A Consistent Weighted Ranking Scheme with an Application to NCAA College Football Rankings

Itay Fainmesser, Chaim Fershtman and Neil Gandal[1]

March 20, 2007

Abstract

The NCAA college football ranking, in which the "so-called" national champion is determined, has been plagued by controversies the last few years. The difficulty arises because there is a need to make a complete ranking of teams even though each team has a different schedule of games with a different set of opponents. A similar problem arises whenever one wants to establish a ranking of patents or academic journals, etc. This paper develops a simple consistent weighted ranking (CWR) scheme in which the importance of (weights on) every success and failure are endogenously determined by the ranking procedure. This consistency requirement does not uniquely determine the ranking, as the ranking also depends on a set of parameters relevant for each problem. For sports rankings, the parameters reflect the importance of winning vs. losing, the strength of schedule and the relative importance of home vs. away games. Rather than assign exogenous values to these parameters, we estimate them as part of the ranking procedure. The NCAA college football has a special structure that enables the evaluation of each ranking scheme and hence, the estimation of the parameters. Each season is essentially divided into two parts: the regular season and the post season bowl games. If a ranking scheme is accurate it should correctly predict a relatively large number of the bowl game outcomes. We use this structure to estimate the four parameters of our ranking function using "historical" data from the 1999-2003 seasons. Finally we use the parameters that were estimated using the historical data (1999-2003) and the outcome of the 2004 regular season to rank the teams for 2004. We then calculate the number of bowl games whose outcomes were correctly predicted following the 2004 season. Our methodology predicted more bowl games correctly in 2004 than any of the six computer ranking schemes used by the BCS.

[1] Fainmesser: Harvard University, ifainmesser@hbs.edu. Fershtman: Tel Aviv University, Erasmus University Rotterdam, and CEPR, fersht@post.tau.ac.il. Gandal: Tel Aviv University, and CEPR, gandal@post.tau.ac.il. We are grateful to Drew Fudenberg for helpful suggestions.

1. Introduction

At the end of the regular season, the two top NCAA college football teams in the Bowl Championship Series (BCS) rankings play for the "so-called" national championship. Nevertheless, the 2003 college football season ended in a controversy and two national champions: LSU and USC. At the end of the 2003 regular season Oklahoma, LSU and USC all had a single loss. Although both the Associated Press (AP) poll of writers and ESPN/USA Today poll of football coaches ranked USC #1, the computer ratings were such that USC ended up #3 in the official BCS rankings; hence LSU and Oklahoma played in the BCS "championship game." Although LSU beat Oklahoma in the championship game, USC (which won its bowl game against #4 Michigan) was still ranked #1 in the final (post bowl) AP poll.[2] The "disagreement" between the polls and the computer rankings following the 2003 college football season led to a modification of the BCS rankings that reduced the weight of the computer rankings.

Why is there more controversy in the ranking of NCAA college football teams than there is in the ranking of other sports' teams? Unlike other sport leagues, in which the champion is either determined by a playoff system or a structure in which all teams play each other (European Soccer Leagues for example), in NCAA college football, teams typically play only twelve-thirteen games and yet, there are 117 teams in (the premier) Division I-A NCAA college football.

Hence, controversies arise because there is a need to make a complete ranking of teams even though there is an "incomplete interaction"; each team has a different schedule of games with a different set of opponents. In a setting in which each team plays against a small subset of the other teams and when teams potentially play a different number of games, ranking the whole group is nontrivial. If we just add up the wins and losses, we obtain a partial (and potentially distorted) measure. Some teams may play primarily against strong teams while others may play primarily against weak opponents. Clearly wins against high-quality teams cannot be counted the same as wins against weak

[2] By agreement, coaches who vote in the ESPN/USAToday poll are supposed to rank the winner of the BCS championship game as the #1 team. Hence LSU was ranked #1 in the final ESPN/USA Today poll.

opponents. Moreover such a measure will create an incentive problem as each team would prefer to play easy opponents.

Similar ranking issues arise whenever one wants to establish ranking of scholars, academic journals, articles, patents, etc.[3] In these settings, the raw data for the complete ranking are bilateral citations or interactions between objects, or individuals. In the case of citations, it would likely be preferable to employ some weighting function that captures the importance of the citing articles or patents. For example, weighing each citation by the importance of the citing article (or journal) might produce a better ranking. Such a methodology is analogous to taking into account the strength of the opponents in a sports setting.

