Rankings, Done Justly (9/21/2014)

J. Patrick Rhamey, Jr., PhD

Background

The continued problem of determining a clear winner has driven the evolution of the college football post-season since the first Rose Bowl between Michigan and Stanford. The two teams, widely perceived to be the most excellent in their respective regions, faced one another in the first contest vaguely resembling a national championship. The excitement of the two teams facing one another was quickly quieted as Michigan crushed Stanford 49-0. Clearly, Stanford was not quite as great as everyone had thought. This continued desire to name a college football team the greatest in the country (as well as determine who numbers 2-25 are) spurred in 1934 the development of the Associated Press poll, with a simple proposition: whichever team ranked first in the poll would be crowned the national champion. By 1950 the Coaches Poll had joined the Associated Press in this activity, providing two competing subjective assessments of college football teams that, on eleven different occasions, have disagreed.

Because the polling method was based on subjective opinions rather than meaningful evidence (not unlike the current playoff selection committee), a national championship game was created in 1992, culminating in 1998 with the BCS which lasted through the 2013 season. The BCS sought to provide a solution to the continuous drama of naming a champion that had plagued football. By adding (mostly) mathematically rigorous computer rankings, many of which are not fully transparent in their methods, the BCS intended to address the subjectivity of the existing ranking system. However, the computer rankings suffered from both a lack of public information on how most of these methods are used, and more importantly, the limited amount of information within a football season immediately calls into question the appropriateness of any statistical model. The average college football fan has a vague sense that "the computers" incorporate wins, losses, scores, offensive yards, defensive yards allowed, and various other statistics. However, each team only has a handful of games on which to generate these mathematical predictions, making such predictions immediately dubious. Furthermore, the dual ranking problems of subjectivity and arbitrariness remain in the selection of *which* statistics matter (which varies across computer rankings), and college football fans are left dissatisfied. This dissatisfaction has led to a plus 1 system of the top four teams, but this new system will only continue perceptions of subjectivity and favoritism, as debates will no longer be over two teams, but four. Gone are the polls and computer rankings in determining those four, replaced with the whims of a selection committee whose composition no one entirely understands and is most certainly subject to bias (ahem Pat Haden).

The Problem

The college football fan doesn't want speculation. They don't want Craig James telling them who is, or is not, worthy to play in a national championship any more than the average Texas Tech fan

wants Craig James telling them who should be their football coach. They don't want Pat Haden's opinion on who deserves to go to a playoff. They don't want football coaches tweaking the Coaches Poll in their interest with their votes. Furthermore, they don't want a computer, using a method that is not even public much less understandable or statistically appropriate, telling them who is and is not the champion of college football. While we've done away with the BCS computers in the plus one, now we have an arbitrarily assigned selection committee mandating a top four teams: a committee composed of individuals with their own interests and machinations no different than the schools that they represent. While we have come a long way from the Rose Bowl game of 1901, 113 years and somehow we still can't get this straight.

The problem is simple: every team in college football can't play every other team. If they could, determining the winner would be simply a matter of tallying up wins and losses, and there would be no debate on which team is the greatest. However, with teams only playing a small sampling of possible opponents, we are left attempting to determine the greatest team on limited information. This leads us to two principles for developing a fair and just ranking:

1) The only information relevant to a poll is how a football team has performed – speculation on how a team will do, or might do, (often referred to as the "look test") has no place in college football.
2) The ranking system must be entirely transparent, every detail available to the public, and easy to understand.

The first principle of speculation is typical of media personalities and the AP poll. Often they base their subject valuations of who is or is not the best team based on how they "looked". However, without any reference to the quality of an opponent, the look test is fairly irrelevant. Alabama "looking good" against the Western Carolina Catamounts is a fairly irrelevant observation, but that doesn't stop Kirk Herbstreidt from deeming it relevant evidence in his ranking of football teams.

