Colley's Bias Free College Football Ranking Method:
The Colley Matrix Explained

Wesley N. Colley

Ph.D., Princeton University

ABSTRACT

Colley's matrix method for ranking college football teams is explained in detail, with many examples and explicit derivations. The method is based on very simple statistical principles, and uses only Div. I-A wins and losses as input — margin of victory does not matter. The scheme adjusts effectively for strength of schedule, in a way that is free of bias toward conference, tradition, or region. Comparison of rankings produced by this method to those produced by the press polls shows that despite its simplicity, the scheme produces common sense results.

Subject headings: methods: mathematical

1. Introduction

The problem of ranking college football teams has a long and intriguing history. For decades, the national championship in college football, and/or the opportunity to play for that championship, has been determined not by playoff, but by rankings. Until recently, those rankings had been strictly the accumulated wisdom of opining press-writers and coaches. Embarrassment occurred when, as in 1990, the principal polls disagreed, and selected different national champions.

A large part of the problem was the conference alignments of the bowl games, where championships were determined. Michigan, for instance, won a national championship in 1997 by playing a team in the Rose Bowl that was not even in the top 5, because the Big 10 champ always played the Pac-10 champ in the Rose Bowl, regardless of national ramifications.

In reaction to a growing demand for a more reliable national championship, the NCAA set up in 1998 the *Bowl Championship Series* (BCS), consisting of an alliance among the Sugar Bowl, the Orange Bowl and the Fiesta Bowl, one of which would *always* pit the #1 and #2 teams in the country against each other to play for the national title (unless one or more were a Big 10 or Pac-10 participant). The question was, how to guarantee best that the true #1 and #2 teams were selected...

By the time of the formation of the BCS (and even long before) many had begun to ask the question, can a machine rank the teams more correctly than the pollsters, with less bias than the humans might have? With the advent of easily accessible computer power in the 1990's, many "computer polls" had emerged, and some even appeared in publications as esteemed as the *New York Times*, *USA Today*, and the *Seattle Times*. In fact, by 1998, many of these computer rankings had matured to the point of some reliability and trustworthiness in the eyes of the public.

As such, the BCS included computer rankings as a part of the ranking that would ultimately determine who played for the title each year. Several computer rankings would be averaged together, and that average would be averaged with the "human" polls (with some other factors) to form the best possible ranking of the teams, and hence determine their eligibility to play in the three BCS bowl games. The somewhat controversial method, despite some implausible circumstances, has worked brilliantly in producing 4 undisputed national champions. With the addition of the Rose Bowl (and its Big 10/Pac-10 alliances) in 2000, the likelihood of a split title seems very small.

Given the importance of the computer rankings in determining the national title game, one must consider the simple question, "Are the computers getting it right?" Fans have doubted occasionally when the computer rankings have seemed to favor local teams, disagreed with one another, or simply disagreed with the party line bandied about by pundits.

Making matters worse is that many of the computer ranking systems have appeared to be byzantine "black boxes," with elaborate details and machinery, but insufficient description to be reproduced. For instance, many of the computer methods have claimed to include a home/away bonus, or "conference strength," or a particular weight to opponent's winning percentage, etc., but without a complete, detailed description, we're left just to trust that all that information is being distilled in some correct way.

With no means of thoroughly understanding or verifying the computer rankings, fans have had little reason to reconsider their doubts. A critical feature, therefore, of the Colley Matrix method is that this paper formally defines *exactly* how the rankings are formed, and shows them to be explicitly bias-free. Fans may check the results during the season to verify that the method is truly without bias.

With luck, I will persuade the reader that the Colley Matrix method:

1. has no bias toward conference, tradition, history, etc.,
   (and, hence, has no pre-season poll),
2. is reproducible,
3. uses a minimum of assumptions,
4. uses no *ad hoc* adjustments,
5. nonetheless adjusts for strength of schedule,
6. ignores runaway scores, and
7. produces common sense results.

2. Wins and Losses Only—Keep it Simple

The most important and most fundamental question when designing a computer ranking system is simply where to start. Usually, in science, one poses a hypothesis and checks it against observation to determine its validity, but in the case of ranking college football teams, there really is no observation—there is no ranking that is an absolute truth, against which to check.

As such, one must form the hypothesis (ranking method), and check it against other rankings systems, such as the press polls, other computer rankings, and, perhaps even common sense, and make sure it seems to be doing something right.

Despite the treachery of checking a scientific method against opinion, we proceed, first by contemplating a methodology. The immediate question becomes what input data to use.

Scores are a good start. One may use score differentials, or score ratios, for instance. One may even invent ways of collapsing runaway scores with mathematical functions like taking the arc-tangent of score ratios, or subtracting the square roots of scores. I even experimented with a histogram equalization method for collapsing runaway scores (which, by the way, produced fairly sensible results).

However, even with considerable mathematical skulduggery, reliance on scores generates some dependence on score margin that surfaces in the rankings at some level. Rightly or wrongly, this dependence has induced teams to curry favor in computer rankings by running up the score against lesser opponents. The situation had degraded to the point in 2001 that the BCS committee instructed its computer rankers either to eliminate score dependence altogether or limit score margins to 21 in their codes.

This is a philosophy I applaud, because using wins and losses only

1. eliminates any bias toward conference, history or tradition,
2. eliminates the need to invoke some *ad hoc* means of deflating runaway scores, and
3. eliminates any other *ad hoc* adjustments, such as home/away tweaks.

By focusing on wins and losses only, we're nearly halfway to accomplishing our goals set out in the Introduction.

A very reasonable question may then be, why can't one just use winning percentages, as do the NFL, NBA, NHL and Major League, to determine standings? The answer is simply that in all those cases, each team plays a very representative fraction of the entire league (more games, fewer teams). In college football, with 117 teams and only 11 games each, there is no way for all teams to play a remotely representative sample. The situation demands some attention to "strength of schedule," and it is herein that lies most of the complication and controversy with the college football computer rankings.

The motivation of the Colley Matrix Method, is, therefore, to use something as closely akin to winning percentage as possible, but that nonetheless corrects efficiently for strength of schedule. The following sections describe exactly how this is accomplished.

3. The Basic Statistic — Laplace's Method

Note to the reader: In the sections to follow, many mathematical equations will be presented. Many derivations and examples will be based upon principles of probability, integral calculus, and linear algebra. Readers comfortable with those subjects should have no problem with the level of the material.

In forming a rating method based only on wins and losses, the most obvious thing to do is to start with simple winning percentages, the choice of the NFL, NBA, NHL and Major League. But simple winning percentages have some incumbent mathematical nuisances. If nothing else, the fact that a team that hasn't played yet has an undefined winning percentage is unsavory; also a 1-0 team has 100% vs. 0% for an 0-1 team: is the 1-0 team really infinitely better than the 0-1 team?

Therefore, instead of using simple winning percentage (n_w/n_{tot}, with obvious notation), I use a method attributable to the famed mathematician Pierre-Simon Laplace, a method introduced to me by my thesis advisor, Professor J. Richard Gott, III.

The adjustment to simple winning percentage is to add 1 in the numerator and 2 in the de-

nominator to form a new statistic,

$$r = \frac{1 + n_w}{2 + n_{tot}}. \tag{1}$$

All teams at the beginning of the season, when no games have been played, have an equal rating of $1/2$. After winning one game, a team has a $2/3$ rating, while a losing team has a $1/3$ rating, i.e., "twice as good," much more sensible than 100% and 0%, or "infinitely better."

The addition of the 1 and the 2 may seem arbitrary, but there is a precise reason for these numbers; namely, we are equating the win/loss rating problem to the problem of locating a marker on a craps table by trial and error shots of dice. What?

This craps table problem is precisely the one Laplace considered. Imagine a craps table (of unit width) with a marker somewhere on it. We cannot see the marker, but when we cast a die, we are told if our die landed to the left or right of the marker. Our task is to make a good guess as to where that marker is, based on the results of our throws. The analogy to football is that we must make a good guess as to a team's true rating based on wins and losses.

At first, our best guess is that the marker is in the middle, at $r = 1/2$. Mathematically, we are assuming a "flat" distribution, meaning that there is equal probability that the marker is anywhere on the table, since we have no information otherwise—that is to say, a uniform Bayesian prior. The average value within such a flat distribution (shown in Fig. 1 at top left) is $1/2$. Computing that explicitly is called finding the expectation value (or weighted mean, or center of mass). If the probability distribution function of rating $\hat{r}$ is $f(\hat{r})$, then in the case of no games played (no dice thrown), $f(\hat{r}) = 1$, and the expectation value of $\hat{r}$ is

$$r = \langle \hat{r} \rangle = \frac{\int_{r_0}^{r_1} \hat{r} \cdot f(\hat{r}) d\hat{r}}{\int_{r_0}^{r_1} f(\hat{r}) d\hat{r}} = \frac{\int_0^1 \hat{r} d\hat{r}}{\int_0^1 d\hat{r}} = \frac{(\hat{r}^2/2)|_0^1}{\hat{r}|_0^1} = 1/2. \tag{2}$$

Now, if the first die is cast to the left of the divider, the probability density that the marker is at the left wall ($\hat{r} = 0$) has to be zero — you can't throw a die to the left of the left wall. From zero at the left wall, the the probability density must increase to the right. That increase is just linear, because the probability density is just the available space to the left of the marker where your die could have landed; the farther you go to the right, the proportionally more available space there is to the left (see Fig. 1, top right).

The analogy with football here is clear. If you've beaten one team, you cannot be the worst team after one game, and the number of available teams to be worse than yours increases proportionally to your goodness, your rating $\hat{r}$.

The statistical expectation value of the location of the marker (rating of your team) for the one

left throw (one win) case is therefore

$$r = \frac{\int_0^1 \hat{r}^2 d\hat{r}}{\int_0^1 \hat{r} d\hat{r}} = \frac{(\hat{r}^3/3)|_0^1}{(\hat{r}^2/2)|_0^1} = 2/3. \tag{3}$$

If we throw another die to the left, we have not a linear behavior in probability, but parabolic, since the probability densities simply multiply,

$$r = \frac{\int_0^1 \hat{r}^3 d\hat{r}}{\int_0^1 \hat{r}^2 d\hat{r}} = \frac{(\hat{r}^4/4)|_0^1}{(\hat{r}^3/3)|_0^1} = 3/4, \tag{4}$$

as shown at the bottom left in Fig. 1.

However, when a die is thrown to the right, we know that the probability at the right wall has to go to zero, and a term growing linearly from right to left is introduced, $(1 - \hat{r})$ (in exact analogy to the left-thrown die). Therefore, if we have thrown one die to the left and one to the right, we have

$$r = \frac{\int_0^1 (1 - \hat{r})\hat{r}^2 d\hat{r}}{\int_0^1 (1 - \hat{r})\hat{r} d\hat{r}} = \frac{(\hat{r}^3/3 - \hat{r}^4/4)|_0^1}{(\hat{r}^2/2 - \hat{r}^3/3)|_0^1} = 1/2, \tag{5}$$

as shown at the bottom right in Fig. 1.

In general, for n_w wins (left throws of the die) and n_ℓ losses (right throws of the die), the formula is

$$r = \frac{\int_0^1 (1 - \hat{r})^{n_\ell} \hat{r}^{n_w} \hat{r} d\hat{r}}{\int_0^1 (1 - \hat{r})^{n_\ell} \hat{r}^{n_w} d\hat{r}} = \frac{1 + n_w}{2 + n_\ell + n_w}, \tag{6}$$

which recovers equation (1). It is an interesting exercise to check a few more examples.

4. Strength of Schedule

The simple statistic developed in the last section would suffice to produce a ranking if we were confident that all teams had played a schedule of similar strength, or for instance a round-robin tournament. While a round-robin with 117 teams would require 6786 games, Division I-A teams play typically a tenth that, so there is absolutely no assurance that the quality of opponents from team to team is close to the same. Contrast this with the NFL, or especially the Major League, where each team plays a very healthy sample of the entire league during the regular season.

This problem is complicated by the addition of still more teams in the form of non-I-A opponents. If one were to use those games as input, he would have to form ratings of all the I-AA teams, which would require ratings of teams in still lesser divisions, since many I-AA teams play

such opponents. Forming sensible ratings which relate Florida State to Emory & Henry is extremely difficult and is frankly beyond the scope of this method. The reason is that my method, in its simplicity, relies on some interconnectedness between opponents, which simply does not exist between a given NAIA squad and a given Division I-A squad—there's barely enough interconnectedness among the I-A teams themselves! Most other computer rankings within the BCS system do endeavor to compute such ratings, and in my opinion, do nearly as good a job as is possible at making sense of such disparate and competitively disconnected teams. To preserve simplicity and total objectivity (no *ad hoc* division adjustment, etc.), my rating system must ignore all games against non-I-A opponents. Therefore, *padding the schedule with I-AA teams contributes absolutely nothing to a team's rating.*

We may then proceed with mathematical adjustments for strength of schedule within Division I-A itself.

The number of wins in equation (1) may be divided into $n_{w,i} = (n_{w,i} - n_{\ell,i})/2 + n_{tot,i}/2$ (which the reader can check). Recognizing that the second term may be written as $\sum^{n_{tot,i}} 1/2$ allows one to identify the sum as that of the ratings of team i's opponents, if those opponents are all random ($r = 1/2$) teams. Instead, then, of using $r = 1/2$ for all opponents, we now use their actual ratings, which gives an obvious correction to $n_{w,i}$.

$$n_{w,i}^{eff} = (n_{w,i} - n_{\ell,i})/2 \; + \sum_{j=1}^{n_{tot,i}} r_j^i, \tag{7}$$

where r_j^i is the rating of the j^{th} opponent of team i. *The second term (the summation) in equation (7) is the adjustment for strength of schedule.*

Now, the rub. When the teams are not random, but ones which have played other teams, which may or may not have played some teams in common with the first team, etc., how does one possibly figure out simultaneously all the r_j^i's which are inputs to the r_i's, which are themselves r_j^i's for other r_i's, etc.?

5. The Iterative Scheme

The most obvious way to solve such a problem is a technique called "iteration," which is employed by several of the other computer ranking methods. The way it works is one first computes the ratings, as if all the opponents were random ($r = 1/2$) teams, using equation (1). Next, each team's strength of schedule is computed according to its opponents' ratings, using equation (7). The ratings are re-computed with the new schedule strengths, and then strengths of schedule are re-computed from the new ratings. With luck, the changes to the ratings get smaller and smaller

with each step of these calculations, and after a time, when the changes are negligibly small for any team's rating (a part in a million, say), one calls the list of ratings, at that point, final.