The weights in the ranking function can be given exogenously, for example when there is a known "journal impact factor" or a previous (i.e., preseason) ranking of teams. Like pre-season sport rankings, journal impact factors are widely available. The problem is that the resulting ranking functions use "exogenous" weights. Ideally, the weight or importance of each game or citation should be "endogenously" determined by the ranking procedure itself. A *consistent* ranking requires that the outcome of the ranking be identical to the weights that were used to form the ranking. This consistency requirement was first employed by Liebowitz and Palmer (1984) when they constructed their academic journal ranking. See also Palacios-Huerta, I., and O. Volij (2004) for an axiomatic approach for determining intellectual influence and in particular academic journal ranking. Their invariant ranking also satisfies the consistency requirement.[4] Finally, the consistency requirement is related to the methodology that the Google search

[3] Citations counts, typically using the Web of Science and/or Google Scholar, are increasingly used in academia in tenure and promotion decisions. Citations counts, typically using the Web of Science and/or Google Scholar, are increasingly used in academia in tenure and promotion decisions. The importance of citations in examining patents is discussed in Hall, Jaffe and Trajtenberg (2000) who find that "citation weighed patent stocks" are more highly correlated with firm market value than patent stocks themselves. The role of judicial citations in the legal profession is considered by Posner (2000).

[4] The consistent weighted ranking we develop can also be interpreted as a measure of centrality in a network. Centrality in networks is an important issue both in sociology and in economics. Our measure is a variant of an important measure of centrality suggested by Bonacich (1985). Ballester, Calco-Armengol, and Zenou (2006) have shown that the Bonacich centrality measure has significant impact on equilibrium actions in games involving networks.

engine uses to rank WebPages. "Google interprets a link from page A to page B as a vote, by page A, for page B. But, Google looks at more than the sheer volume of votes, or links a page receives; it also analyzes the page that casts the vote. Votes cast by pages that are themselves "important" weigh more heavily and help to make other pages "important"."[5]

In the case of patents or journals articles, the problem is relatively simple: either there is a citation or there is no citation. The problem is more complex in the case of sports rankings. The outcomes of a game include the result - winning, losing, not playing, and in some cases, the possibility of a tie. Additionally, it is important to take into account the location of the game, since there is often a "home field" advantage. An analogy for wins and losses also exists for the case of academic papers. One could in principle use data on rejections and not just publications in formulating the ranking. A rejection would be equivalent to losing and would be treated differently than "not playing" (i.e., or not submitted).[6]

This paper presents a simple consistent weighted ranking (CWR) scheme to rank agents or objects in such interactions and applies it to NCAA division 1-A college football. The ranking function we develop has four parameters: the value of wins relative to losses, a measure that captures the strength of the schedule, and measures for the relative importance of "home vs. away" wins and "home vs. away" losses. Rather than assign exogenous values to these parameters, we estimate them as part of the ranking procedure.

In most ranking problems, there are not explicit criteria to evaluate the success of proposed rankings. NCAA college football has a special structure that enables the evaluation of each ranking scheme. Each season is essentially divided into two parts: the regular season and the post season bowl games. We estimate the four parameters of our ranking function using "historical" data from the regular season games from 1999-2003. The regular season rankings associated with each set of parameter estimates is then

[5] Quote appears at http://www.google.com/technology/.
[6] A paper that was accepted by the RAND Journal of Economics without ever being rejected would be treated differently than a paper that was rejected by several other journals before it was accepted by the RAND Journal. But this is, of course, a hypothetical example since such data are not publicly available.

evaluated by using the outcomes of the bowl games for those five years. For each vector of parameters, the procedure uses the regular season outcomes to form a ranking among the teams for each season. If a ranking is accurate it should correctly predict a relatively large number of bowl game outcomes. Our methodology is such that the optimal parameter estimates give rise to the best overall score in bowl games over the 1999-2003 period.

Our estimated parameters suggest the "loss penalty" from losing to a very highly rated team is just 1/3 the "loss penalty" of losing to a team with a very low rating. Hence, our estimates suggest that it indeed matters to whom one loses: the strength of the schedule is important in determining the ranking. Further, our estimates are such that a team is heavily penalized for a home loss, relative to a road loss, while a home win is rewarded only slightly less than a win "on the road."

The wealth of information and rankings available on the Internet suggests that the rating of college football teams attracts a great deal of attention.[7] There are, however, just six computer ranking schemes that are employed by the BCS. Comparing the CWR ranking to these six rankings indicates that over a five year period, the CWR ranking did approximately 12-15 percent better (in predicting correct outcomes) than the other ratings. This comparison is, of course, somewhat unfair, because our optimization methodology chose the parameters that led to the highest number of correctly predicted bowl games during the 1999-2003 period.