The computers, however, provided us with a way out of the Sunday morning speculation of the pollsters. But despite their potential for less bias, more anger was often directed at the computer rankings than the polls. There are two clear problems with the computer rankings that defy our second principle: they are often secretive and always complex. If those that are secretive were not, the underlying statistical methodologies would likely make them inaccessible to the average football fan, further breeding existing dissatisfaction. In addition to problems of inaccessibility, the rankings are unnecessarily complicated including a number of both mathematical and theoretical assumptions in their creation with which most any fan would disagree. For example, many operate off the false assumption that we can compare certain detailed statistics, like the score or total offense, across games, and thereby determine the best teams. Any thoughtful fan knows this to be absurd. Comparing total offense from a game in the snow in Ann Arbor to 75 degrees and sunny in Los Angeles is so beyond incomparable as to be entirely irrelevant. Notre Dame's games are a good example of incomparability. Notre Dame narrowly defeated Pitt in 2012. Notre Dame also defeated Oklahoma (not so narrowly). Does that mean that Pitt is as good or better than Oklahoma? Of course not. You cannot compare the specific statistics across games. The propensity of certain coaches to run up the score on certain teams,

due to past team histories, rivalries, or the dislike of the opposing coach, can alone make comparing the score across games in similar conditions unreliable (i.e. Spurrier playing Georgia). The position of teams in a ranking shouldn't be a reflection of weather or questionable sportsmanship. From game to game in college football, there is simply too much of what statisticians call "noise" – all those intervening factors that make each game, and the statistics that game produces, entirely unique. All the computer rankings do is highlight the original problem of rankings – limited amounts of information. They attempt to organize this information coherently, but instead, due to the complexity of the game, end up reinforcing existing problems of arbitrariness using statistical model inappropriately given a very limited number of observations.

While the selection committee does away with this arbitrariness, we are back to where we were before the BCS and remain faced with the problem of speculation and opinion, now coupled with a disconcerting potential lack of transparency. While no one will likely contest that the winner of four team championship will be the best team *of those four teams*, there will always be a team number five angrily protesting their exclusion on subjective grounds. The problems we've faced over the past century remain intact, and given the possibility of a dark horse candidate winning a championship the playoff engenders, debates are likely to only become more heated with our new format.

The Solution

At first, it would seem that our difficulties are insurmountable. The game is too complex and our information too limited, leaving us to the original method of 1901 - the speculation of experts. However, the solution to the problem is also profoundly simple. Every college football team does not play every other team, *but* they are all interconnected by a web of wins and losses. Alabama plays Mississippi State, Mississippi State plays Troy, Troy plays Florida International, Florida International plays Duke. Based on the outcomes of these games, we can generate a simple ranking of which team is best, even though all our teams do not play one another. However, as with the problem in the current computer rankings, we must determine what information from these games is relevant. Because of the problem of comparing detailed statistics across different settings from game to game, we must limit ourselves to the simple principle that a win is a win, not because there is not important descriptive information in *how* a team wins, but because there are so many other intervening factors. Continuing our example, let's say Alabama beats Mississippi State. Alabama gets a win. Then we can adjust the value of that win based on how Mississippi State performs. If Mississippi State wins the rest of its game, than that win by Alabama is worth a lot. If, however Mississippi State loses a number of games, the value of Alabama's win will deteriorate, particularly if Mississippi State loses to teams with less than stellar win-loss records.

Of the existing methods available, the logic (albeit not the method) is actually fairly similar to the Colley Matrix, one of the 7 previous BCS computer methods. The Colley Matrix is perhaps the best existing ranking beyond the one presented below because it is based off this simple assumption of ranking on wins and losses and is open and transparent in its method (Colley provides the details of his