Here is a very simple example of the iterative technique, after only one week of play, where a team that has played is either 0-1 against a 1-0 team, or vice versa. Before any iterations, the basic Laplace statistic from equation (1) is computed for each team. Letting $r_{W,0}$ be the initial rating of a winning team, and $r_{L,0}$ be the initial rating of the losing team, one finds that equation (1) initially gives

$$\begin{array}{c}\text{Initial ratings}\\ r_{W,0} = (1+1)/(2+1) = 2/3 \approx 0.6667 \\ r_{L,0} = (1+0)/(2+1) = 1/3 \approx 0.3333.\end{array} \tag{8}$$

The first adjustment for strength of schedule (there's been only one game, so schedule strength is just the rating of the one opponent) is made by computing n_w^{eff} for each team, using equation (7):

$$\begin{array}{c}\text{First Correction}\\ n_{w,W,1}^{eff} = (1-0)/2 + 1/3 = 5/6 \\ n_{w,L,1}^{eff} = (0-1)/2 + 2/3 = 1/6.\end{array} \tag{9}$$

Because the 1-0 team beat a 0-1 team, worse than an average team, the 1-0 team is punished, and given only $5/6$ of a win, whereas the losing team lost to a 1-0 team, better than an average team, and is rewarded by suffering only $5/6$ of a loss. One can see how the method explicitly gives to one team only by taking from another.

The next step is to re-compute the ratings, given the new n_w^{eff} values. Plugging back into equation (1) yields:

$$\begin{array}{c}\text{Ratings After First Iteration}\\ r_{W,1} = (1+5/6)/(2+1) = 11/18 \approx 0.6111 \\ r_{L,1} = (1+1/6)/(2+1) = 7/18 \approx 0.3889.\end{array} \tag{10}$$

Let's look at just one more iteration.

$$\begin{array}{c}\text{Second Correction}\\ n_{w,W,2}^{eff} = (1-0)/2 + 7/18 = 8/9 \\ n_{w,L,2}^{eff} = (0-1)/2 + 11/18 = 1/9.\end{array} \tag{11}$$

$$\begin{array}{c}\text{Ratings After Second Iteration}\\ r_{W,2} = (1+8/9)/(2+1) = 17/27 \approx 0.6296 \\ r_{L,2} = (1+1/9)/(2+1) = 10/27 \approx 0.3704.\end{array} \tag{12}$$

If one examines the ratings of the winning team after the zeroth, first and second iterations, one finds that the values $r_{W,\{0,1,2\}} \approx \{0.6667, 0.6111, 0.6296\}$, show first a correction down, then

a correction up, by a lesser amount. Corrections that alternate in sign, and shrink in magnitude are hallmarks of *convergence*, meaning that with each iteration, the scheme is closer to finding a final, consistent value. Table 1 shows how these numbers converge to a part in a million after 11 iterations.

In fact, one can demonstrate that the final ratings in this simple case are explicitly the sums of converging series (compare to Table 1),

$$\begin{aligned} r_L &= \tfrac{1}{2}\left[1 - \tfrac{1}{3} + \tfrac{1}{9} - \tfrac{1}{27} + \cdots\right] \\ &= \tfrac{1}{2}\sum_{n=0}^{\infty}(-1/3)^n \\ &= \tfrac{1}{2}\cdot\tfrac{1}{1+1/3} = \tfrac{3}{8}, \end{aligned} \tag{13}$$

where the last line is the standard formula for the sum of a geometric series. In this simple case, the iterative method converges rapidly and stably, as a classic alternating geometric series.

Also note that the results converge to an average rating of 1/2, which is the same average as if there had been no game played at all; average rating has been conserved.

The ratings may converge nicely, but how can one know that these are the *right* answers? Furthermore, is the method extensible to the prodigiously more complicated case of 117 teams having played 11 or 12 games each?

6. The Colley Matrix Method

The previous section showed how an iterative correction for strength of schedule could provide consistent results that make intuitive sense for the simple one game case, but left us with the question of how do we know that the result is really right?

Let us return, then, to the example of the two teams 1-0, and 0-1 after their first game. Referring to equations (1) and (7), we have

$$\begin{aligned} r_W &= \tfrac{1+1/2+r_L}{2+1} \\ r_L &= \tfrac{1-1/2+r_W}{2+1}. \end{aligned} \tag{14}$$

A simple rearrangement gives

$$\begin{aligned} 3r_W - r_L &= 3/2 \\ -r_W + 3r_L &= 1/2, \end{aligned} \tag{15}$$

a simple two-variable linear system. Plugging in the results from the iterative technique (Table 1), one discovers that indeed $r_W = 5/8$ and $r_L = 3/8$ work exactly.

This exercise illustrates that linear methods can be used for two teams, but begs the question, can the ratings of many teams, after many games, be computed by simple linear methods?

6.1. The Matrix Solution

Returning to equations (1) and (7), using the same definitions for r_i and r_j^i, one finds that equations (1) and (7) can be rearranged in the form:

$$(2 + n_{tot,i})r_i - \sum_{j=1}^{n_{tot,i}} r_j^i = 1 + (n_{w,i} - n_{\ell,i})/2, \tag{16}$$

which is a system of N linear equations with N variables.

It is convenient at this point to switch to matrix form by rewriting equation (16) as follows,

$$C\vec{r} = \vec{b}, \tag{17}$$

where $\vec{r}$ is a column-vector of all the ratings r_i, and $\vec{b}$ is a column-vector of the right-hand-side of equation (16):

$$b_i = 1 + (n_{w,i} - n_{\ell,i})/2. \tag{18}$$

The matrix C is just slightly more complicated. The i^{th} row of matrix C has as its i^{th} entry $2+n_{tot,i}$, and a negative entry of the number of games played against each opponent j. In other words,

$$\begin{aligned} c_{ii} &= 2 + n_{tot,i} \\ c_{ij} &= -n_{j,i}, \end{aligned} \tag{19}$$

where $n_{j,i}$ is the number of times team i has played team j.

The matrix C is defined as the **Colley Matrix**. *Solving equations (17)–(19) is the method for rating the teams.* In practice, the matrix equation is solved in double precision by Cholesky decomposition and back-substitution (faster and more stable than Gauss-Jordan inversion, for instance [e.g., Press et al. 1992]). The Cholesky method is available, because the matrices are not only (obviously) symmetric and real, but are also positive definite, which will be discussed in the next section.

6.2. Equivalence of the Matrix and Iterative Methods

The matrix method has been shown to agree with the iterative method in the simple one game case. The question is whether the agreement extends to more complex situations. In a word, “yes,” but why?

There is no dazzingly elegant answer here. If the iterative scheme converges, then equations (1) and (7) are more nearly mutually satisfied with every iteration; otherwise the ratings would have to diverge at some point. When the ratings have converged, and iterating no longer introduces any changes to the ratings, the equations themselves have become simultaneously satisfied—the convergent ratings values have solved equations (1) and (7). Because those equations are identically the same ones solved by the matrix method, the matrix and iterative solutions must be identical as long as the iterative method remains convergent.

The question then becomes shifted to one of the convergence itself, which has been discussed only by example to this point. The convergence is due principally to $n+2$ denominator in equation (1). I shall not give a rigorous proof as to *exactly* why this is so, but rather motivate the idea in a less rigorous way.

The initial ratings are the final ratings plus some error.

$$\vec{r} = \vec{\rho} + \vec{\delta}, \tag{20}$$

where $\vec{\rho}$ is the vector of the true (final) ratings, and $\vec{\delta}$ the vector of errors. Let us consider the simple case where $\vec{\delta}$ has only one non-zero component, say $\delta_1 \neq 0$. Assuming a round-robin schedule (the slowest to converge), the iterations would proceed as in Table 2. The convergence in Table 2 is slow, with the errors decreasing by $\sim n/(n+2)$ in each iteration. In practice, the college football schedule is not round-robin outside of each conference, so the convergence factor is more like $\sim (n-2)/(n+2) \approx 10/14 \approx 0.71$, so convergence to a part in 10^7 occurs in about 48 iterations. In the 2000 season, for instance, the number of iterations required before median ratings correction fell below 10^7 was 60, so this very simple estimate is correct to 20% for a typical case.

Of course, in reality there are errors in more than one of the ratings, but, because the equations are linear, the principle of superposition applies, and the above calculation changes very little.

While the preceding is no proof that the scheme is always convergent, the round-robin case is the slowest to converge, and even in that case, we have shown that any error in a single rating does vanish over time, and superposition extends that to errors in multiple ratings. In the sparser case of an actual college football season the convergence can be slightly faster for some teams.

It should be noted that if it weren't for the $+2$ in the denominator of equation (1), the convergence would not occur, since, in the round-robin case, the error decrement would be by a factor of $\sim n/n = 1$, so if this method were based strictly on winning percentages, rather than Laplace's formula, it would fail.

The iterative scheme produces convergent ratings, which upon convergence simultaneously satisfy equations (1) and (7), which are exactly those that the matrix method solves; therefore, the iterative and matrix solutions must be equivalent (and, in practice, are equivalent).

6.3. Examples of the Colley Matrix Method

We now consider two examples to illustrate the matrix method in action. First, let us return to our friend, the simple two team, one game case. There, we discovered that the ratings could be determined from the linear system

$$\begin{aligned} 3r_W - r_L &= 3/2 \\ -r_W + 3r_L &= 1/2, \end{aligned} \tag{21}$$

Rewriting this in matrix form,

$$\begin{bmatrix} 3 & -1 \\ -1 & 3 \end{bmatrix} \begin{bmatrix} r_W \\ r_L \end{bmatrix} = \begin{bmatrix} 3/2 \\ 1/2 \end{bmatrix} \tag{22}$$

we recognize that r_W and r_L can be determined by simply inverting the matrix and multiplying by the solution vector on the right hand side.

$$\begin{bmatrix} 3/8 & 1/8 \\ 1/8 & 3/8 \end{bmatrix} \begin{bmatrix} 3/2 \\ 1/2 \end{bmatrix} = \begin{bmatrix} 5/8 \\ 3/8 \end{bmatrix}, \tag{23}$$

verifying the iterative and linear solutions.

But let us now consider a more complex example, a five team field, teams *a*–*e*. Their records against each other are shown thus (the initial and final ratings are listed for reference, the latter of which will be solved for forthwith):

team	*a*	*b*	*c*	*d*	*e*	record	initial rating	final rating
a	○	○	W	L	L	1-2	0.4	0.4130
b	○	○	L	○	W	1-1	0.5	0.5217
c	L	W	○	W	L	2-2	0.5	0.5000
d	W	○	L	○	○	1-1	0.5	0.4783
e	W	L	W	○	○	2-1	0.6	0.5870.

The matrix equation, according to equations (17)–(19) is written

$$\begin{bmatrix} 5 & 0 & -1 & -1 & -1 \\ 0 & 4 & -1 & 0 & -1 \\ -1 & -1 & 6 & -1 & -1 \\ -1 & 0 & -1 & 4 & 0 \\ -1 & -1 & -1 & 0 & 5 \end{bmatrix} \begin{bmatrix} r_a \\ r_b \\ r_c \\ r_d \\ r_e \end{bmatrix} = \begin{bmatrix} 1/2 \\ 1 \\ 1 \\ 1 \\ 3/2 \end{bmatrix}, \tag{24}$$

Solving the matrix equation yields sensible results:

$$\vec{r} = \{19, 24, 23, 22, 27\}/46 \approx \{0.413, 0.522, 0.500, 0.478, 0.587\} \tag{25}$$

As with the two team case, the ratings average to exactly 1/2 (23/46). Notice that team *b*, having played a 2-1 team and a 2-2 team, is rated higher than team *d*, having played a 1-2 team and a 2-2 team, despite identical 1-1 records: an example of how strength of schedule comes into play. Finally, there is consistency in that team *c* has a rating of exactly 1/2, because that team is 2-2 against teams whose ratings average to exactly 1/2.

7. Comments on the Colley Matrix Method

There has been discussion of the fact that the matrix method (and the iterative method) conserves an average ranking of 1/2. The reason I emphasize that point is that, as such, the ratings herein require no renormalization. All teams started out with a rating of 1/2, and only by exchange of rating points with other teams does one team's rating change. The first subsection below shows why the rating scheme explicitly conserves total ratings points. The subsection which follows establishes that the matrix in equation (17) is indeed positive definite, which allows for quick and stable solution of the matrix equation. The last subsection shows that the matrix C is singular if winning percentages are used in place of Laplace's formula, and thus, the method cannot be used with straight winning percentages.

7.1. Conservation of Average Rating

Why does the matrix solution always preserve the average rating of 1/2? We can tackle this problem by examining the construction of the matrix C. From the definition of the matrix C in equation (19), it follows that the matrix can be represented as:

$$C = 2I + \sum_{k}^{\text{all games}} G^k, \tag{26}$$

where I is the identity matrix, and G^k is a matrix operator added for each game k. G^k always has the form that $G^k_{ii} = G^k_{jj} = 1$, and $G^k_{ij} = G^k_{ji} = -1$, with all other entries 0, like this:

$$G^k = \begin{bmatrix} & \vdots & & \vdots & \\ \cdots & 1 & \cdots & -1 & \cdots \\ & \vdots & & \vdots & \\ \cdots & -1 & \cdots & 1 & \cdots \\ & \vdots & & \vdots & \end{bmatrix}. \tag{27}$$

Carrying out the multiplication $\vec{r}\,' = G^k\vec{r}$, we find that the i^{th} and j^{th} entries in $\vec{r}\,'$ are $r_i - r_j$, and $r_j - r_i$, respectively, while all other entries in $\vec{r}\,'$ are obviously zeroes. The sum of all the $\vec{r}\,'$ values, is, therefore, zero, no matter what the values of $\vec{r}$. Hence,

$$\sum_{i=1}^{N}(C\vec{r})_i = \sum_{i=1}^{N}(2I\vec{r})_i = 2\sum_{i=1}^{N} r_i. \tag{28}$$

What about the other side of equation (17), $\vec{b}$? The definition of b_i from equation (18) is $b_i = 1 + (n_{w,i} - n_{\ell,i})/2$. It is easy to see that the total of the b_i's must be N, since each win by one team must be offset by a loss by that team's opponent, so that the total number of wins must equal the total number of losses, and therefore, $\sum b_i = N$.

So, summing both sides of equation (17), we have

$$\begin{aligned} \textstyle\sum_i (C\vec{r})_i &= \textstyle\sum_i b_i \\ 2\textstyle\sum_i r_i &= N; \end{aligned} \tag{29}$$

$$\Rightarrow \frac{\sum r_i}{N} = \frac{1}{2}, \tag{30}$$

i.e., the average value of r is exactly one-half.

7.2. The Colley Matrix is Positive Definite

In order to use Cholesky decomposition and back substitution to solve the matrix equation (17), the matrix C must be positive definite, such that, for any non-trivial vector $\vec{v}$, this inequality holds

$$\vec{v}^{\,T}(C\vec{v}) > 0. \tag{31}$$

Recalling our separation of C into $2I + \sum_k G^k$, and noting that matrix multiplication is distributive, $(A + B)\vec{v} = A\vec{v} + B\vec{v}$, we can examine the inequality piece-wise. Obviously the matrix $2I$ is

positive definite, which leaves the G^k's. In the subsection above, we discovered that the multiplication $G\vec{r}$ yielded zeroes in all entries, except the i^{th} and j^{th} which contained $r_i - r_j$ and $r_j - r_i$, respectively. The product $\vec{r}^{\,T}(C\vec{r})$ is thus computed as

$$\vec{r}^{\,T}(G^k\vec{r}) = r_i(r_i - r_j) + r_j(r_j - r_i) = (r_i - r_j)^2 \geq 0. \tag{32}$$

Since $(r_i - r_j)^2 \geq 0$, and $\vec{r}^{\,T}(2I\vec{r}) > 0$, the matrix C must be positive definite.