Finally we use the 2004 season, which was not used in estimating the parameters of the ranking, and perform a simple test. Using the estimated parameters, we employ the CWR and the outcome of the 2004 regular season in order to determine the ranking of the teams for 2004 season. We then evaluate our ranking scheme by using it to predict the outcome of the 2004 post season (bowl) games. Our CWR ranking scheme predicted more bowl game outcomes correctly than any of the computer rankings used in the BCS rankings for

[7] See http://homepages.cae.wisc.edu/~dwilson/rsfc/rate/index.shtml for the numerous rankings. Fair and Oster (2002) compares the relative predictive power of the BCS ranking schemes.

2004 period. This is, of course, only a one year comparison and clearly not statistically significant evidence regarding the quality of the various rankings. On the other hand the forecasting ability of our CWR scheme should improve as more seasons (data) are included in the estimation stage.

2. The BCS Controversies

Unlike other sports, there is no playoff system in college football. Hence, it was not always easy for the coaches' and writers' polls to agree on a national champion or the overall ranking. The BCS rating system which employs both computer rankings and polls was first implemented in 1998 to address this issue and try to achieve a consensus national champion, as well as help choose the eight teams that play in the four premier (BCS) bowl games.[8] Nevertheless, the 2003 college football season ended in controversy and two national champions: LSU and USC. The polls rated USC #1 at the end of the regular season, but only one of the computer formulas included in the 2003 BCS rankings had USC among the top two teams. While all three teams had one loss, the computer rankings indicated that Oklahoma and LSU had played a stronger schedule than USC.

The disagreement between the polls and the computer rankings led to a modification of the method used to calculate the BCS rankings following the 2003 college football season. Up until that time, the computer rankings made up approximately 50 percent of the overall BSC ratings. The 2004 BCS rankings were based on the following three components, each with equal weights:[9] (I) The ESPN/USA Today poll of coaches, (II) The Associated Press poll of writers, (III) Six computer rankings. Hence, the weight placed on the computer rankings was demoted.[10]

Following the 2004 season, the BCS system again came under scrutiny. The complaint involved California (Cal) which appeared to be on the verge of its first Rose Bowl

[8] There are now five BCS bowl games.
[9] See http://www.bcsfootball.org/news.cfm?headline=40 for details.
[10] If the new system had been used during the 2003 season, LSU and USC would have played in the 2003 BCS championship game.

appearance since 1959. Despite Cal's victory in its final game, it fell from 4th to 5th in the final BCS standings and lost its place to Texas, which climbed to 4th, despite being idle the final weekend. Texas thus obtained the BCS' only at-large berth and an appearance in the Rose Bowl, and Cal lost its place in a BCS bowl game.[11]

The controversy was due to the changes in the polls over the last week of the season. In the BCS ranking released following the week ending November 27, Cal was ranked ahead of Texas. There were only a few games the following weekend. Cal played December 4 against Southern Mississippi because an earlier scheduled game between the teams had been rained out by a hurricane. Cal beat Southern Mississippi on the road 26-16,[12] while Texas did not play. Nevertheless, Cal fell and Texas gained in the AP and USA Today/ ESPN polls. The BCS computer ranking of the two teams was unchanged between the November 27 and December 4 period. If there had been no changes in the polls, Cal would have played in the Rose bowl. Given its drop to 5th, Cal ended up playing in a minor (non BCS) bowl.[13] Table 1 below summarizes the changes that occurred in the polls and computer rankings between November 27 and December 4.

In part because of the "Cal" controversy following the 2004 season, the AP announced that it would no longer allow its poll to be used in the BCS rankings and ESPN withdrew from the coaches' poll. Although the BCS eventually added another poll, a better solution might have been to give more importance to computer rankings. Despite the criticism of computer rankings, they are the only ones that can be transparent and based on measurable criteria, which is to say, impartial.

[11] This discussion should not be taken as a criticism of Texas. If the BCS had taken the top eight teams for its four bowl games that year, both Cal and Texas would have played in a BCS bowl game, perhaps against each other in the Rose Bowl.
[12] Southern Mississippi finished the regular season 6-5 and later won its bowl game.
[13] This had financial implications beyond the "pride" of competing in a top (BCS) bowl. Playing in a minor (non BCS) bowl typically means much smaller payouts for the schools involved. There are also claims that donations to universities increase and the demand for attending a university increases in the success of the football team. Frank (2004) finds no statistical support for this claim.