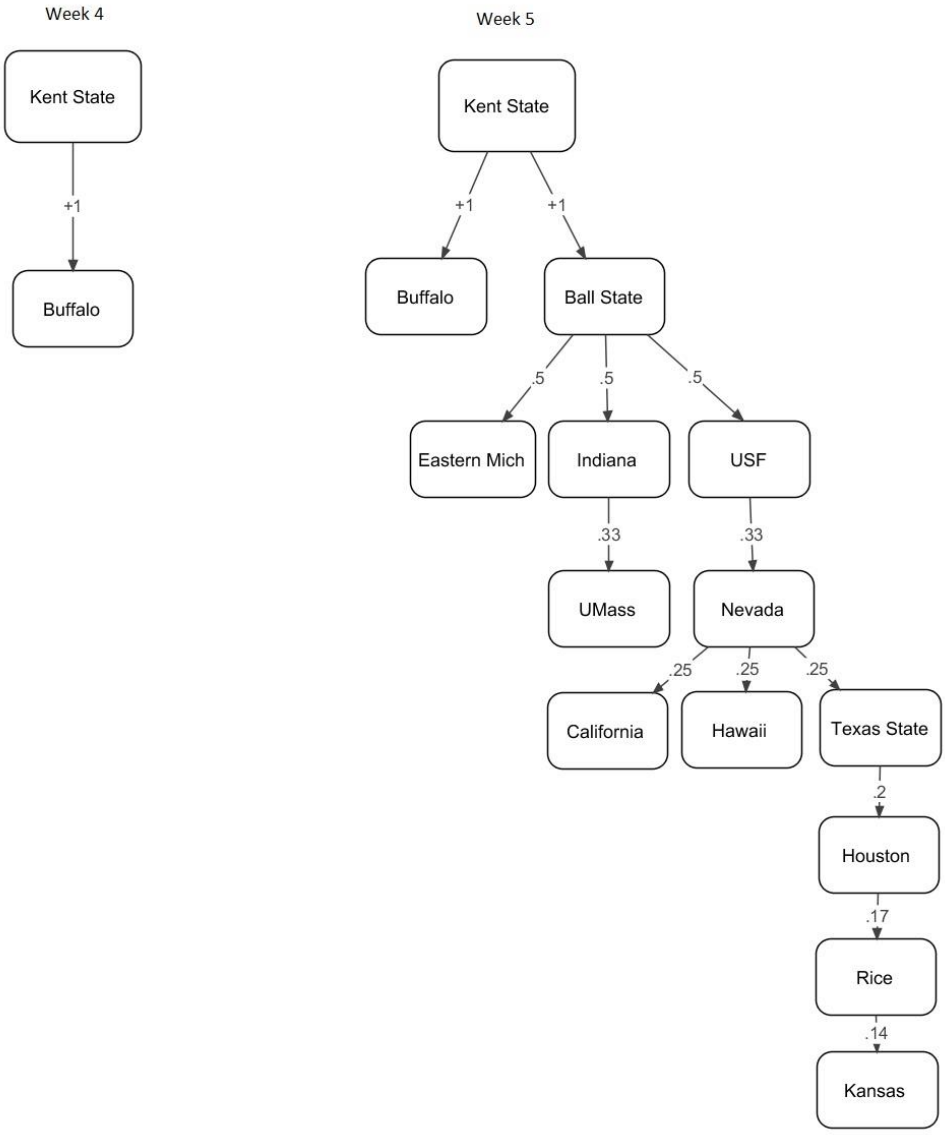
ranking here). However, the Colley Matrix uses a complex iterated set of equations with a few rather arbitrary mathematical assumptions rather than simply looking at the network of wins and losses. A simplistic way to explain the difference between the win-loss method proposed below and Colley's method would be that, when a team A beats team B that beats team C, the win-loss method proposed here adds up the distance across each of those wins (from A to C) to give a win score for team A. The Colley Method uses a system of equations treating these wins and losses as equal, using them to come up with a strength of schedule estimate, and then running the equation over again to calculate a predicted value based on wins and losses adjusted by opponents rating (in other words, as Colley describes it, "the probability of goodness" extrapolated off a team's win and loss record).