7.3. Singularity of C for Straight Winning Percentages

We have seen that the iterative method would fail if straight winning percentages were used (i.e. if one removed the $+1$ and $+2$ from the numerator and denominator in equation [1]). In performing the same exercise with the matrix method, equation (19) would change C into a singular matrix!

For the one game case, the result is obvious.

$$C = \begin{bmatrix} 1 & -1 \\ -1 & 1 \end{bmatrix}, \tag{33}$$

obviously singular. In the general case, it's easy to see from equation (19) that if one removes the addend of 2 from c_{ii}, the total of the c_{ij}'s for any row or any column is zero. Therefore, in performing the legal row operation of adding all the rows, one has produced a row of all zeros; hence the matrix is singular. The matrix method, like the iterative method, cannot work with simple winning percentages.

8. Performance

With the mathematics of the ranking method on a firm foundation, the next question is how well does the method perform. Answering that question is difficult because we do not actually know the "truth." We simply do not know in any precise way how good Western Michigan is relative to Hawaii; therefore there is really nothing to check our results against.

The best we can do is compare our results to the venerable press polls, just as a mutual sanity check. Before I even begin, I should caution that the press polls are artificially correlated simply because the coaches are well aware of the AP poll, and the press writers are well aware of the Coaches' poll. Nonetheless we shall proceed.

Table 3 compares the final Colley rankings against the final AP rankings for 1998–2002 and against the Coaches' Poll rankings for 1999–2002. The first thing to notice is that the national

champion is agreed upon every year in all three ranking systems, which is ultimately the most important test for our purposes. Just scanning down the lists, the rankings are usually quite consistent within a few places, with an occasional outlier.

To quantify the agreement, I have found it more useful to think in terms of ranking ratios, or percentage differences, rather than simple arithmetic ranking differences. I have previously used the median absolute difference as a figure of merit, but have concluded this statistic poorly describes the behavior of the ranking comparisons. To show that, I have plotted in Fig. 2 the arithmetic differences *(top)* and ratios *(bottom)* of my rankings vs. the press rankings for 1999–2002 (all lumped together). In the plots, the histograms for the AP and Coaches' rankings are over-plotted (they're very similar).

Throwing all six groups together (three years, AP, Coaches), one can compute the direct average and variance of the distributions, which are listed as "avg" and "s" on the plots (yes, s is the square-root of the variance, and has a $\sqrt{n/(n-1)}$ factor). The corresponding normal curve is plotted as a dotted line. Another way of finding the mean and variance of a distribution is to fit the integral of the normal curve—that's $P(x)$ for you calculator statisticians, or the error function with $\sqrt{2}$'s in the right places for you math pedants—to the cumulative distribution. I have plotted those resulting normal curves as solid lines at top and bottom. If a distribution is normal, these two methods should be nearly identical. Obviously at top, the two curves are not so identical, but at bottom things seem much better.

Of course there are dozens of formal ways to check "Gaussianity" of a distribution, but I don't want to belabor the point. I just want to illustrate that the ratios are much better behaved than the arithmetic differences. What does this mean? It means that it's much more accurate to say my rankings disagree with the press rankings by a typical percentage, rather than by a typical number.

To compute that typical percentage, one may average the absolute differences of the logs of the rankings, as such,

$$\eta = \exp\left(\frac{1}{25}\sum_{i=1}^{25} |\log j_C(\text{team}_i) - \log i|\right), \tag{34}$$

where i is the press ranking (either AP or coaches), and $j_C(\text{team}_i)$ is the Colley ranking of that team. It's hard to come up with a good name for this statistic, η; perhaps "mean absolute ratio." Anyway, I list that statistic at the bottom of Table 3 for each of the 4 years of the BCS. The values are typically about 1.25, which means that my rankings agree with the press rankings within about 25% in either direction, so you might have an error of 1 ranking place at around #5, but about 5 places by #20. Inspecting the columns of Table 3 shows this to be quite a good description of the relative rankings.

Is that a *good* agreement?

Who knows, really? But to me (at least) the agreement is surprisingly good:

- The press polls started with a pre-season poll, with all the pre-conceived notions of history and tradition such an endeavor demands, then week by week allowed their opinions and judgments to migrate, being duly impressed or disappointed in the styles of winning and losing by certain teams, being more concerned about recent games than earlier ones, perhaps mentally weighting games seen on television as more important, perhaps having biases (good or bad) toward local schools one sees more often... *ad nauseam*.
- My computer rankings started with nothing, literally no information, but then, given only wins and losses, generated a ranking with pure algebra.

That two such processes produce even remotely consistent results is, frankly, remarkable to me.

I hope in this section we can agree to have learned, despite a lack of "truth" data, comparison of the press polls and my rankings shows both that the press and coaches must not be too loony, and that the Colley Matrix system yields common sense results.

9. Conclusions

Colley's Bias Free College Football Ranking Method, based on solution of the Colley Matrix, has been developed with several salient features, desirable in any computer poll that claims to be unbiased.

1. It has no bias toward conference, tradition, history, or prognostication.
2. It is reproducible; one can check the results.
3. It uses a minimum of assumptions.
4. It uses no *ad hoc* adjustments.
5. It nonetheless adjusts for strength of schedule.
6. It ignores runaway scores.
7. It produces common sense results that compare well to the press polls.

This information, the weekly poll updates, as well as useful college football links may be found on the Internet home for Colley's Rankings:

`http://www.colleyrankings.com/`.

WNC would like to thank A. Peimbert and J. R. Gott for their contributions in many lively

discussions on the subject of rankings. All programming for this method was done by WNC in IDL, FORTRAN, C, C++, Perl and shell script in the Solaris Unix and Linux environments.

REFERENCES

Press, W. H., Teukolsky, S. A., Vetterling, W. T., & Flannery, B. P., 1992, *Numerical Recipes in FORTRAN*, Cambridge University Press, Cambridge, UK, pp. 89-91

Convergence of Ratings via Iteration

Iteration	Winning Team's Rating	Losing Team's Rating
0	0.666667	0.333333
1	0.611111	0.388889
2	0.629630	0.370370
3	0.623457	0.376543
4	0.625514	0.374486
5	0.624829	0.375171
6	0.625057	0.374943
7	0.624981	0.375019
8	0.625006	0.374994
9	0.624998	0.375002
10	0.625001	0.374999
11	0.625000	0.375000

Table 1: Convergence of ratings to final, stable values, after 11 iterations, for the simple two team, one game case. The initial ratings are 2/3 for the winner and 1/3 for the loser, before any adjustment for schedule strength. Moving down the table are successive adjustments for strength of schedule. Because the winning team beat a below average (0-1) team, while the losing team lost to an above average (1-0) team, the final ratings are lower for the winning team, and greater for the losing team than were the initial ratings.

Iterative Convergence of Ratings in a Round-Robin

iter.	team 1	other teams
1.	$r \to \rho + \delta_1$ $n_w^{eff} = \nu_w^{eff}$	$r = \rho$ $n_w^{eff} \to \nu_w^{eff} + \delta_1$
2.	$r = \rho$ $n_w^{eff} = \nu_w^{eff} + n\delta_1/(2+n)$	$r = \rho + \delta_1/(2+n)$ $n_w^{eff} = \nu_w^{eff} + (n-1)\delta_1/(2+n)$
3.	$r = \rho + n\delta_1/(2+n)^2$ $n_w^{eff} = \nu_w^{eff} + n(n-1)\delta_1/(2+n)^2$	$r = \rho + (n-1)\delta_1/(2+n)^2$ $n_w^{eff} = \nu_w^{eff} + [(n-1)^2+n]\delta_1/(2+n)^2$
4.	$r = \rho + n(n-1)\delta_1/(2+n)^3$	$r = \rho + [(n-1)^2+n]\delta_1/(2+n)^3$
	$\vdots$	$\vdots$

Table 2: Iterative convergence in a round-robin. Ratings r and effective wins n_w^{eff} are computed from equations (1) and (7) in each iteration. Starting with the correct (final) values, ρ and ν_w^{eff}, an error δ_1 is added to team 1's rating. Column 1 gives the propagation of that error in team 1 through 4 iterations; Column 2 does the same for all other teams (whose errors will be equivalent). As one moves down the table, the errors shrink (slowly).

Comparison of Final Rankings with Press Polls

press rank	Colley Ranking for Teams with Given Press Rank								
	1998	**1999**		**2000**		**2001**		**2002**	
	AP	**AP**	**Coaches**	**AP**	**Coaches**	**AP**	**Coaches**	**AP**	**Coaches**
1	1	1	1	1	1	1	1	1	1
2	2	5	2	3	3	3	3	3	3
3	3	2	5	2	2	4	4	2	2
4	6	11	11	5	6	2	2	4	4
5	8	3	3	6	5	6	6	5	5
6	4	7	7	4	4	10	10	6	11
7	10	4	4	7	8	8	5	11	6
8	5	6	6	8	11	5	8	8	8
9	11	10	10	11	7	7	7	7	7
10	9	8	8	9	13	11	13	12	12
11	7	9	9	13	9	13	11	10	14
12	13	13	12	22	22	9	9	14	17
13	12	12	15	19	19	15	15	13	13
14	17	15	13	17	12	12	12	19	20
15	16	14	14	10	17	16	16	17	18
16	22	24	24	12	10	14	17	18	19
17	14	28	23	15	30	17	14	9	9
18	19	23	17	20	23	31	31	20	24
19	20	17	28	29	15	18	18	24	32
20	23	18	22	30	20	20	20	21	23
21	24	21	18	23	29	30	21	15	21
22	15	27	27	28	27	24	28	25	28
23	27	22	21	14	18	28	30	28	15
24	21	16	29	27	14	33	19	32	26
25	26	26	16	18	32	19	24	23	25
η									
Colley vs. poll	1.224	1.309	1.281	1.287	1.331	1.262	1.232	1.200	1.253
AP vs. Coaches	n/a	1.071		1.098		1.037		1.082	

Table 3: Comparison of final rankings to AP Poll for 1998–2002, and to the Coaches' Poll for 1999–2002. At bottom is a statistic η, described in the text. Essentially, it is the typical ratio of the Colley ranking to the poll ranking, or vice versa, so that the larger of the two always in the numerator, (specifically η is the exponent of the mean of the absolute values of the logs of the ratios), so $\eta = 1.25$ means the rankings would differ by typically one place at #4, and 5 places at #20.

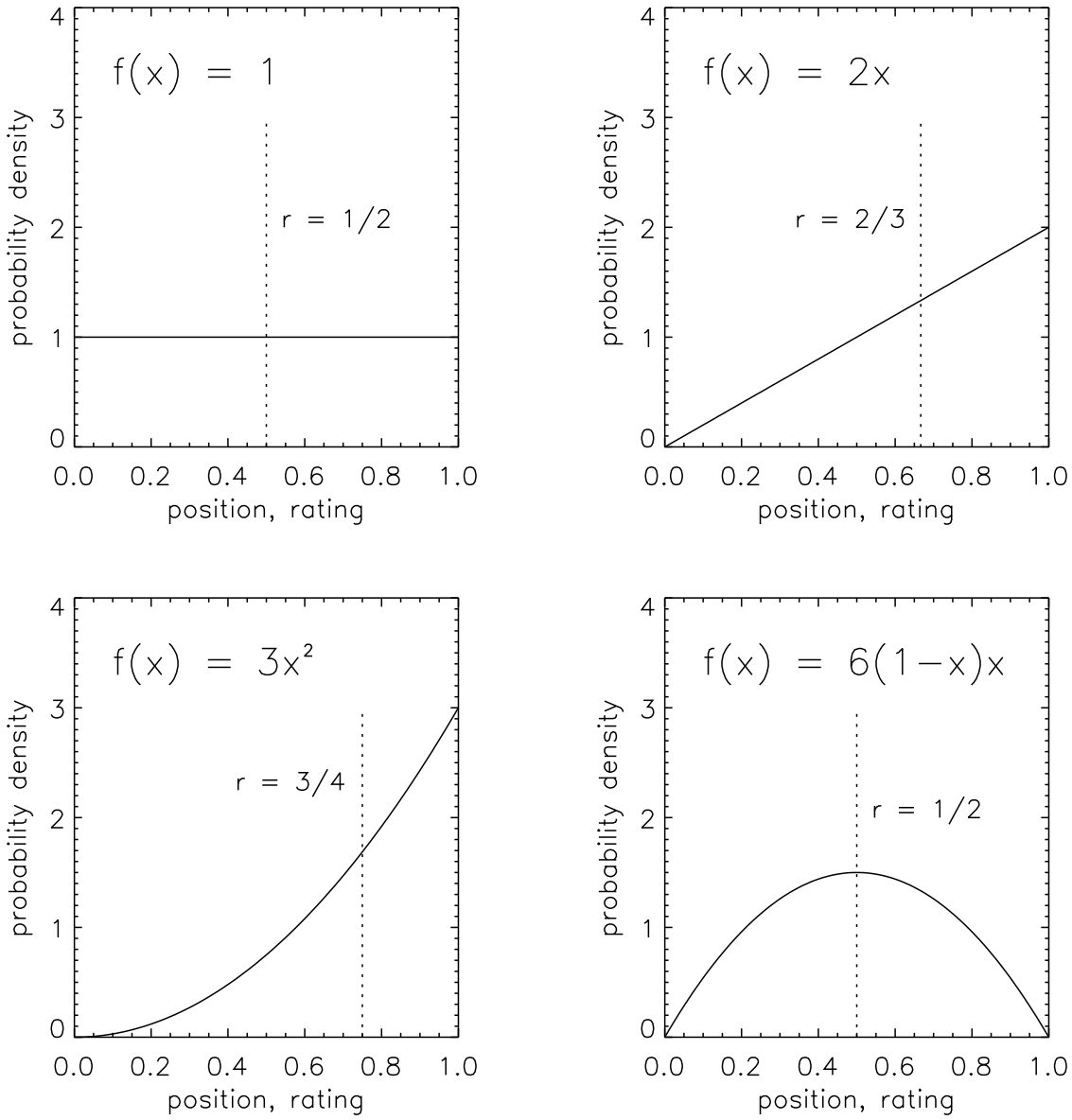

Fig. 1.— Probability distribution functions for the Laplace dice problem, analogous to rating by wins and losses. At top left is the initial condition {no dies thrown; no games played}. At top right is {one die left; one game won}: the probability density must be zero at the left. At bottom left is {two dies left; two games won}: the probability densities multiply. At bottom right is {one die left, one die right; one game won, one game lost}: the probability density must be zero at the left and at the right. The functions here have been normalized to have an integral of one, which is irrelevant in section 3 of the text, because the normalizations explicitly divide out.