Games through	November 27	December 4	Actual Change (% change)
Polls			
Cal (AP)	1410	1399	-11 (-0.8%)
Texas (AP)	1325	1337	+12 (+0.9%)
Cal (ESPN/USA)	1314	1286	-27 (-2.2%)
Texas (ESPN/USA)	1266	1281	+15 (+1.2%)
BCS Computer Ranking: No change in California's and Texas' rankings			
Games: California 26 Southern Mississippi 16; Texas (idle)			

Table 1: Changes in Ratings between November 27 and December 4

3. The CWR Ranking Methodology

3.1 Development of a Consistent Ranking

We develop our formal ranking in three steps. We first consider a simple bilateral interaction like citations (cited articles or patent citations). This is relatively a simple case because either object *i* cites object *j* or it does not cite object *j*. We then consider a sports setting; in this case, there is a winner and a loser or no game. (In some sports settings, there is the possibility of a tie.[14]) In the final stage we incorporate the possibility of two types of games; home games and away games. This means that winning (or losing) a home game can have a different weight than winning (or losing) an away game.

Consider a group $N \equiv \{1,....,n\}$ of agents (or objects), with the relation $a_{ij} \in \{0,1\}$ for every $i, j \in N$. For example, N is a set of patents or articles, $a_{ij} = 1$ if patent or article j cites patent (or article) i and $a_{ij} = 0$ otherwise. Our dataset is hence uniquely defined by the matrix $A = \lfloor a_{ij} \rfloor$. We interpret each $a_{ij} = 1$ as a positive signal regarding object i. The objective is to define a rating function: $R : A \to R^n$ which generates a rating (and not just a ranking) for every agent that summarizes the information in A.

[14]In NCAA college football, a game tied at the end of regulation goes into overtime and the overtime continues until there is a winner.

There are many possible ways to define the function R; the most trivial (and commonly used) is the summation $r_i(A) = \sum_{j \neq i} a_{ij}$, $i = 1,...,n$, which is just a count; an example is the number of citations that each article receives. The advantage of such a ranking is its simplicity but it ignores much of the information embodied in A. Such a ranking may be appropriate when the "interactions" between the objects are not important; for example, when ranking bestsellers, a simple count of sales is probably appropriate. In other situations the identity or the "importance" of *j* should be taken into account when aggregating the a_{ij}. For example, in forming a ranking based on citations one may want to take into account the "importance" of the citing patent or article.

One possible resolution is achieved by using an exogenous weighting vector, describing the agents' "importance." Examples include "Journal Impact Factors" or the use of polls (or previous rankings) in college football. Letting m_j be agent's j subjective significance, we can normalize the count in the following way:

$$r_i(A,m) = \sum_{j \neq i} m_j a_{ij} \quad , \quad i = 1,...,n$$

However, this ranking function is not "consistent". The rating used to determine each agent's influence (m_j) differs from the final rating (r_j) of the agents. This "inconsistency" can be fixed by requiring that the weight given to each a_{ij} is identical to the rating itself (see Liebowitz and Palmer [1984]), i.e. the rating function z(A,z) should satisfy the following consistency requirement:

$$z_i(A,z) = \sum_{j \neq i} a_{ij} z_j \ .$$

To guarantee uniqueness, we can employ a simple normalization requiring, for example, that $\Sigma z_i = 1$ and $\min_{i=1,..,n} z_i = 0$. Specifically,

(1) $$z_i(A,z) = \frac{\sum_{j \neq i} a_{ij} z_j + g}{\sum_i \left(\sum_{j \neq i} a_{ij} z_j + g \right)} \ , \ i = 1,...,n \text{, where } \min_{i=1,..,n} z_i = 0 \text{,}$$

where g is endogenously determined in order to enable a solution to the system (i.e., it is determined by the condition $\min_{i=1,..,n} z_i = 0$). In order to solve system (1) we need to simultaneously determine the ratings of all agents (and g), since the ratings themselves are also the weights needed in the calculations.

Equation (1) is also related to Google's ranking of web pages (See Brin and Page (1998) and the Wikipedia entry on PageRank). In Google, the "page rank" value of webpage i is $z_i=(1-d)/N + d\sum_{j=1}^{N} l_{ij} z_j$, where N is the number of web pages, d is an exogenous constant, and l_{ij} = (1/# of outgoing links from webpage j) if webpage j links to webpage i, and 0 otherwise.[15] The "Google" normalization is that the sum of the page ranks equals one, i.e., $\sum_{i=1}^{N} l_{ij} = 1$.

3.2 Incorporating Wins and Losses

Our discussion up to this point considered the case when $a_{ij} \in \{0,1\}$. But in a sports match, the outcome can be win, lose, or do not play. Teams also might play more than one game against each other. To accommodate this we modify the ranking in the following way: For every $i, j \in N$, $a_{ij} \in Z^+$ indicates the number of times team i won against team j and $\bar{a}_{ij} \in Z^+$ indicates the number of times team i lost to team j, so the matrix $\bar{A} = [\bar{a}_{ij}]$ is added to the dataset and identifies losses while the matrix A is defined as above and identifies the wins.[16] Returning to the analogy of ranking articles, if it would have been feasible to use both acceptance and rejection data, the $\bar{A}$ matrix would be the "rejection" matrix.