All of Colley's extra steps, however, aren't necessary (and are in fact problematic because he is asking an enormous amount from very little data). We can start by drawing lines between teams that have played each other. As of this week, every FBS team is connected in some way by having played at least one other FBS team. We can then add up the distances between each team and every other team in the FBS following the path of wins, or tracing the path of who beats who beats who. It's like a game of six degrees of separation where we're trying to figure out which team is Kevin Bacon. We figure out which team is best by adding up the degrees of separation between a team and other teams in the FBS. The fewer steps it takes to reach more teams will give us a higher win score. Likewise, we can do the same by adding up the steps connecting teams by losses. For example, if Alabama defeats 10 teams, and each of those teams beat 9 teams, Alabama is connected to 100 teams within two degrees of separation and would be ranked very highly. If Alabama beats 10 teams, but then those 9 teams lose all their games, Alabama is only connected to 10 teams, and would not be ranked highly. In each example, Alabama has the same record, but the wins are valued differently based on the performance of the defeated teams. Through this degrees of separation method (called "average reciprocal distance") we can add up the value of each team's wins and losses, and therefore their rank based on their total body of work. Because adding up all the wins and losses across 128 interconnected football teams is hard, or impossible, to do by hand, we use a computer.[1] But, this is not a computer ranking. The computer is just a big calculator adding up the wins and losses, unlike the BCS rankings which use selected statistics to estimate a relative assignment of team rank, or a prediction of their "goodness".

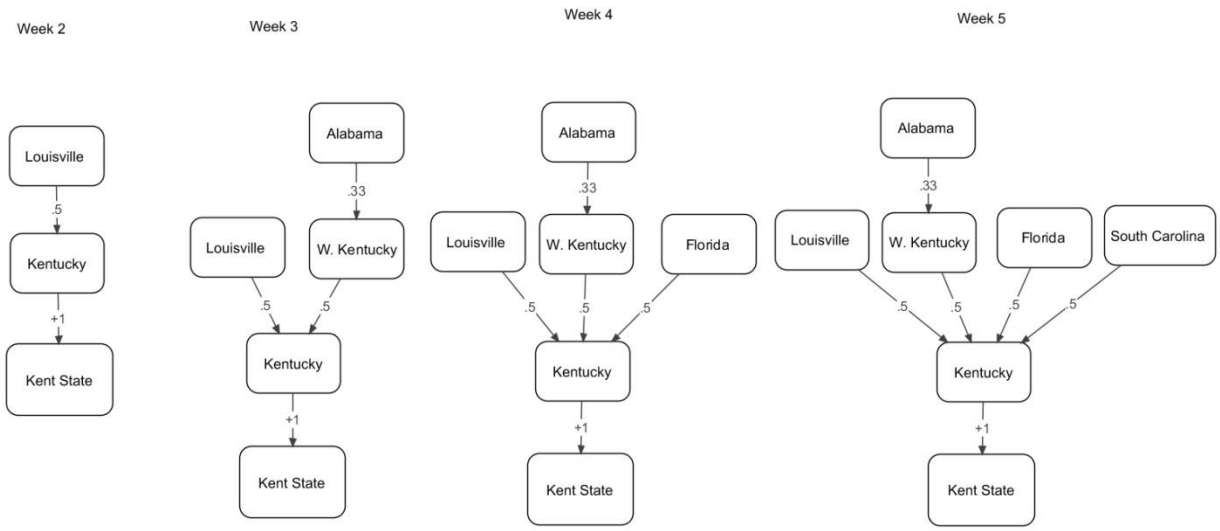
The simplified illustration below provides an example of how we can come up with a just and fair ranking after the first five weeks for the mighty Kent State Golden Flashes in their 2012 season. Adding up wins first, week 1 is against an FCS opponent, so it is ignored. Week two the Golden Flashes lose to Kentucky, and week 3 they have a bye, so after the first three weeks Kent State has a win score of 0. Finally in week 4 the Golden Flashes get their first win against an FBS opponent with Buffalo. Kent State get +1 for defeating Buffalo, and now has a win score of 1. Week 5 the Golden Flashes defeat Ball State, which adds another +1 for a win, but then partial credit for each team Ball State defeated (.5 for EMU, Indiana, USF) and then additional partial credit for each team in the chain of wins all the way to Kansas, where Kent State gets credit for the win divided by the number of degrees of separation. By doing this, the win over Buffalo remains "1", as Buffalo has defeated no FBS opponent, but the win over

[1] The program used to calculate these values is called UCI, using the measure of centrality for each team through average reciprocal distance.