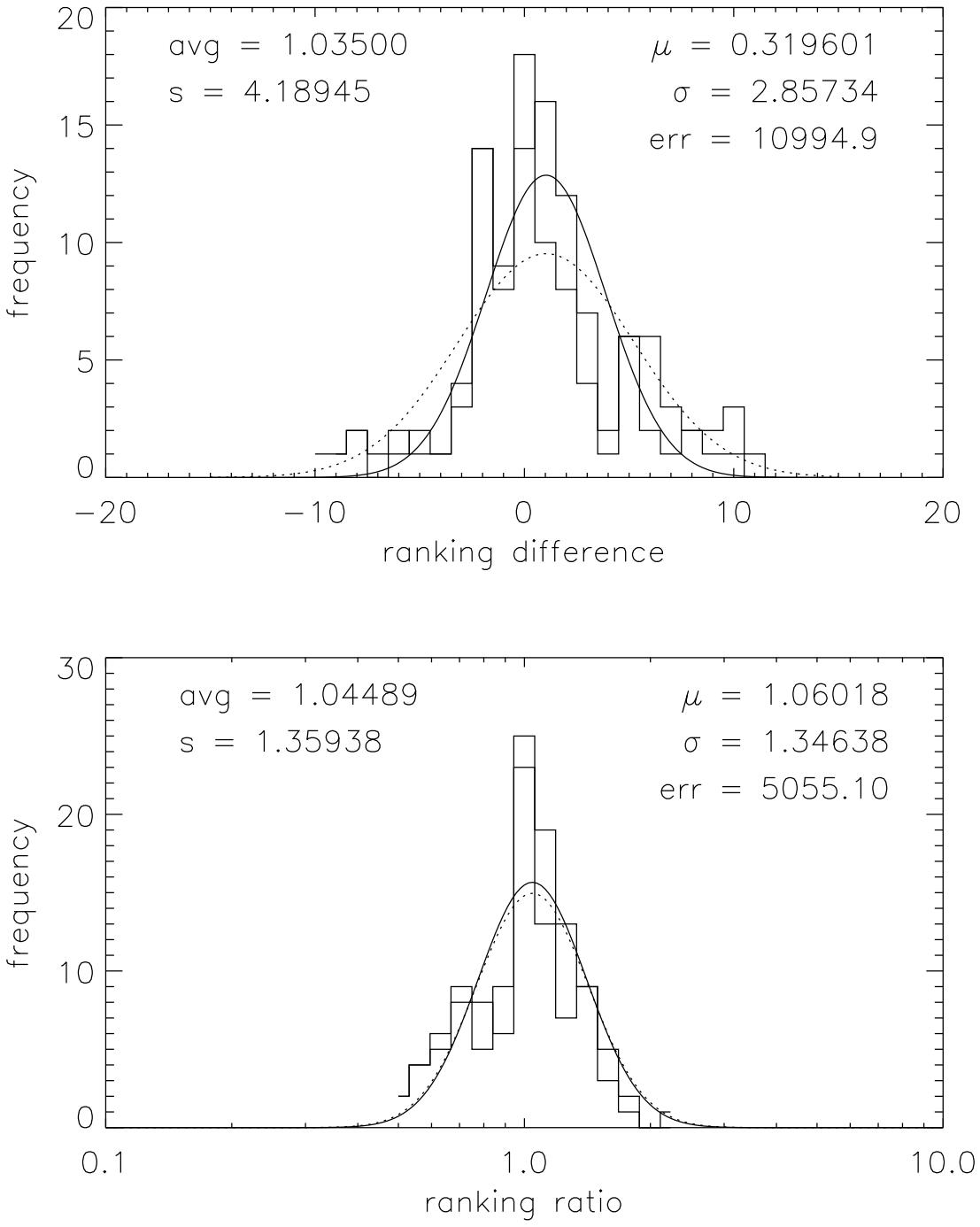

Fig. 2.— Two different ways for comparing the Colley Rankings with the Press Polls. *(at top)* Arithmetic differences, $(\text{Colley} - \text{press})$, between the final rankings for 1999–2002 for both the AP and Coaches' polls. Over-plotted are the normal curves from direct measurement of mean ($=$ *avg*) and standard deviation ($= s$), and from fitting for the mean ($= \mu$) and standard deviation ($= \sigma$). *(at bottom)* Same plots, but for ratios $(\text{Colley} \div \text{press})$ in logarithmic space.

SIAM REVIEW
Vol. 35, No. 1, pp. 80–93, March 1993

004

THE PERRON–FROBENIUS THEOREM AND THE RANKING OF FOOTBALL TEAMS*

JAMES P. KEENER†

Abstract. The author describes four different methods to rank teams in uneven paired competition and shows how each of these methods depends in some fundamental way on the Perron–Frobenius theorem.

Key words. Perron–Frobenius theorem, paired comparisions, ranking, orderings

AMS(MOS) subject classifications. 15-01, 15A48

1. Introduction. Throughout the fall of every year, arguments rage over which is the best college football team. The AP and UPI polls add to the confusion because they are based on votes which are certainly not objective. Many newspapers publish one or more additional indices that rank the top football teams, but these are not understood or accepted by the general public as easily as the polls, because they are usually based on "mathematical formulas." Given the general level of appreciation of mathematics among sports fans, these rankings are usually shrouded in mystery.

I first became interested in the problem of ranking football teams a few years ago when the football team at a rival campus won the national championship because it was the only undefeated team in the country. I wanted to know if a mathematically based ranking scheme would agree with the conclusions of the UPI and AP voters. What I learned (beyond what I hoped I would find!) is that a number of ranking schemes rely in some fundamental way on the Perron–Frobenius theorem, and that with the problem of ranking of teams in uneven paired competition I had discovered a marvelous way to motivate students to learn about a beautiful theorem that has in recent times fallen into relative obscurity.

An uneven paired competition is one in which the outcome of competition between pairs of teams (also called paired comparisons) is known, but the pairings are not evenly matched. That is, the competition is not a round robin in which each team is paired with every other team an equal number of times.

A good ranking scheme has a large number of potential uses. For example, it could be used to rank football teams, to create a tennis ladder, or to determine the research strength of mathematics departments. However, ranking schemes remove some, but not all, subjectivity, and different ranking schemes can give vastly different answers about who is number one, depending on the factors that are emphasized by the scheme.

This paper is about the ranking methods that I use. I use them not because they solve with certainty the problem of which team is number one, but because the mathematics is fun and well motivated. These methods are excellent vehicles by which to introduce students to interesting and important mathematical ideas, including the Perron–Frobenius theorem, the power and inverse power methods for finding eigenvalues of a matrix, and fixed point theorems for nonlinear maps. I find that the few minutes I spend each week during the fall collecting and entering data for my computer program are justified by the increased student interest in the mathematics of the methods generated by my weekly posted rankings. It is not difficult for students to write their own computer program to test some of these ideas on their favorite competition.

*Received by the editors January 2, 1992; accepted for publication (in revised form) August 25, 1992.
†Department of Mathematics, University of Utah, Salt Lake City, Utah 84112.

In this paper four different ranking schemes are described. The first, in §2, formulates the ranking problem as a linear eigenvalue problem and makes direct use of the Perron–Frobenius theorem. In §3, a nonlinear generalization of the first method is described. This method makes use of successive approximations to find a fixed point of a nonlinear map. The third and fourth methods attempt to assign a probability to the outcome of a contest, and make indirect use of the Perron–Frobenius theorem. Finally, in §6 we show the results of these four schemes when applied to the 1989 NCAA football schedule.

2. The direct method. The first method we describe is perhaps the most direct ranking method. To each participant in a contest we wish to assign a score that is based on the interactions with other participants. The assigned score should depend on both the outcome of the interaction and the strength of its opponents. If we suppose there is a vector of ranking values $\mathbf{r}$, with positive components r_j indicating the strength of the jth participant, then we define a score for participant i as

$$s_i = \frac{1}{n_i} \sum_{j=1}^{N} a_{ij} r_j, \tag{2.1}$$

where a_{ij} is some nonnegative number depending on the outcome of the game between participant i and participant j, N is the total number of participants in the competition, and n_i is the number of games played by participant i. The matrix A with entries a_{ij} is often called a preference matrix. For example, for football we could pick a_{ij} to be 1 if team i won the game, $\frac{1}{2}$ if the game ended in a tie, and zero otherwise. The division by n_i is to prevent teams from accumulating a large score by simply playing extra games.

Now we propose that the strength (or rank) of a participant should be proportional to its score, that is,

$$A\mathbf{r} = \lambda \mathbf{r}, \tag{2.2}$$

where A is the matrix with entries a_{ij}/n_i. In other words, the ranking vector $\mathbf{r}$ is a positive eigenvector of the positive matrix A.

The Perron–Frobenius theorem tells us when this problem has a solution, as follows.

THEOREM. *If the (nontrivial) matrix A has nonnegative entries, then there exists an eigenvector $\mathbf{r}$ with nonnegative entries, corresponding to a positive eigenvalue λ. Furthermore, if the matrix A is irreducible, the eigenvector $\mathbf{r}$ has strictly positive entries, is unique and simple, and the corresponding eigenvalue is the largest eigenvalue of A in absolute value (i.e., is equal to the spectral radius of A).*

To clarify the nomenclature, we refer to a vector with nonnegative entries as a nonnegative vector, and a vector with positive entries as a positive vector. We also introduce a partial order on the set of nonnegative vectors by saying that $\mathbf{p} > \mathbf{q}$ whenever $\mathbf{p} - \mathbf{q}$ is a positive vector and $\mathbf{p} \geq \mathbf{q}$ whenever $\mathbf{p} - \mathbf{q}$ is nonnegative.

The following are equivalent ways to describe an irreducible matrix.

(i) A is irreducible if for any two numbers i and j there is an integer $p \geq 0$ and a sequence of integers $k_1, k_2, \ldots, k_p$, so that the product $a_{ik_1} a_{k_1 k_2} \ldots a_{k_p j} \neq 0$.

(ii) A is irreducible if there is no permutation that transforms the matrix A into a block matrix of the form

$$\begin{matrix} A_{11} & A_{12} \\ 0 & A_{22} \end{matrix},$$

with A_{11} and A_{22} square matrices.

(iii) The nonnegative matrix A is irreducible if for any $\mathbf{r} \geq 0$, $A\mathbf{r} > 0$.

For paired competitions, if we take $a_{ij} = 0$ or 1 for a loss or a win, respectively, then the matrix A is irreducible if there is no partition of the teams into two sets S and T such that no team in S plays any team in T or every game between one team from S and one team from T resulted in a victory for the team in S. In particular, for this preference matrix to be irreducible, there can be no winless teams.

The proof of this theorem, while found in a number of older books, has not been included in most recent linear algebra books, so we include it in the appendix for completeness.

To calculate the eigenvector $\mathbf{r}$ we can use another powerful idea, namely, the power method [12], [10]. Since the ranking vector $\mathbf{r}$ is a simple eigenvector and corresponds to the largest eigenvalue of A, it follows that

$$\lim_{n\to\infty} \frac{A^n\mathbf{r}_0}{|A^n\mathbf{r}_0|} = \mathbf{r} \tag{2.3}$$

for any nonnegative vector $\mathbf{r}_0$.

Now comes the important question of how to pick the entries of the matrix A, and here there is room for subjectivity. We suggested earlier the choice $a_{ij} = 1$ if team i beat team j, $a_{ij} = \frac{1}{2}$ if team i and j tied, and $a_{ij} = 0$ otherwise. With this choice, if we guess an initial ranking vector $\mathbf{r}_0$ with all entries equal to one, then the ith component of $A\mathbf{r}_0$ is the winning percentage for team i. The ith component of vector $A^2\mathbf{r}_0$ is the average winning percentage of the teams that team i defeated. In some sense, $A^2\mathbf{r}_0$ contains information about the strength of schedule. I have heard it suggested by a nationally prominent football coach that $A^2\mathbf{r}_0$ should be used to determine a national champion. While this is a better ranking than the winning percentage $A\mathbf{r}_0$, it places a very high premium on strength of schedule. Of course, he did not express his scheme in mathematical notation, and, therefore, did not see the obvious generalization of using $A^n\mathbf{r}_0$ with large n. We now know that in the limit of n going to infinity, $A^n\mathbf{r}_0/|A^n\mathbf{r}_0|$ converges to the unique positive eigenvector of A, and this eigenvector gives a positive ranking for teams.

The idea of using the matrix A to find a ranking vector has been around for some time. Kendall and Babington Smith [6] considered the ranking $\mathbf{r} = A\mathbf{r}_0$, and the idea of powering the matrix A to find a ranking vector was initiated by Wei [13] and Kendall [5], and revisited often [1], [4], [9].

This simple choice for the entries a_{ij} leaves much to be desired. It is adequate for sports such as baseball where teams play each other often during a season. If teams play each other more than once, then a_{ij} is the total number of victories of team i over team j. With an increasing number of games, a_{ij} becomes a better indicator of the comparative strength of the two teams. But in football where teams play each other only once per season, there is information in the game score that is discarded when credit is given only for the win. For example, under this simple scheme, whether a score is nearly even or quite lopsided, all of the credit for the win goes to the winner. Also, a winless team has rank zero and, therefore, contributes nothing to the score of its opponents, and a matrix with a winless team is not irreducible. In fact, beating a winless team is more harmful than not playing that team at all because the winning team earns no points and its average point earning decreases.

A better method is to distribute the one point per game between two competing teams in a continuous, rather than discrete way. One way to assign a value to a_{ij} is to distribute the one point on the basis of the game score. If team i scored S_{ij} points and

team j scored S_{ji} points in their encounter, we might award $a_{ij} = S_{ij}/(S_{ij} + S_{ji})$ points to team i. This is slightly unfair because in a close defensive game with final score 3–0, the winner takes all, even though the two teams were evenly matched. To prevent this, we might consider an award of $a_{ij} = (S_{ij} + 1)/(S_{ij} + S_{ji} + 2)$ to team i, for example.

With such a scheme there is another weakness, namely, for a good team to show its dominance and get an appropriate score for the win, it can show no mercy. To avoid having teams run up a score to improve their ranking, the one point could be distributed in a nonlinear way. For example, the choice

$$a_{ij} = h\left(\frac{S_{ij}+1}{S_{ij}+S_{ji}+2}\right), \tag{2.4}$$

$$h(x) = \tfrac{1}{2} + \tfrac{1}{2}\operatorname{sgn}\left(x - \tfrac{1}{2}\right)\sqrt{|2x-1|}$$

has the features that it is continuous, $h(\frac{1}{2}) = \frac{1}{2}$, and away from $x = \frac{1}{2}$, h goes rapidly to zero or 1. A sketch of $h(x)$ is shown in Fig. 1. With an award distribution as in (2.4), to obtain a good score it is important to win, but not as useful to run up the score.