---

[15] By definition, l_{ii} = (1/# of outgoing links from webpage i). The "damping factor," d, is typically set equal to 0.85.

[16] Note that for every i,j $\bar{a}_{ij} = a_{ji}$, therefore there is no necessity in defining the new matrix $\bar{A}$. However, it will make the presentation of the system of equations clearer, especially when we introduce further extensions.

As before, our objective is to define a consistent ranking function $R:\left\langle A,\overline{A}\right\rangle \to R^n$. Allowing for different coefficients for wins and losses, equation (1) now becomes:

$$(2)\qquad z_i\left(A,\overline{A},z\right)=\frac{\sum_{j\neq i}a_{ij}z_j-b\sum_{j\neq i}\overline{a}_{ij}\left(\gamma-z_j\right)+g}{\sum_i\left(\sum_{j\neq i}a_{ij}z_j-b\sum_{j\neq i}\overline{a}_{ij}\left(\gamma-z_j\right)+g\right)},\ i=1,...,n,\ \min_{i=1,..,n}z_i=0.$$

There are two new parameters in this ranking function; b and γ. These parameters account for the importance of losses relative to wins. As b and γ increase, the rating gives higher weight to losses. The parameter γ has an additional interpretation; keeping $b\cdot\gamma$ constant, a large γ means that our ranking function primarily depends on the number of losses, while a small γ implies that the ranking is sensitive to whom one loses. To insure that winning increases a team's rating and losing decreases a team's rating, it must be the case that $b>0$ and $\gamma>\max_i z_i$. Clearly different values of these parameters yield different ratings.

3.3 Home Field Advantage

In addition to the large set of possible outcomes, the location of the game may affect the outcome as well. Winning at "home" is easier than winning on the road. Since the location of the game is known, we can incorporate it in the ranking function by giving different weights to wins and losses at home and away games. This means that in addition to providing weights for the relative importance of wins vs. losses, weights must also be employed for the importance of "home games" vs. "away games". We split each matrix $A\left(\overline{A}\right)$, into home wins (losses) and away wins (losses). Thus, for every pair of teams $i,j\in N$, there are four relevant values $a_{ij}^{home},a_{ij}^{away},\overline{a}_{ij}^{home},\overline{a}_{ij}^{away}\in Z^+$ which (respectively) describe the number of times team *i* won at home, won away, lost at home,

and lost away, against team *j*. The four data matrices are: $A^{home}, A^{away}, \overline{A}^{home}, \overline{A}^{away}$ and we modify the ranking function as follows:

(3)

$$z_i\left(A^{home}, A^{away}, \overline{A}^{home}, \overline{A}^{away}, z\right) =$$

$$= \frac{\left[\sum_{j\neq i} a_{ij}^{away} z_j + h^w \sum_{j\neq i} a_{ij}^{home} z_j\right] - b\left[\sum_{j\neq i} \overline{a}_{ij}^{away}\left(\gamma - z_j\right) + h^l \sum_{j\neq i} \overline{a}_{ij}^{home}\left(\gamma - z_j\right)\right] + g}{\sum_i \left(\left[\sum_{j\neq i} a_{ij}^{away} z_j + h^w \sum_{j\neq i} a_{ij}^{home} z_j\right] - b\left[\sum_{j\neq i} \overline{a}_{ij}^{away}\left(\gamma - z_j\right) + h^l \sum_{j\neq i} \overline{a}_{ij}^{home}\left(\gamma - z_j\right)\right] + g\right)}$$

Again, $\min_{i=1,..,n} z_i = 0$.

Road wins and road losses are normalized to one. Hence the parameters h^w and h^l account for the weight of home wins (losses) relative to away wins (losses) in calculating the ratings. Again different values of these parameters yield different ratings. We do not assume any specific values of these parameters, but rather employ the unique data to estimate them.

4. Estimation and Evaluation of Ranking Parameters

Equation (3) is our ranking function, but it requires an input of four exogenous parameters: b, γ, h^w, and h^l. Determining the values of these parameters might be considered a task for football analysts. We clearly do not claim to possess such expertise. Instead, we propose to estimate these parameters using data from previous seasons.

The NCAA college football season is set up in a unique way that facilitates the evaluation of different ranking schemes. There are essentially two rounds in the college football season. In the first round, there are regular season games; in the second round, there are the so-called bowl games. Teams that play well during the regular season are invited to bowl games.

This setting provides us with a natural experiment to test the different ranking schemes. The regular season ranking determines the relative strength of the teams. The performance of each ranking can be evaluated by its implied prediction of the bowl game outcomes. If a ranking is reasonably good, then in a bowl game involving the #3 and #9 teams, the probability that the team ranked #3 wins the game should be more than 50%. We can thus use the results of the bowl games to evaluate the quality of the pre-bowl rankings or to estimate the relevant parameters.

