bias free college football rating method: The Colley matrix explained, www.colleyrankings.com/matrate.pdf.

Park J. and Newman M.E.J. (2005) A network-based ranking system for US college football. *Journal of Statistical Mechanics: Theory and Experiment* 10, 14 pages.

Trono J.A. (2010) Rating/ranking systems, post-season bowl games, and 'the spread'. *Journal of Quantitative Analysis in Sports* 6, Issue 3, Article 6.