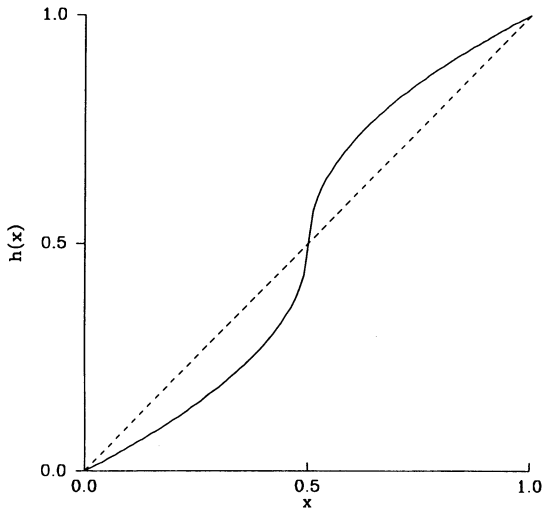

FIG. 1. *Plot of $h(x)$ as a function of x(solid curve) and the line $y(x) = x$(dashed) shown for comparison.*

3. A nonlinear scheme. Although the Perron–Frobenius scheme seems well motivated, after examining the results for a number of years its weaknesses became apparent. (By weakness, I mean that coaches and fans object to certain features of the ranking, not that there is a mathematical deficiency.) With this method, strength of schedule is quite important. If a strong team plays mostly weak opponents, with few strong opponents, it cannot earn a high ranking. This is because a team can never earn enough points playing weaker opponents to increase its earned score. Of course, this is not all bad, since simply because a team is undefeated does not mean it should have the highest rank, particularly if it did not play a difficult schedule. We have found that there is often an enormous difference in difficulty of schedule between some of the top ranked football teams.

There are a number of ways to address this problem. We could use this scheme to determine a national champion anyway and hope that coaches will eventually come to

understand that to earn a high ranking they cannot pad their team's schedule with weak opponents. This might also force some conferences with only one or two strong teams to consider realignment.

But another dilemma exists, and that is if a team does reasonably well against strong opponents, even though it may lose many or even most of its games, it can still earn a high ranking. For example, with the above linear method it is not unusual to find teams with losing records ranked among the top twenty-five teams. The reason for this is the decision implicit in the scheme to base ranking on a point system whereby one must earn points to improve one's rank. The teams that can optimize earning points are by definition the better teams. This may not be all bad, because some teams that are indeed very good nonetheless have losing records.

Since it is not likely that anytime in the near future coaches will be motivated by this ranking scheme to adjust their schedules, we decided to generalize this method to avoid the "problem" that a strong team with a weak schedule may be underrated. The idea is to calculate the rank for each team as

$$r_i = \frac{1}{n_i} \sum_{j=1}^{N} f(e_{ij} r_j), \tag{3.1}$$

where e_{ij} is a number that is determined from the outcome of the game between team i and team j, r_j is again the positive rank of team j, and f is some continuous monotone increasing function with $f(0) = 0$, and $f(\infty) = 1$. The advantage of this method is that now a team can earn up to a maximum of one point for each game it plays either by doing well against a highly ranked team, or by clobbering a poor team, but at least there is a way to have a weak schedule and still earn a good score.

We can again use interesting mathematics to conclude that a positive ranking vector $\mathbf{r}$ exists. If we define the nonlinear function of $\mathbf{r}$,

$$F_i(\mathbf{r}) = \frac{1}{n_i} \sum_{j=1}^{N} f(e_{ij} r_j), \tag{3.2}$$

then F is a bounded, nonlinear map of the positive orthant into itself. If we further suppose that $f(0) > 0$, and that $f(x)$ is a strictly concave function satisfying $f(tx) > tf(x)$ for all $t, 0 < t < 1$, then there is a unique fixed point of the map $F(\mathbf{r})$ in the positive orthant that can be found by successive approximation starting with any positive vector $\mathbf{r}_0$, whereby

$$\lim_{n \to \infty} F^n(\mathbf{r}_0) = \mathbf{r}. \tag{3.3}$$

The assumption $f(0) > 0$ implies that a team earns something just for showing up. Concavity is not strictly necessary to have a reasonable ranking, but it does guarantee a unique ranking vector. The proofs of these facts are relatively simple, and are relegated to the appendix. The equation $F(\mathbf{r}) = \mathbf{r}$ is a nonlinear eigenvector problem for which we seek a positive eigenvector, so this result can be viewed as a nonlinear generalization of the Perron–Frobenius theorem.

For this problem we found, after considerable experimentation, that

$$f(x) = \frac{.05x + x^2}{2 + .05x + x^2}, \tag{3.4}$$

and

$$e_{ij} = \frac{5 + S_{ij} + S_{ij}^{2/3}}{5 + S_{ji} + S_{ij}^{2/3}}, \tag{3.5}$$

work reasonably well. By reasonably well, we mean only that it gave results that aroused the ire of fewer people than did the linear method. A plot of $f(x)$ is shown in Fig. 2. The function $f(x)$ in (3.4) is not strictly concave, and neither is $f(0) > 0$, but it is close enough that the iterations (3.3) converge to a useful ranking vector. The function e_{ij} is shown plotted in Fig. 3 as a function of score S_{ij} for different values of S_{ji} fixed at 0, 10, 20, 30, and 50. Note that $e_{ij} = 1$ when $S_{ij} = S_{ji}$, that e_{ij} is an increasing function of S_{ij} and a decreasing function of S_{ji}.

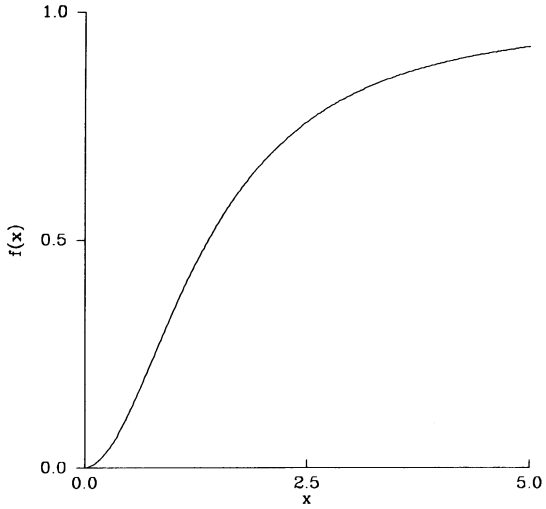

FIG. 2. *Plot of the function $f(x)$ as a function of x.*

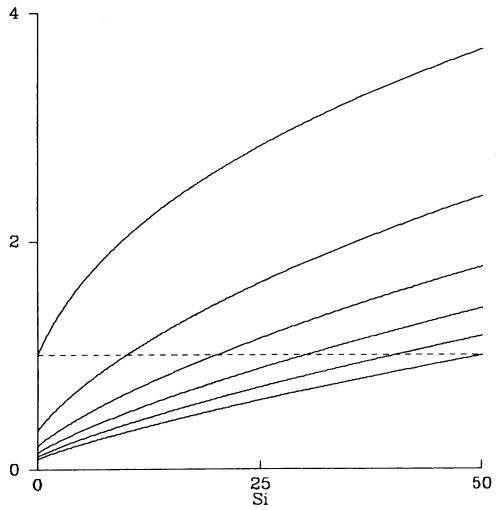

FIG. 3. *Plots of $e_{ij}(S_{ij}, S_{ji})$ as a function of S_{ij} with $S_{ji} = 0, 10, 20, 30, 40,$ and 50. Since $e_{ij}(x, x) = 1$, the value of S_{ji} can be identified by its intersection with the level 1 (dashed line).*

This choice for the map F in (3.2) points out the subjective nature of the methods described so far. The methods that follow are less subjective because they have an improved theoretical basis.

4. Assessing the probability of winning. Many people like to use ranking systems to predict the outcome of games between rivals, and so determining the probability that team i will beat team j is of primary interest. To this end, it would be nice if the ranking vector $\mathbf{r}$ could be given some probabilistic interpretation.

Suppose the ranking vector $\mathbf{r}$ is defined so that the probability π_{ij} that team i beats team j is

$$\pi_{ij} = \frac{r_i}{r_i + r_j}. \tag{4.1}$$

Since $\pi_{ij} + \pi_{ji} = 1$, it follows that

$$\pi_{ji} r_i - \pi_{ij} r_j = 0. \tag{4.2}$$

Unfortunately, we do not know π_{ij}, but if we did, we could find $\mathbf{r}$.

The relationship (4.2) between probability and the ranking vector is one of many possibilities in the class of so-called linear models having the form $\pi_{ij} = \prod(v_i - v_j)$, where $\mathbf{v}$ is the ranking vector [11]. The identification $r = e^v$ shows that (4.2) is a linear model. Other possibilities for the function $\prod$ are the Heaviside function, or

$$\prod(v) = \int_0^v e^{-x^2} dx. \tag{4.3}$$

The model (4.3) (due to Mosteller [8]) is motivated by the idea that the ith team has an actual performance that is a random variable with mean v_i and variance σ^2, σ being the same for each team. Then, if $\sigma = 1$, the probability π_{ij} that team i beats team j is

$$\pi_{ij} = \frac{1}{\sqrt{2\pi}} \prod \left(\frac{v_i - v_j}{\sqrt{2}} \right).$$

An interesting mathematical problem is to use statistical tests to determine the best linear model $\prod$. Bradley [2] gives a test of the hypothesis that the model (4.1) (known as the Zermelo model [14]) is correct.

If we use game scores to estimate π_{ij}, a reasonable estimate for π_{ij} is

$$\pi_{ij} = \frac{S_{ij}}{S_{ij} + S_{ji}}, \tag{4.4}$$

and (4.2) becomes

$$S_{ji} r_i - S_{ij} r_j = 0. \tag{4.5}$$

If teams i and j do not play each other, we take $S_{ij} = S_{ji} = 0$. Since, in any season there are many more games than there are teams, (4.5) gives many more equations than there are unknowns. Perhaps we can find the "best" solution of the overdetermined system (4.5) using a least squares method.

The least squares solution of all of the equations of the form (4.5) is trivial, $\mathbf{r} = \mathbf{0}$, and this is not the desired solution. Instead, we seek to minimize the squared error

subject to the constraint that $\mathbf{r}$ has norm 1. Thus, using the Lagrangian multipliers, we seek to minimize

$$\sum_{ij}(S_{ji}r_i - S_{ij}r_j)^2 - \mu\left(\sum_{i=1}^{N} r_i^2 - 1\right). \tag{4.6}$$

After differentiating (4.6) with respect to $\mathbf{r}$, we find that a minimum occurs only if $\mathbf{r}$ satisfies the matrix equation

$$B\mathbf{r} = \mu\mathbf{r}, \tag{4.7}$$

where the matrix B has entries b_{ij} given by

$$b_{ii} = \sum_k S_{ik}^2, \quad b_{ij} = -S_{ij}S_{ji}, \quad i \neq j. \tag{4.8}$$

To understand the solution properties of (4.7), we notice some important properties of the matrix B. The matrix B is invertible whenever the columns of the matrix associated with (4.5) are linearly independent, and it is reasonable to assume that this occurs naturally with enough games. The matrix B has positive diagonal and nonpositive off diagonal entries.

For some number $\lambda_0 > 0$, the shifted matrix $B' = B + \lambda_0 I$ is diagonally dominant. Then, for the vector $\mathbf{r}_0$ with all entries equal to 1, $B'\mathbf{r}_0$ has all positive entries. Now, notice what happens to the faces of the positive orthant under transformation by B'. If $\mathbf{r}_j$ has all entries positive except its jth entry which is zero, then the jth component of $B'\mathbf{r}_j$ is negative or zero. We will assume that there are enough entries in the matrix B so that the jth component of $B'\mathbf{r}_j$ is strictly negative, and then none of the faces of the positive orthant are invariant. In other words, the boundary of the positive orthant is mapped by the matrix B' to the exterior of the positive orthant. Since there is at least one vector, namely $\mathbf{r}_0$, that maps from the positive orthant into the positive orthant, it follows that B' maps the positive orthant to a cover of the positive orthant. Necessarily, B'^{-1} maps the positive orthant *into* the positive orthant and is therefore a positive map, meaning that its nonzero entries are positive. We conclude from the Perron–Frobenius theorem that B'^{-1} has a positive eigenvector $\mathbf{r}$, and that its corresponding eigenvalue is the largest in absolute value of the eigenvalues of B'^{-1}. As a result, $\mathbf{r}$ is the unique positive eigenvector of B', and the corresponding eigenvalue is the smallest eigenvalue in absolute value of B'.

The vector $\mathbf{r}$ is easily calculated by the inverse power method, since

$$\lim_{n\to\infty} \frac{(B + \lambda_0 I)^{-n}\mathbf{r}_0}{|(B + \lambda_0 I)^{-n}\mathbf{r}_0|} = \mathbf{r}. \tag{4.9}$$

Of course, we should never calculate the inverse of $B + \lambda_0 I$ explicitly, but rather calculate its LU decomposition, and then perform the inverse iteration using forward and backward substitution.

5. A maximum likelihood estimate. Suppose that the probability that team i beats team j is π_{ij}, and that the outcome of the contest between team i and team j is given by a_{ij}. For now we will take $a_{ij} = 1$ if team i beat team j, and zero otherwise. If the result of the contest between two teams is a Bernoulli trial with the outcome determined by the values π_{ij}, then the probability of the event a_{ij} is

$$P = \prod_{i<j} \binom{a_{ij} + a_{ji}}{a_{ij}} \pi_{ij}^{a_{ij}} \pi_{ji}^{a_{ji}}. \tag{5.1}$$

We now suppose that the ranking vector $\mathbf{r}$ has the property that

$$\pi_{ij} = \frac{r_i}{r_i + r_j}, \tag{5.2}$$

so the probability that the outcome is represented by a_{ij} is

$$P(\mathbf{r}) = \prod_{i<j} \binom{a_{ij} + a_{ji}}{a_{ij}} \left(\frac{r_i}{r_i + r_j}\right)^{a_{ij}} \left(\frac{r_j}{r_i + r_j}\right)^{a_{ji}}. \tag{5.3}$$

Since the outcome a_{ij} is known to have occurred, we pick $\mathbf{r}$ so that $P(\mathbf{r})$ is as large as possible. The resulting vector $\mathbf{r}$ is called the maximum likelihood solution.

The problem of choosing $\mathbf{r}$ to maximize $P(\mathbf{r})$ is quite old. This model and an iterative method for its solution was first proposed by Zermelo in 1926 [14] and then rediscovered by Ford in 1955 [7] and is often called a Bradley–Terry model [3], [11]. We give a new proof of existence and uniqueness of the solution here.

With the choice (5.2) and since the matrix A is fixed, it is equivalent to maximize the function

$$F_A(\mathbf{r}) = \prod_{i<j} \left(\frac{r_i}{r_i + r_j}\right)^{a_{ij}} \left(\frac{r_j}{r_i + r_j}\right)^{a_{ji}}, \tag{5.4}$$

or

$$\ln F_A(\mathbf{r}) = \sum_{i<j} (a_{ij}(\ln r_i - \ln(r_i + r_j)) + a_{ji}(\ln r_j - \ln(r_i + r_j))). \tag{5.5}$$

To show that a maximum exists, we assume that the matrix A is irreducible. Clearly, the function $F_A(\mathbf{r})$ is continuous and bounded on the interior of the positive orthant. While it is not defined on the faces of the positive orthant, if A is irreducible we can define $F_A = 0$ on the faces of the positive orthant as the continuous extension of $F_A(\mathbf{r})$. That is, if $\mathbf{r}_0$ is on a face of the positive orthant, then one of its elements, say r_i, is zero, and another of its elements, say r_j, is nonzero. Because the matrix A is irreducible, there is a sequence of indices $i_0, i_1, \ldots, i_k$, with $i_0 = i$ and $i_k = j$ with the property that $a_{i_p i_{p+1}} > 0$ for $p = 0, 1, \ldots, k-1$. Necessarily, there are consecutive integers $m < n$ for which $r_{i_m} = 0$, and $r_{i_n} > 0$.