Approximately 50% of the teams participate in bowl games Since we use these bowl games in estimating the parameters, our ranking may not be that accurate for the teams below the median and caution should be used when comparing the rankings of the lower ranked teams. But that does not pose a problem, since the ranking of the bottom half of barrel is much less important.

We use the 1999-2003 seasons to estimate the parameters: b, γ, h^w and h^l.[17] For a given set of parameters, we construct, for every year, a unique pre-bowl consistent rating. The second step is to examine the bowl games and determine which set of parameters provide the best prediction. There are clearly different ways to evaluate the performance of each rating system; we adopt for this paper a simple rule that selects the parameters that predict the highest number of bowl game results correctly over the five year period but we also discuss some alternative estimation methodologies.

For every set of parameters we assign a grade G(b, γ, h^w, h^l) which is defined by the number of bowl games (during the 1999-2003 period) predicted correctly by the ranking derived from these parameters. A correct prediction means that the winner of the bowl game is the higher ranked team at the end of the regular season. Fortunately bowl games are played at neutral sites (i.e., no home field advantage for either team) so the prediction of the outcome of the bowl games depends only on the teams' relative ranking.

[17] Some of the bowl games of the 2003 season, for example, take place in early January 2004. For ease of presentation we refer to them as games of the 2003 season.

Denote team "a_i" ("b_i") as the team that wins (loses) bowl game i. Formally, our estimation method minimizes the following function (over the N bowl games)

$$(4) \qquad \sum_{b \in N} \sum_{a \in \{a | a \in N, a \text{ beat } b\}} \left(1 - \phi(z_a, z_b)\right)^2 ,$$

where for each bowl game, $\phi(z_a, z_b) = 1$ if $z_a(b, \gamma, h^w, h^l) > z_b(b, \gamma, h^w, h^l)$ and $\phi(z_a, z_b) = 0$ otherwise. Our estimation methodology can be thought of as a nonlinear general method of moments (GMM) estimator, where the estimates are such that the distance between the data (the actual outcomes of the bowl games) and the model predictions are minimized.

Following the 1999 season there were 24 bowl games, following the 2000-2001 seasons there were 25 bowl games each year, while following the 2002-2003 seasons there were 28 bowl games each year. Thus the maximum overall score for the 1999-2003 period is 130, the number of bowl games during that period. We then sum up the number of correct predictions for the five years of bowl games associated with each set of parameter estimates. This gives us a grade, G(b, γ, h^w, h^l), for every set of parameters.

We first chose relatively broad intervals for the parameters in order to find areas which provided the best grade. The values chosen for the initial grid (see Table 3 below) were as follows: b which accounts for the importance of losses relative to wins was allowed to vary between 0.1 and 2.8. This means that the importance of losses relative to wins could vary between 10% and 280%. γ was allowed to vary between from 0.02 to 0.32. A γ of 0.32 is roughly 15 times the rating of the most highly ranked team; hence the range for γ is also very large. h^w and h^l were chosen to allow a large range as well.

	b	γ	h^w	h^l
Lower bound	0.1	0.02	0.1	0.1
Upper bound	2.8	0.32	2.8	2.8
Broad Grid intervals	0.3	0.05	0.3	0.3
Narrow Grid intervals	0.1	0.01	0.1	0.1

Table 3: Initial Grid and Intervals

Using the results from the initial grid, we changed and narrowed parameter range and increased the resolution around two distinct areas that yielded high grades.[18] The best predictions were given by two sets of parameters in two areas of the grid; these two distinct areas yielded 82 correct predictions over the five year period (out of a possible 130), or 63%. The two sets of parameters shown in Table 4 are at the center of the two regions with the highest scores:

Parameters	b	γ	h^w	h^l
Estimates Set 1	1.1	0.03	0.9	1.8
Estimates Set 2	1.9	0.03	2.6	1.6

Table 4: Optimal Parameter Estimates

In the first set of "optimal" parameter estimates, $h^w<1$ while $h^l>1$. The large difference between h^w and h^l suggests that a team is heavily penalized for a home loss, relative to a road loss (which is normalized to one) and that a home win is rewarded only slightly less than a road win (which is normalized to one). For this set of parameters, losing at home is a key factor in assessing a team's rating. When b is close to 1, wins and losses affect the ratings symmetrically. Hence, b=1.1 suggests that ratings are just slightly more sensitive to losses than wins.