Ball State is worth “4.42” (adding up credit for all the teams Ball State beat, the teams those teams beat, etc.). At the end of week four, Kent State now has a win score of 5.42.



We then need to figure out the cost of Kent State’s loss to Kentucky in Week 2. When the Golden Flashes lose to Kentucky, Kentucky already has a loss of its own to undefeated Louisville. As a result, Kent State now has a loss score of 1.5 (1 for Kentucky, .5 for Kentucky losing to Louisville). However, as Kentucky continues to lose in weeks 3, 4, and 5, that loss becomes more costly for Kent State. It grows from 1.5 in week 2, to 2.33 in week 3 when Kentucky loses to Western Kentucky (who lost to Alabama), to 2.88 in week 4 when Kentucky loses to Florida, to 3.38 by Week 5. Subtracting their loss score (3.38) from their win score (5.42), Kent State at the end of Week 5 has an overall score of 2.04. Adding up wins and losses for every team in the nation by this process gives us an overall score for each team and their overall rank.



Simply put, we can create a ranking in the most intuitively obvious way possible - based on who beats who. No subjective value judgments about how good of a win it was, what that win means for future games, or how well your Heisman candidate happened to perform. No longer are teams allowed to ride high in the polls based on their pre-season speculation. All teams begin on even ground, and only their success, or failure, provides a means of advancement. In other words, you have to work to be number 1; no one is going to hand it to you.

The implications of the win-loss method are both very little, and very great, depending upon the school. Going back to previous seasons, using the win-loss method doesn't appear to have a dramatic impact on the selection of the top two teams. Alabama and LSU are still the best teams of the 2011 season. Auburn is still the number 3 team going into the bowl season in 2004. It turns out that by the end of the season, the pollsters and the BCS often get number 1 and 2 right (except of course when they disagree). While the method presented here is more intuitive, simple, and reasonable than existing methods, it doesn't necessarily dramatically affect the big outcome. However, there is more to the college post-season than assigning number 1, and with the new method of four teams being selected proper ranking becomes even more important. Furthermore, bowl committees remain heavily influenced by rankings when determining at-large bids, and invitation to a bowl can mean millions in revenue for the participants.

Now that we live in a world where the top four teams rather than two are what truly matter, having an accurate, non-arbitrary means of determining those top four should be at the forefront of college football. A selection committee, however, is an indication that athletic competition and excellence are backseat to revenue in college football. While the win-loss method designates the same top two as the AP in the 2013, 2011 and 2004 seasons, it has different suggestions for the third and fourth seeds. In 2013, the win-loss method would have chosen Mizzou and Alabama to join Auburn and Florida State in a playoff, the AP would have chosen Alabama and Michigan State. In 2011, the BCS, AP, and Coaches Polls would have sent Oklahoma State and Stanford, but the win-loss method would have sent Arkansas and Kansas State. While all methods agree the third seed in 2004 is Auburn, the AP and

Coaches polls would have sent one loss California, overlooking undefeated Utah, the win-loss method fourth seed. Overlooking Utah may not have mattered to the cause of crowning a final champion, but it certainly matters to Utah in terms of prestige, financial gain, and rewarding the hard work and accomplishments of their players, left out by no fault of their own. Lower profile teams, from lower profile conferences are likely to be left out of the minds of selection committee members, just as AP voters inflate high profile teams like USC and Michigan. A similar criticism could be lodged in defense of Arkansas if there had been four team playoff in 2011 or Mizzou in 2013: if three of the four best teams in college football are in the same conference, I am confident the third team would be overlooked by the selection committee under political pressure from other conferences.

There are other benefits to the win-loss method beyond its simplicity and objectivity. For starters, like the computer methods, it's fun for the football fan to have a full 128 team ranking of college football teams to see how your favored team is fairing, but unlike the computer methods, it's simple and easy to understand. But all these small gains take a back seat to defending the meaningful work of the players. It's time to end the arbitrariness and subjectivity in college football, and put the accomplishments of the athletes on the field ahead of the posturing of media, school administrators, and conference officials. It's such a simple idea, but a little simplicity is what college football needs.