We write

$$F_A(\mathbf{r}) = \left(\frac{r_{i_m}}{r_{i_m} + r_{i_n}}\right)^{a_{i_m i_n}} \phi(\mathbf{r}), \tag{5.6}$$

and observe that $\phi(\mathbf{r})$ is positive and bounded in the interior of the positive orthant. It follows that $\lim_{r \to r_0} F_A(\mathbf{r}) = 0$. As thus extended, the function $F_A(\mathbf{r})$ is continuous and bounded on the closed and bounded set $\sum = \{\mathbf{r} | r_i \geq 0, \sum_i r_i = 1\}$, $F_A(\mathbf{r})$ is strictly positive on the interior of the set $\sum$, and is zero on the boundary of the set $\sum$. It follows that $F_A(\mathbf{r})$ attains a maximum on the interior of the positive orthant. (This part of the proof is from Ford's work [7].)

To find an extremum we differentiate the function $\ln F_A(\mathbf{r})$ to find

$$\frac{\partial}{\partial r_k} \ln F_A(\mathbf{r}) = \frac{\alpha_k}{r_k} - \sum_j \frac{A_{jk}}{r_j + r_k}, \tag{5.7}$$

where $\alpha_k = \sum_j a_{jk}$, and $A_{jk} = a_{jk} + a_{kj}$. Consequently, the maximizing vectors $\mathbf{r}$ must satisfy the nonlinear system of equations

$$\frac{\alpha_k}{r_k} - \sum_j \frac{A_{jk}}{r_j + r_k} = 0. \tag{5.8}$$

Zermelo [14] and Ford [7] used an iterative method to solve (5.8). In my opinion, it is just as easy to solve (5.8) by integrating the system of differential equations

$$\frac{dr_k}{dt} = \frac{\alpha_k}{r_k} - \sum_j \frac{A_{jk}}{r_j + r_k}, \tag{5.9}$$

using one's favorite numerical integrator, starting from any initial point in the interior of the positive orthant. We are assured that the solution of the differential equation system (5.9) will approach a steady state because it is a gradient system, and along trajectories

$$\begin{aligned} \frac{d \ln F_A(\mathbf{r})}{dt} &= \sum_k \frac{\partial}{\partial r_k} (\ln F_A(\mathbf{r})) \frac{dr_k}{dt} \\ &= \sum_k \left(\frac{\partial}{\partial r_k} (\ln F_A(\mathbf{r})) \right)^2, \end{aligned} \tag{5.10}$$

which is positive except at an extremum of $\ln F_A(\mathbf{r})$. Hence $F_A(\mathbf{r})$ increases along trajectories of (5.9).

Finally, we can show that the maximum of $\ln F_A(\mathbf{r})$ is unique. We calculate the Hessian H of $\ln F_A(\mathbf{r})$ at any extremum to be $H = (h_{ik})$, where

$$h_{ik} = \frac{\partial \ln F_A(\mathbf{r})}{\partial r_i \partial r_k} = \left(-\frac{\alpha_i}{r_i^2} + \sum_j \frac{A_{ij}}{(r_i + r_j)^2} \right) \delta_{ik} + \frac{A_{ik}}{(r_i + r_j)^2}. \tag{5.11}$$

By virtue of (5.8),

$$\alpha_i = \sum_j A_{ij} \left(\frac{r_i}{r_i + r_j} \right) > \sum_j A_{ij} \left(\frac{r_i}{r_i + r_j} \right)^2, \tag{5.12}$$

and the off-diagonal elements of H are positive. Observe also that H has a null space, since $H\mathbf{r} = 0$ for any vector $\mathbf{r}$ satisfying (5.8). This null space results from the invariance of (5.4) under changes of the scale of $\mathbf{r}$.

Now we want to find the eigenvalues of H. Notice that for any sufficiently large positive λ_0, the matrix $-H + \lambda_0 I$ is diagonally dominant. However, because all of its off-diagonal elements are negative, the matrix $-H + \lambda_0 I$ maps the boundary of the positive orthant to the exterior of the positive orthant. It follows from our friend the Perron–Frobenius theorem that $(-H + \lambda_0 I)^{-1}$ is a map of the positive orthant into itself, having a unique positive eigenvector with corresponding eigenvalue μ_1, with $\mu_1 > \mu_2 \geq \cdots \geq \mu_n$. (This is the same application of the Perron-Frobenius theorem as used in §4.) Therefore, the eigenvalues of H are

$$\lambda_0 - \frac{1}{\mu_1} > \lambda_0 - \frac{1}{\mu_2} \geq \cdots \geq \lambda_0 - \frac{1}{\mu_n}. \tag{5.13}$$

The maximizing vector $\mathbf{r}$ satisfying (5.8) also satisfies $H\mathbf{r} = 0$ so that $\mathbf{r}$ is an eigenvector of $(-H + \lambda_0 I)^{-1}$ as well, and being positive, must correspond to its largest eigenvalue. It follows that $\lambda_0 - (1/\mu_1) = 0$, and the remaining eigenvalues of H must be strictly negative. Thus, on the surface $\sum$, all extrema for $\ln F_A(\mathbf{r})$ are local maxima. We conclude that there is, therefore, exactly one extremum.

The motivation for this model was based on the assumption that the numbers a_{ij} were integers. But clearly, there is nothing in the proof of existence and uniqueness that forces this requirement. For the purpose of ranking football teams it is preferable to use a different determination for a_{ij}. A choice that works well is

$$a_{ij} = \frac{S_{ij}}{S_{ij} + S_{ji}}. \tag{5.14}$$

6. Putting it all together. There are 106 Division I-A college football teams in the United States, which during each season play about 570 games, including bowl games. Schedules for the coming season and results from the previous season are available annually in the NCAA Football book (available from NCAA Publications, Mission, KS 66201).

In Table 1, we present the results of the above ranking schemes for the 1989 season. In Table 1 there are eight columns. The top 40 teams are ordered in the table according to percentage of wins. W-L-T refers to the win-lose-tie record for the 1989–90 season (including bowl games). Columns labelled 1–4 show the integer rank of the team for methods 1–4, respectively, and the columns labeled UPI and AP are the final poll results for those teams that were ranked.

For this table, methods 1–4 are defined as follows:

(1) The direct linear method based on the eigenvalue problem (2.2) with entries a_{ij} chosen using (2.4);

(2) The nonlinear method (3.1) with $f(x)$ satisfying (3.6) and scoring factors e_{ij} satisfying (3.7);

(3) The least squares estimate of probabilities (4.6);

(4) The maximum likelihood method (Bradley–Terry model) (5.3) with entries a_{ij} satisfying (5.15).

What can we conclude from all of this? First, there is no unique way to devise a ranking scheme. The different ranking schemes give different rankings because they weigh important factors differently. Each of the schemes proposed here have strengths and weaknesses, but invariably when a method is tweaked to get rid of some "undesirable" feature, another "counterintuitive" result shows up. After studying these methods for awhile, it is also apparent that intuition is not a good guide to determining a ranking. With 106 teams there are just too many factors to consider. On the other hand, the numbers are not biased; they simply report the results of the algorithm.

It is interesting to compare the results of the ranking algorithms with the UPI and AP polls. First, it is obvious that there is much more variation between the ranking schemes than between the polls, suggesting that the two polls are not independent. Second, there are noted differences between the polls and the ranking schemes. For example, counting the number of teams whose poll rankings do not lie within the range of rankings from the four mathematical schemes, we find nine teams for whom the polls are "too high" and four teams for whom the polls are "too low."

TABLE 1
Ranking of NCAA *Division* I-A *football teams for* 1989 *season.*

Team	W-L-T	#1	#2	#3	#4	UPI	AP
Notre Dame	12-1-0	3	6	5	3	3	2
Miami (Fla)	11-1-0	8	2	1	1	1	1
Tennessee	11-1-0	4	10	8	11	5	5
Colorado	11-1-0	11	5	10	7	4	4
Fresno State	10-1-0	48	20	47	43		
Florida St.	10-2-0	2	3	2	5	2	3
Auburn	10-2-0	6	8	4	8	6	6
Alabama	10-2-0	9	11	6	12	7	9
Arkansas	10-2-0	10	16	11	17	13	13
Michigan	10-2-0	16	12	12	9	8	7
Illinois	10-2-0	20	18	20	14	10	10
Nebraska	10-2-0	29	9	13	13	12	11
Clemson	9-2-0	5	7	9	6	11	12
Houston	9-2-0	15	1	3	2		14
USC	9-2-1	1	4	7	4	9	8
Northern Illinois	7-2-0	65	52	59	68		
BYU	10-3-0	22	21	27	25	18	
Virginia	9-3-0	13	22	19	19	15	18
Texas Tech	9-3-0	21	33	25	27	16	19
Hawaii	9-3-1	33	28	46	41		
Eastern Michigan	6-2-1	79	49	83	90		
Penn State	8-3-1	19	15	18	18	15	15
Pittsburgh	8-3-1	41	36	37	34	19	17
Syracuse	7-3-0	45	39	38	45		
West Virginia	7-3-1	37	27	35	38		
Washington	8-4-0	7	13	14	10	19	20
Arizona	8-4-0	12	24	28	20		
Oregon	8-4-0	17	19	23	21		
Texas A & M	8-4-0	18	17	16	16		20
Duke	8-4-0	26	38	34	37		
Michigan State	8-4-0	32	14	15	15	16	16
Ohio State	8-4-0	51	42	48	30		
Air Force	8-4-1	30	30	32	33		
Mississippi	7-4-0	24	34	21	31		
Oklahoma	7-4-0	57	29	44	29		
Ball State	6-3-2	81	59	90	89		
Georgia Tech	6-4-0	25	40	30	42		
Arizona State	6-4-1	39	57	52	44		
Virginia Tech	6-4-1	36	35	42	26		
Florida	7-5-0	27	25	22	23		
Number of upsets		102	110	108	100		

There are other ways that one might try to rank teams. For example, a method that is unrelated to those presented here is to try to minimize the number of upsets. An upset occurs when a team is ranked higher than a team to which it lost. At the bottom of the columns in Table 1 are listed the number of upsets for each of the algorithms used here. The number of upsets cannot be zero since there is no well ordering, but by assigning an objective function that measures the degree of an upset, we can devise algorithms to find the best ranking with respect to that particular measure.

Appendix A. Proof of the Perron–Frobenius theorem. Let $\sum$ be the set of all non-negative vectors with Euclidean norm one. For each vector $\mathbf{s}$ in the set $\sum$ let σ^* be the positive number for which $A\mathbf{s} \leq \sigma\mathbf{s}$ whenever $\sigma \geq \sigma^*$. If $\mathbf{s}$ has zero entries then σ^* may be infinite. Since $\sum$ is a closed and bounded set, the smallest value of σ^* is attained for some vector $\mathbf{s}^*$ in $\sum$. We claim that $\mathbf{s}^*$ is a positive eigenvector of A.

Suppose that $A\mathbf{s}^* \leq \sigma^*\mathbf{s}^*$ but $\mathbf{s}*$ is not an eigenvector of A. Then some, but not all, of the relations in the statement $A\mathbf{s}^* \leq \sigma^*\mathbf{s}^*$ are equalities. (If there were no equalities, the number σ^* would be incorrectly chosen.) After permutation, we can write the relations

$A\mathbf{s}^* \leq \sigma^*\mathbf{s}^*$ in the form

$$\begin{aligned} A_{11}\mathbf{s}_1 + A_{12}\mathbf{s}_2 &< \sigma^*\mathbf{s}_1, \\ A_{21}\mathbf{s}_1 + A_{22}\mathbf{s}_2 &= \sigma^*\mathbf{s}_2. \end{aligned} \tag{A.1}$$

Since A is irreducible, A_{21} is not identically zero, so we can reduce at least one component of the vector $\mathbf{s}_1$, thereby changing at least one of the equalities to a strict inequality, without changing any of the original strict inequalities. After this change in $\mathbf{s}^*$ we rescale the vector to have norm one. Proceeding inductively, we can continue to modify the vector $\mathbf{s}^*$ until all of the relations in $A\mathbf{s}^* \leq \sigma^*\mathbf{s}^*$ are strict inequalities, but of course, this contradicts the definition of σ^*; so we are done.

To prove uniqueness, we note that a nonnegative eigenvector $\mathbf{r}$ must have all positive entries. Suppose there are two linearly independent eigenvectors of A, $\mathbf{r}_1$ and $\mathbf{r}_2$, satisfying $A\mathbf{r}_1 = \lambda_1\mathbf{r}_1$, and $A\mathbf{r}_2 = \lambda_2\mathbf{r}_2$, and suppose that $\mathbf{r}_1$ has strictly positive entries. If the entries of $\mathbf{r}_2$ are all of one sign, then without loss of generality they can be taken as positive. The vector $\mathbf{r}(t) = r_1 - t\mathbf{r}_2$ has nonnegative entries for all t in some range $0 \leq t \leq t_0$ with $t_0 > 0$, and $\mathbf{r}(t_0)$ has some zero entries but is not identically zero, while for $t > t_0$, $\mathbf{r}(t)$ has some negative entries. Then $A\mathbf{r}(t_0) = \lambda_1(\mathbf{r}_1 - t_0\lambda_2/\lambda_1\mathbf{r}_2)$ has only positive entries. By the maximality of t_0, it must be that $|\lambda_2| < |\lambda_1|$. But if both $\mathbf{r}_1$ and $\mathbf{r}_2$ have only positive entries, we can interchange them in the above argument to conclude that $|\lambda_1| < |\lambda_2|$. This is, of course, a contradiction. We conclude that the positive eigenvector is unique and all other eigenvectors have eigenvalues that are smaller in absolute value. A minor modification of this argument shows that the largest eigenvalue is simple. For, if $\mathbf{r}_2$ is a generalized eigenvalue of A satisfying $A^k\mathbf{r}_2 = \lambda_1^k\mathbf{r}_2$ for some $k > 1$, then $A^k\mathbf{r}(t_0) = \lambda_1^k\mathbf{r}(t_0)$ is strictly positive, contradicting the definition of t_0.

Appendix B. Proof of the nonlinear fixed point theorem (nonlinear generalization of the Perron–Frobenius theorem). Suppose F is a positive, monotone, and strictly concave mapping of a finite-dimensional space to itself. That is, $F(\mathbf{r}) > 0$ for all $\mathbf{r} > 0$, $F(\mathbf{p}) > (\geq)F(\mathbf{q})$ whenever $\mathbf{p} > (\geq)\mathbf{q}$, and $F(t\mathbf{r}) > tF(\mathbf{r})$ for $0 < t < 1$.

To see that there is at least one positive fixed point, let $\mathbf{r}_0$ have all entries equal to 1 and notice that $F(\mathbf{r}_0) < 1$. Define the sequence of vectors $\mathbf{r}_k$ by successive approximation

$$\mathbf{r}_k = F(\mathbf{r}_{k-1}), \tag{B.1}$$

and notice that $\mathbf{r}_k < \mathbf{r}_{k-1}$. The monotone decreasing sequence of vectors $\{\mathbf{r}_k\}$ is bounded below by $F(\mathbf{0}) > \mathbf{0}$, and therefore converges to some positive vector $\mathbf{r}$. Since F is continuous, $\mathbf{r}$ is a fixed point of F.