In order to interpret γ, we need to know that the highest rating in 2004 was approximately 0.02. This means that other things being equal, the "loss penalty" from losing to a very highly rated team is γ - .02 = .01, which is 1/3 the "loss penalty" of losing to a team with a very low rating (γ - 0 = .03). Hence, the relatively low γ suggests that it indeed matters to whom one loses. (A relatively high γ implies that the ranking is more sensitive to the number of losses, rather than to whom one loses.)

[18] The search algorithm was written in Matlab. The algorithm and the complete set of results for the whole broad and narrow grids are available upon request.

In the second set of parameters, b and h^w are both higher, while γ and h^l are essentially unchanged relative to the first set of parameters. Hence in both cases, the estimated parameters suggest that a team should be heavily penalized for a home loss (h^l is relatively large) and it indeed matters to whom one loses (γ is relatively small).

The two different sets of parameters give similar results because of the substitutability among b and h^w. For example, as b rises from 1.1 to 1.9, more weight is given to losses relative to wins. This effect is offset in large part by a higher value of h^w (0.9 in the first set of parameters and 2.6 in the second set of parameters), which increases the importance of the home wins, relative to home losses.

In the appendix, we choose two relatively high and two relatively low values for each parameter in order to provide a sense as to the shape of the objective function. The "low" parameters employed in Table A1 in the appendix are b=0.7, γ=0.03, h^w=0.7, h^l=0.7, while the high parameters are b =2.2, γ=0.27, h^w=2.8, h^l=2.8. Table A1, which presents the results descending by grade, makes it clear that low values of γ and high values of h^l are critical for maximizing the number of correctly predicted games.

5. Alternative Estimation Methods

There are several possible ways to use the regular season ratings to forecast the bowl games results. In section 4, we employed a quite straightforward methodology; the estimated parameters were those that predicted the highest number of bowl game outcomes correctly. An alternative method is to use the rating (rather than the ranking) of two teams to predict the probability that team a beats team b in the bowl game. For example, if z_i, $i \in \{a,b\}$ is the rating of team i, then $\Pr\{a \text{ beats } b \mid z_a, z_b\} = \frac{z_a}{z_a + z_b}$. In order to evaluate the quality of a prediction of a given rating schedule for the bowl games,

one could then use a least squares method. The objective function to be minimized would then be $\sum_{b \in N} \sum_{a \in \{a | a \in N, a \text{ beat } b\}} \left(1 - \frac{z_a}{z_a + z_b}\right)^2$.

On one hand, this method uses more data than the method we chose since it exploits the whole cardinal rating rather than just the ordinal ranking that we used in the previous section. On the other hand, there is a fundamental problem with this methodology: when we use the rating itself to form $z_a/(z_a+z_b)$, the estimation method places more weight on bowl games involving lower ranked teams. This is because a given point spread in the rankings between two teams will yield a $z_a/(z_a+z_b)$ value much closer to ½ for the higher ranked teams than for teams lower in the ranking.

This problem indeed occurs in practice because closely ranked teams typically play each other in bowl games. When we employed the alterative estimation scheme, we obtained the following two sets of parameters estimates.

Parameters	b	γ	h^w	h^l
Estimates Set 1	0.7	0.03	2.8	2.8
Estimates Set 2	0.1	0.12	2.8	2.8

Table 5: Alternative Methodology: Parameter Estimates

If we look at the first set of parameter estimates, we see that the estimate of b is somewhat less than our preferred results. Thus, compared to our preferred estimates, these estimates place more weight on winning than losing. Additionally, the estimate of h^w is much higher. Since h^l is high as well, home games are much more important than "road" games. The parameter estimates are intuitive, since this (alternative) methodology places greater weight on the relatively weak teams, and these teams typically lose on the road. Hence, there is very little information available from road games.

In the second set of parameter estimates, the estimate of b is very small and the estimate of γ is very large. Hence the methodology in this case basically counts the number of wins. Again, the intuition is that weaker teams have fewer wins, so any win is very valuable, regardless of the opponent.

An analogous problem would arise if we would use the ranking (rather than the rating) to form $z_a/(z_a+z_b)$. Such a method places more emphasis on teams that finish near the top. For example, in a bowl game between the top two ranked teams, the "expected" probability that team number #1 will win in the methodology using the alternative ranking is $z_a/(z_a+z_b)=2/(2+1)=2/3$. On the other hand, in a game between teams ranked #15 and #16, the "expected" probability that team number #15 will win is $16/(16+15)=0.52$. This discussion suggests that our methodology is more attractive than the alternative methodology.