The positive fixed point $\mathbf{r}$ is unique. If not, there is a positive vector $\mathbf{q}$ satisfying $F(\mathbf{q}) = \mathbf{q}$. Since $\mathbf{r} \neq \mathbf{q}$, one of the inequalities $\mathbf{r} \leq \mathbf{q}$ and $\mathbf{q} \leq \mathbf{r}$ must fail to hold. Without loss of generality suppose that $\mathbf{q} \leq \mathbf{r}$ does not hold. Now, there is a maximal t_0 with $0 < t_0 < 1$ so that $t\mathbf{q} \leq \mathbf{r}$ for all t in $0 \leq t \leq t_0$. Therefore,

$$\mathbf{r} = F(\mathbf{r}) \geq F(t_0\mathbf{q}) > t_0F(\mathbf{q}) = t_0\mathbf{q}, \tag{B.2}$$

contradicting the maximality of t_0.

Acknowledgment. Thanks to Joe Keller for introducing me to this fascinating topic over ten years ago, and to Fred Phelps for his ideas on how to define a nonlinear ranking scheme.

REFERENCES

[1] C. BERGE, *The Theory of Graphs and Its Applications*, Wiley and Sons, New York, 1962.

[2] R. A. BRADLEY, *Incomplete block rank analysis: on the appropriateness of a model for the method of paired comparisons*, Biometrics, 10 (1954), pp. 375–390.

[3] R. A. BRADLEY AND M. E. TERRY, *The rank analysis of incomplete block diagrams*. I. *The method of paired comparisons*, Biometrica, 39 (1952), pp. 324–345.

[4] H. A. DAVID, *The Method of Paired Comparisons*, Griffin, London, 1969.

[5] M. G. KENDALL, *Further contributions to the theory of paired comparisons*, Biometrics, 11 (1955), p. 43.

[6] M. G. KENDALL AND B. BABINGTON SMITH, *On the method of paired comparisons*, Biometrica, 31 (1939), p. 324.

[7] L. R. FORD, JR., *Solution of a ranking problem from binary comparisons*, Amer. Math. Monthly, 64 (1957), pp. 28–33.

[8] F. MOSTELLER, *Remarks on the method of paired comparisons*: I. *The least squares solution assuming equal standard deviations and equal correlations*, 16 (1951), pp. 3–9.

[9] T. L. SAATY, *Rank according to Perron: a new insight*, Math. Magazine, 60 (1987), p. 211.

[10] G. W. STEWART, *Introduction to Matrix Computations*, in Computer Science and Applied Mathematics, W. Rheinboldt, ed., Academic Press, New York, 1973.

[11] M. STOB, *A supplement to "A mathematicians guide to popular sports,"* Amer. Math. Monthly, 91 (1984), pp. 277–281.

[12] R. S. VARGAS, *Matrix Iterative Analysis*, Prentice-Hall, Englewood Cliffs, NJ, 1962.

[13] T. H. WEI, *The algebraic foundations of ranking theory*, Cambridge University Press, London, 1952.

[14] E. ZERMELO, *Die Berechnung der Turnier-Ergebnisse als ein Maximumproblem der Warscheinlichkeitsrechnung*, Math. Z., 29 (1926), pp. 436–460.

An overview of some methods for ranking sports teams

Soren P. Sorensen
University of Tennessee
Knoxville, TN 38996-1200
soren-sorensen@utk.edu

Introduction

The purpose of this report is

- to argue for an open system for ranking sports teams,
- to review the history of ranking systems, and
- to document a particular open method for ranking sports teams against each other.

In order to do this extensive use of mathematics is used, which might make the text more difficult to read, but ensures the method is well documented and reproducible by others, who might want to use it or derive another ranking method from it. The report is, on the other hand, also more detailed than a ”typical” scientific paper and discusses details, which in a scientific paper intended for publication would be omitted.

We will in this report focus on NCAA 1-A football, but the methods described here are very general and can be applied to most other sports with only minor modifications.

Predictive vs. Earned Ranking Methods

In general most ranking systems fall in one of the following two categories: predictive or earned rankings. The goal of an earned ranking is to rank the teams according to their past performance in the season in order to provide a method for selecting either a champ or a set of teams that should participate in a playoff (or bowl games). The goal of a predictive ranking method, on the other hand, is to provide the best possible prediction of the outcome of a future game between two teams.

In an earned system objective and well publicized criteria should be used to rank the teams, like who won or the score difference or a combination of both. By using well defined criteria for the ranking then teams know exactly what the consequences of a win or at loss will be. This is done in most football conferences to select the conference champion or at least the two teams selected for a championship game. In general an earned ranking system allocates a number (often called the power ranking), which is used the order teams in a linear sequence.

Most systems found on the WWW is predictive, even many of the BCS systems. In order to make a predictive system as accurate as possible it is allowed to include any information, which is deemed useful, like the strength of the quarterback, yards earned, number of fumbles etc. In particular it is very common to put more weight on recent games than older games. This allows a more precise extrapolation the next weeks games. In a more advanced predictive system the teams are not necessarily linearly ordered. One can easily imagine situations, where a good predictive system will

predict, that A beats B, B beats C, but C beats A. This is not possible in a pure earned system.

Unfortunately most WWW ranking systems are a strange mix of the two types of systems described here. The BCS ranking, that determines which teams have earned the right to play in the various bowls, in particular the bowl determining the national championship, should be a pure earned ranking system,. But may, if not most, of the BCS systems seem to be predictive. It is even so bad, that a particular web site ranks the BCS systems according to how well they predict next weeks games!

The method described below are all intended as earned systems and now efforts are spent on trying to optimize for predictive capabilities.

Open vs. Closed Methods

Currently the Bowl Championship Series (BCS) system cite: BCS is based on 4 components, 1) 2 subjective polls by coaches and journalists, b) 8 computer programs, c) a well defined, but primitive, method for calculating the strength of schedule, and 4) the number of losses of each team.

Here we will be concern ourselves with the 8 BCS computer based methods for ranking the teams.

Billingsley cite: Billingsley .

Dunkel cite: Dunkel .

Anderson & Hester / Seattle Times cite: Seattle Times

Massey cite: Massey

Matthews cite: Matthews

The New York Times cite: NYT

Rothman cite: Rothman (public)

Sagarin cite: sagarin

Of all these systems only Rothman offers the code to others. However, Massey has a fairly detailed description of his system so it should be possible to recreate his rankings. Since all the other systems seem to be well guarded secrets it is not possible for others to check the calculations, or try to estimate what effect wins or losses in the next weeks will have. It is especially problematic in the case of Jeff Sagarin, who is making a living of publishing ratings for various sports in USA Today. Since Mr. Sagarin has an economical advantage of keeping his system proprietary, it will be very difficult to obtain a detailed description or the actual source text of his system.

This secrecy hurts the computer rankings tremendously, since they create the impression that they are un-understandable and unfair and rewards ”running up the score”.

It is also a very uncommon situation. In other sports, where rankings play an important role (chess, golf, tennis, etc.) the ranking method is completely public and can be checked by any interested person.

I will therefore suggest, that *the current system is replaced with an open system based on publicly accessible code with a detailed mathematical description*. Ideally the whole package should be available for download from the BCS WWW site, so coaches, players, journalists and fans can perform the rankings themselves. The criteria the teams should be ranked on should be well published and should be quantifiable in terms of either simple formulas or tables.

For this reason I have chosen to give a detailed report on the ranking system I use. As will be seen, it is based on the work of others. It does, however, contain a few original, minor ideas.

I will also propose, that the NCAA should be responsible for the WWW posting of all NCAA results in a standard ASCII based format, that can be used by everybody ranking sports teams.

An overview of ranking methods

Specialized Tournament Systems

Often a ranking between teams is obtained by letting them play in tournaments with the purpose of either establishing a ranking between them or at least define ”the best team”. This tournament can either be stretched over the whole season or, more commonly, is used in a playoff at the end of the season.

Round Robin System

All teams in the tournament play each other and the ranking is determined by the number of points each team accumulates (see the section on accumulative point systems). This is a very ”fair” system, but requires a large amount of games.

Cup System

The tournament is played in ”rounds”. Only winners are allowed to play in the next round. Eventually only one team is left and is declared the winner. The theme can be varied by introducing a losers bracket, so a team is only eliminated after two loses. This is a very popular system, since it will find an undisputed ”best” team using the least amount of games. However, due to the unavoidable fluctuations in performance of each team from game to game it happens very often, that a better teams looses a match to an inferior opponent and is then eliminated. This is not fair from a ranking point of view, but has a lot of appeal from a spectator point of view.

Monrad System

The Monrad system is a very interesting variation of the cup system, which to my knowledge is only used on a regular basis in chess tournaments. In the first round all teams are paired randomly. The winner gets 1 point and the looser zero. In each successive round all teams with the same number of points are paired randomly (except that teams which earlier have played each other can not be paired if there are other pairing possibilities). This system has the advantage, that all teams keep playing, in contrast to the cup system, and as the season (or tournament) advances teams with equal strength will be meeting each other. There are no limitations to the number of rounds that can be played, but eventually teams have to be paired if they have similar, but not necessarily identical, number of points. The team with the largest number of points after a predefined set of rounds is the winner.

This system would be ideally suited for NCAA football, if only tradition and logistics would not interfere. Early in the season teams would play 5-6 Monrad games within their conference and the rest of the season they would be paired nationally, but with a pairing preference for teams geographically close. This would result in some very exciting games in November and would create optimally matched bowl games.

Accumulative Point Systems

Most sports use ranking systems based on the idea, that for each match or tournament the team (or player) acquires a certain number of points depending on their performance and the teams ranking is based on the total number of point they have accumulated during the season. If several teams have the same number of points additional objective criteria are used, like winner of mutual game or accumulated score difference.

In most soccer leagues the winning team gets 2 (or 3) points and the looser none and each team gets 1 point for a tie. If all teams play each other (round robin as described above) this provides a very simple and effective method for evaluating the integrated performance of each team. Typical sizes of leagues range from 6 to 24. Usually the best teams from a league will move up in a higher league next season and the lowest ranked teams will be moved to a lower league. The winner of the highest league will be the overall winner.

Golf and tennis are using systems where each player accumulates points based on their placement in each tournament according to published tables. Prestigious tournaments will provide more points and small local tournament will of course provide less points. Usually the points are accumulated over a sliding time interval of one year.

The advantage of the accumulative point system is their simplicity. It is easy for each team or the spectators to figure out their accumulated score and therefore their ranking. All participants know what they gain or loose by winning or loosing a game. The disadvantage is, that the point allocation, especially in tournaments for large systems like in golf and tennis, becomes somewhat arbitrary and not based on the actual strength of the participants. It is also sometimes possible to rake up a lot of points by carefully selecting weak tournaments or by playing a large amount of tournaments.

Elo Systems

The Elo rating system was first used by the International Chess Federation in 1970 to rank chess players. The system was proposed by Arpad E. Elo. It is partly based on earlier work done by Anton Hoesslinger. The official description of the system as it is used in chess can be found at cite: EloUSCF . Jones has given a nice overview of the system in cite: RoyJones .

The basic idea in the system is to continuously change a players rating R_p based on whether she performs better or worse than expected in tournaments or matches. For a new player with a total of N matches, where $N \leq N_{cut}$ $N_{cut} = 20$ the rating R_p is calculated as

$$R_p = \langle R_c \rangle + \alpha \frac{N_w - N_l}{N}$$

where $\langle R_c \rangle$ is the arithmetic average of the competitor's ratings at the time of the match, N_w and N_l is the number of wins and losses, respectively, and $\alpha = 400$ is an initial scale factor. This is the basic Ingo system named after the place of origin, Ingolstadt in Germany, of it's inventor Anton Hoesslinger.

Elo's important improvement to this system was to introduce the *Win Expectancy Function* W_e, which is defined as

$$W_e(\Delta R = \frac{1}{1 + 10^{-\frac{\Delta R}{\alpha}}}$$

where

$$\Delta R = R_p - R_c$$

For a tournament with M matches played by a player with a rating R_p the new ranking $R_{p,new}$ after the tournament becomes

$$R_{p,new} = R_p + K(S - \Sigma_{i=1}^{M} W_{e,i}(\Delta R_i$$

where the score is defined as (N_t is the number of tied games)

$$S = N_w + \frac{1}{2}N_t$$

and the sum i runs over each of the games the player played in the tournament. K is in principle a constant, but is in reality varied slightly depending on the rating of the player.

The Elo system seem especially well suited to sports with a large number of participants (> 10,000), where methods based on linear algebra have problems due to memory limitations in computers. However, the idea of introducing the probability function W_e is very powerful and can be used in other ranking system. In particular Massey has used part of these ideas in his very interesting BCS ranking system.

Global Optimization Systems

Ordinal Ranking

Select a ranking that minimizes the number of violations. A violation is a game where a team with a lower ranking defeats a team with a higher ranking.

.....

The Ranking Model

Let us consider a set of teams T consisting of N_T teams playing a total of N_G games between each other. Depending on the nature of the sport the term ”team” can refer to either an individual (chess, boxing, singles tennis etc.) or a set of individuals (football, baseball, basketball, doubles tennis etc.). We will only consider games consisting of a set of two teams, but the method outlined in this paper can easily be generalized to consider games consisting of $n > 2$ teams (common in track and field, swimming etc.).

In game g ($g = 1,...,N_G$) the home team is denoted t_h ($h = 1,...,N_T$) and the away team is t_a ($a = 1,...,N_T$). In this game the home team obtains a score of S_h and the away team a score of S_a. The game is played at time T_g within a given season y. Results of games are assumed to be available from a total of N_y seasons ($y = 1,...,N_y$). The score of the winner is S_w and the score of the looser is S_l. We will assume for simplicity that the winner of a game is the team with the larger score. The margin of victory or point spread ΔS can be defined in two different ways.

$$\Delta S_{ha} = S_h - S_a$$

or

$$\Delta S_{wl} = S_w - S_l$$

ΔS_{wl} will always be non-negative, whereas ΔS_{ha} will be positive if the home team wins and negative if the away team wins. The relation between them is

$$\Delta S_{ha} = \begin{cases} \Delta S_{wl} & \text{if } S_h \geq S_a \\ -\Delta S_{wl} & \text{if } S_h < S_a \end{cases}$$

Basic Model Assumptions

The assumptions of the current ranking model are:

1. Only games played between two teams t_h and t_a in the set T are considered ($t_h, t_a \in \mathsf{T}$).
2. Only games played within a given season y are considered, except for rankings performed early in the season, where results of games from season $y - 1$ can be used.
3. The outcome or result $R_g(S_h, S_a$ of game g is a real function depending only on the final scores S_h and S_a.
4. The result R_g does *not* depend on the time of the game T_g within the season nor on any other variable related to the game.
5. The ranking of the teams in the set will be accomplished by allocating the i'th team t_i a strength or power rating r_i, where $r_i, i = 1,...,N_T$ is a set of real numbers . The teams will then be ranked (= ordered) according to the value of their strength.
6. The result R_g of game g is a measurement of the strengths r_h and r_a of the two teams with an associated measurement error σ_g.

In the following the discussion will be based on examples from football, but the formalism is completely general and could be used for any binary game resulting in a final set of scores. It follows from assumptions 4, that the current ranking method does not take into account other "unofficial" statistics from a game, like half time score, yardage gained or lost, fumbles etc. While these variables might very well be of importance for a prediction algorithm, we consider them irrelevant and unfair to use for a ranking algorithm, which purpose is to evaluate which team is the best or which set of teams

are the best. In this latter case the teams need to know exactly what they are being evaluated on.

The Game Outcome Function

Let us now consider in more detail the result R_g of game g. As stated in assumption 3 the result $R_g(S_h, S_a)$ is a real function of the two scores S_h and S_a. R_g will be considered a measurement of the strength of the two teams. As usual with any physical measurement the result R_g does not provide an exact measurement of the strengths of the two team, but an uncertainty σ_g is associated with each measurement g. There is, unfortunately, not a universally accepted result function. Some commonly used functions are discussed below.

Win-Loss system (WL)

It only matters which team wins (= which team has the higher score), but the actual scores do not matter.

$$R_g^{WL}(S_h, S_a) = \begin{cases} 1 \text{ if } S_h > S_a \\ 0 \text{ if } S_h = S_a \\ -1 \text{ if } S_h < S_a \end{cases}$$

This system is often referred to as the JWB system (Just Win Baby). Effectively this is how many fans view the game outcome, since the only thing that matters is whether you win or not. Not how you win or by how much.

Score Difference system (SD)

The result of the game is defined as the difference between the scores of the two teams:

$$R_g^{SD}(S_h, S_a) = S_h - S_a = \Delta S_{ha}$$

In football this system is often referred to as the BOMB Index (Bowden - Osborne Memorial Blowout Index), since it is perceived, that Florida State and Nebraska used to run up the score against weak opponents, which will help their ranking in this system.

Truncated Score Difference system (TSD)

A modification the SD system, where the score difference is truncated at some value $\Delta S_{\max}$ in order to avoid to heavy an emphasis on games, where one team ”runs up” the score:

$$R_g^{TSD}(S_h, S_a|\Delta S_{\max}) \equiv \Delta S_t \equiv \begin{cases} \Delta S_{\max} & \text{if } \Delta S_{ha} \geq \Delta S_{\max} \\ \Delta S_{ha} & \text{if } |\Delta S_{ha}| < \Delta S_{\max} \\ -\Delta S_{\max} & \text{if } \Delta S_{ha} \leq -\Delta S_{\max} \end{cases}$$

In football typical values of maximum point spread $\Delta S_{\max}$ ranges from 21-35. ΔS_t is called the truncated point spread. Please note that the game outcome now also depends on the game-independent parameter $\Delta S_{\max}$.

A mathematically more elegant way of providing this cutoff is to use the hyperbolic tangent function

$$R_g^{TSDT}(S_h, S_a|\Delta S_{\max}) = \Delta S_{\max} \tanh\left(\frac{\Delta S_{ha}}{\Delta S_{\max}}\right)$$

with

$$R_g^{TSDT} \simeq \Delta S_{ha} \quad \text{for } |\Delta S_{ha}| \ll \Delta S_{\max}$$

and

$$R_g^{TSDT} \to \Delta S_{\max} \quad \text{for } \Delta S_{ha} \to \infty$$

Simple Hybrid WL-SD system

Since both the pure WL and SD system tend to favor a particular, but maybe extreme, view of the game outcome, other systems have introduced a simple linear combination of the two systems:

$$R_g^{WLSD}(S_h, S_a|B_w) = \begin{cases} B_w + \Delta S_{ha} & \text{if } S_h > S_a \\ 0 & \text{if } S_h = S_a \\ -B_w + \Delta S_{ha} & \text{if } S_h < S_a \end{cases}$$

In football typical values for the ”bonus” B_w for a win is 50-100. For $B_w = 0$ the system reduces to the SD system and for $B_w \gg |\Delta S|$ it is identical to the WL system.

Score Ratio system (SR)

Instead of forming the difference between the score, as in the SD system, the ratio between the scores is the game outcome.

$$R_g^{SR}(S_h, S_a) = \frac{\Delta S_{ha}}{S_w}$$

In this case $-1 \leq R_g^{SR} \leq 1$, with $|R_g^{SR}| = 1$ if the loosing team does not score any points (a shot-out).

The rare case of $S_h = S_a = 0$ is not defined in this system and it can therefore not be used in sports, where this result is possible.

Linear Win - Difference - Ratio system (LWDR)

All the above mentioned methods for evaluating the outcome of a game have their virtues. It is therefore natural to form a outcome function, that combines all of them. The simplest way of doing this is by forming a linear combination of the three types of outcome:

$$R_g^{LWDR}(S_h, S_a | B_w, \Delta S_{\max}, B_r) = \begin{cases} B_w + \Delta S_t + B_r \dfrac{\Delta S_{ha}}{S_w} & \text{if } S_h > S_a \\ 0 & \text{if } S_h = S_a \\ -B_w + \Delta S_t + B_r \dfrac{\Delta S_{ha}}{S_w} & \text{if } S_h < S_a \end{cases}$$

There are three parameters in this approach: 1) the win bonus B_w, 2) the maximum point spread $\Delta S_{\max}$, and 3) the scoring ratio weight factor B_r. They are, however, not independent since a scaling of all of them will lead to the same ranking. Since the sum of the three parameters is equal to the maximum value of the game outcome function $R_g^{\max}$, it is convenient to constrain the three parameters by choosing $R_g^{\max}$ to have a convenient value like 100.

$$B_w + \Delta S_{\max} + B_r = R_{\max}$$

In this paper we will choose the following values

$$B_w = 50 \quad \Delta S_{\max} = 25 \quad B_r = 25$$

A subsequent paper will discuss techniques for choosing the most optimal values for $(B_w, \Delta S_{\max}, B_r)$.

The Outcome Prediction Function

The outcome prediction function $P_g(r_h, r_a | \alpha_j)$ estimates the outcome of a game g between the two teams t_h and t_a based on their strength rating r_h and r_a, respectively. In addition P can depend on a number of additional parameters, depending on the choice of game outcome function, R_g. In the case the WDR game outcome function R_g^{WDR} is used the parameter vector is $\vec{\alpha} = (B_w, \Delta S_{\max}, B_r)$. In addition other parameters, like the home field advantage B_h, can be added to the parameter set, if needed. Actual choices of P will be discussed later in this paper, but first we will outline the method for determining the strength ratings, r_i, independent of the functional expression of P.

During a season with N_g games a total of N_g measurements of the N_t strength ratings r_i is performed and we want to determine the vector $\vec{r}$ so the difference between the game result R_g and the prediction (or hypothesis) P_g is as small as possible for as many games as possible. This can be done in a number of ways. We choose to minimize the sum of the square of the difference between the game result and the prediction. This is the so-called Least Squares Method.

minimize: $$\chi^2(\vec{r}) = \sum_{g=1}^{N_g} \left[\frac{R_g(\vec{\alpha}) - P_g(\vec{r} \mid \vec{\alpha})}{\sigma_g} \right]^2$$

The measurement error σ_g for each game will be discussed later.

Other methods are based on the maximum norm .....

Linear chi-square method

If we furthermore assume, that the outcome prediction function depends *linearly* on the strength ratings, r_i, we can use the general framework for linear least squares fits. This is a very strong assumption and a later paper will discuss non-linear approaches. We will furthermore assume P_g does not depend on $\vec{\alpha}$, but only depends on the relative difference between the strength ratings of the two teams participating in the game

$$P_g = r_h - r_a = \sum_{t=1}^{N_t} \delta_{gt} r_t \quad \text{where} \quad \delta_{gt} = \begin{cases} 1 & \text{if } t = h \\ -1 & \text{if } t = a \\ 0 & \text{if } t \neq h, a \end{cases}$$

This reduces the problem to the minimization of

$$\chi^2(\vec{r}) = \sum_{g=1}^{N_g} \left[\frac{R_g(\vec{\alpha}) - \sum_{t=1}^{N_t} \delta_{gt} r_t}{\sigma_g} \right]^2$$

Following the standard χ^2 nomenclature we introduce the design matrix **A** with elements

$$A_{gt} = \frac{\delta_{gt}}{\sigma_g}$$

the weighted result vector **b** with elements

$$b_g = \frac{R_g}{\sigma_g}$$

and the strength parameter vector **r**. In matrix notation the problem can now be written as

$$\chi^2 = |\mathbf{A} \bullet \mathbf{r} - \mathbf{b}|^2$$

where the $|\quad|^2$ symbol indicates the Euclidean norm in the vector space spanned by all the games.

Single Value Decomposition Solution

Using the Single Value Decomposition algorithm the vector **r**, that minimizes χ^2 is

$$r_t = \sum_{g=1}^{N_g} \left(\frac{U_{tg}\ b_g}{v_g} \right) V_{tg}$$

where **U**, **V** are the SVD matrices and **v** is the single value vector defined as

$$A_{gt} = \sum_{i=1}^{Nt} v_g U_{gi} V_{ti}$$

or

$$\mathbf{A} = \mathbf{U} \bullet [\mathrm{diag}(v_g)] \bullet \mathbf{V}^T$$

We obtain the **U** and **V** matrices and the **v** vector using the routines described in Press et al. [1992].

The advantage of using the linear χ^2 method is speed, since it only takes a few seconds on a normal PC to solve for the rankings. The SVD algorithm is the standard tool for solving linear χ^2 problems due to its robustness. For further details see press et al. [1992].

The Game Weight Factors

The game weight factors, defined as

$$w_g = \frac{1}{\sigma_g^2}$$

provide the option to let various games influence the rankings differently. It is very common in other ranking models to make the weights time-dependent

$$w_g(t) \propto \exp(T - t_g)$$

where T is the time where the ranking is performed. This will put more emphasis on games played recently. This method is, however, more appropriate for prediction model than for ranking models. The current ranking model does not incorporate any time-dependence in the weights, with the exception of the initial period as explained later.

If two teams of very similar strength are playing each other, we consider the outcome of this game a ”good” measurement of the relative strengths of the two teams. In contrast we consider mis-matched games between opponents with very different strengths as ”bad” measurements. In mathematical terms this implies, that we will associate a larger weight w_g (or equivalently a smaller uncertainty σ_g) to games between teams where the difference between the rankings r_h and r_a is small. This can, however, only be done in an iterative procedure, since the rankings r_h and r_a are not initially. We are

therefore using the following method:

Initial values of the ranking vector $\mathbf{r}^{1}$ are determined by assuming $w_g^{(1} = 1$. Afterwards the χ^2 system is solved again, but this time with the following weights

$$w_g^{(2} = \exp\left(-\frac{|r_h - r_a|}{\alpha_w}\right)$$

where α_w is determined from the condition

$$\exp\left(-\frac{\max\{r_t - \min\{r_t}{\alpha_w}\right) = \frac{1}{\beta_w^2}$$

In other words, we consider a game between the best and the worst team to have an measurement error β_w times larger than when two completely equal teams play. β_w is a free parameter in this ranking model. The numerical examples show later have used the value $\beta_w = 3$.

Early season rankings

Early in the season, when only few games have been played, the design matrix **A** has a rank smaller than N_t. This has the effect, that the relative ranking of many sub-sets of teams are undetermined. For NCAA 1-A division football teams, where $N_t = 112$, the number of independent sub-sets of teams as function of number of the number of weeks played is shown in the table below (based on the 1996-98 seasons):

Number of independent sets

week	1996	1997	1998
1	109	104	104
2	78	78	65
3	36	36	18
4	3	2	1
5	1	1	1

Only after 5 playing weeks will it be possible to obtain a solution, where all teams are ”connected”. For NCAA 1-A football the onset of full connectiveness seems empirically to coincide with a condition, that the average number of games per team is 2.5 - 3.0 or a total of 139 to 170 games between the 112 teams (only games where both teams are division 1-A counts).

It is, however, often required to obtain a ranking early in the season before the full connectedness haven been obtained. Since we are requiring, that only the final results of games can be used in the ranking this can only be obtained by using game results from the previous season(s). So early in the season results from the previous season are also included in the ranking, but with a decreasing weight factor. The total weight factor on each game is

$$w_{tot} = w_g w_s$$

where w_g is defined above and

$$w_s = \begin{cases} 1 & \text{for games in the current season} \\ \max\left[0, \left(1 - \frac{\langle N_{gt} \rangle}{R_{cut}} \right)^2 \right] & \text{for games in the previous season} \end{cases}$$

where $\langle N_{gt} \rangle$ is the average number of games played by each team and $R_{cut} = 4.5$ represents a cut-off value for the inclusion of previous seasons games.

Input

Currently it is cumbersome to obtain NCAA results, especially for the lower divisions. I would therefore like to propose, that NCAA post results of all NCAA games (football, basketball, tennis, etc.) in a standard ASCII based format, that can be used directly a input to ranking programs.

For each sport two files should be created for each season: a team file and a game file.

Team File

The Team file should contain the following information on each team:

1. *Team index*. A unique integer index running from 0
2. *Official Team Name*. (Tennessee, James Madison, etc.)
3. *Conference Name*. (The Southeastern Conference, The Atlantic Ten Conference, etc.)
4. *Division*. (1-A, 1-AA, etc.)

The format of the file should be (generic C printf statement):

```
fprintf( TeamFile, ” %5d %-25s %-35s %-25s\n”,
   Index, Name, Conference, Division );
```

Game File

The Game file should contain the following information on each game:

1. Date
2. Away team name
3. Away team score
4. Home team name
5. Home team score
6. Away team index
7. Home team index
8. Playing field code (= 1 if home team was home (default), = 2 if the game was a bowl or play-off

game, and = 3 if game for some other reason was played on a neutral field)

The format of the file should be (generic C printf statement):

```
fprintf( GameFile,
   ” %4d %2d %2d %-25s %3d - %-25s %3d |%5d %5d %1d\n”
   Year, Month, Day,
   AwayTeamName, AwayTeamScore,
   HomeTeamName, HomeTeamScore,
   AwayTeamIndex, HomeTeamIndex, FieldCode );
```

Summary

References

BCS The official web site for the BCS is: http://www.abccfb.com/. The current rules for the BCS are documented at http://www.cae.wisc.edu/~dwilson/rsfc/rate/newbcsrelease.html

Billingsley The description of the Billingsley system can be found at: http://www.cfrc.com/html/searchof.htm

Dunkel No current, active links to a description of the Dunkel system could be found.

Seattle Times The description of the Anderson & Hester / Seattle Times system can be found at: http://ww1.sportsline.com/u/football/college/polls/seattletimes/

Massey http://www.mratings.com/theory/massey.htm

Matthews http://www.expertpicks.com/

NYT

Rothman http://www.cae.wisc.edu/~dwilson/rsfc/rate/rothman.html

sagarin http://www.usatoday.com/sports/football/sfc/sfcjsx.htm

EloUSCF http://www.uschess.org/ratings/info/system.html

RoyJones http://www.brainiac.com/royjones/