6. Evaluating the Performance of the CWR Ranking Methodology

Finally, we now compare our ranking methodology with the rankings of the experts. The six computer rankings included in the BCS rankings are:[19]

- AH- Anderson & Hester ratings (http://www.andersonsports.com/football/ACF_SOS.html),
- RB - Richard Billingsley ratings (http://www.cfrc.com/),
- CM - Colley Matrix ratings (http://www.colleyrankings.com/matrate.pdf),
- KM - Kenneth Massey ratings (http://www.mratings.com/rate/cf-m.htm),
- JS – Jeff Saragin ratings, (http://www.usatoday.com/sports/sagarin.htm),
- PW - Peter Wolfe ratings (http://www.bol.ucla.edu/~prwolfe/cfootball/ratings.htm).

In Table 6, we report the number of correct predictions for the CWR as well as the six BCS ranking schemes for the 1999-2003 bowl games. Table 6 shows that over a five year

[19] There are many other computer rankings in addition to the six used by the BCS. Massey, for example, includes 97 rankings on his comparison page. See, for example, the ratings comparison page at the end of the regular season in 2003, available at http://www.masseyratings.com/cf/compare2003-15.htm.

period, the CWR rankings do approximately 12-15 percent better (in predicting correct outcomes) than the other ratings for which we have complete data. This comparison is, of course, somewhat unfair, because our optimization methodology chose the parameters that led to the highest number of correctly predicted bowl games during the 1999-2003 period. Despite this caveat, the results suggest that there may be benefits from using historical data to estimate the parameters of ranking schemes.

Ranking	CWR 1	CWR 2	AH	CM	KM	RB	PW	JS
1999	15	17	14	12	14	NA[20]	NA	NA
2000	17	16	15	13	12	16	NA	NA
2001	15	16	14	14	15	12	NA	NA
2002	17	17	14	15	13	13	14	13
2003	18	16	15	15	19	21	17	19
Total 1999-2003	**82**	**82**	**72**	**71**	**73**	**NA**	**NA**	**NA**

Table 6: Bowl Games Predicted Correctly for the 1999-2003 Seasons[21]

Finally we use the 2004 season, which was not used in estimating the parameters of the ranking, and perform a simple test. Using the parameters that we estimated in section 4 and the outcome of the 2004 regular season, we rank the teams. We then calculate the number of bowl games whose outcomes were correctly predicted following the 2004 season and we compare our result with the number of correct predictions from the six computer ranking schemes employed in the BCS ranking. As Table 7 indicates our methodology predicted 15 or 16 out of the 28 bowl games in 2004 correctly, while the six computer ranking schemes used by the BCS predicted between 10-14 games correctly. We should add a word of caution here: while these results are interesting, they do not necessarily suggest any significant difference between our ranking schemes and those of the computer ranking schemes used by the BCS since the comparison is only for a single season. Nevertheless, our algorithm is constructed such that the parameters can be reevaluated every year with more data and the forecasting ability of our CWR scheme should improve as more seasons (data) are included in the estimation stage.

[20] NA= Data Not Available.
[21] CWR 1 refers to the first set of parameters discussed in section 4, while CWR 2 refers to the second set of parameters in that section.

Ranking	CWR 1	CWR 2	AH	CM	KM	RB	PW	JS
# of correct predictions	15	16	12	13	11	14	10	14
% of correct predictions	0.54	0.57	0.43	0.46	0.39	0.50	0.36	0.50

Table 7: Bowl Games Predicted Correctly for the 2004 Season

7. Concluding Remark:

The paper presents a consistent weighted rating scheme and showed how the results could be applied in developing useful rankings in sports settings. Our algorithm is such that the parameters can be reevaluated every year with more data. Hence, with more data we would expect the estimation to yield better predictions. While the focus of this paper is sport tournaments, a similar algorithm can be used for academic ranking of papers, journals or patents and may provide better insights than the commonly used citation counts.

References

Ballester, C., Calvo-Armengol, A., and Y. Zenou, 2006, “"Who's Who in Networks. Wanted: the Key Player, Econometrica, 74:1403-1417

Bonacich, P., 1987, “Power and Centrality: A Family of Mesaures,” The American Journal of Sociology, 92:1170-1182.

Brin, S., and L. Page, 1998, “The Anatomy of a Large-Scale Hypertextual WSEb Search Engine,” Standford University, mimeo

Fair R., and J. Oster, 2002 “Comparing the Predictive Information Content of College Football Ratings,” mimeo, available at http://papers.ssrn.com/sol3/papers.cfm?abstract_id=335801.

Frank, R., 2004 “Challenging the Myth: A Review of the Links Among College Athletic Success, Student Quality, and Donations,” Knight Foundation, executive summary available at http://www.knightfdn.org/default.asp?story=athletics/reports/2004_frankreport/summary.html.

Hall, B., Jaffe, A., and M. Trajtenberg, 2000, “Market Value and Patent Citations: A First Look,” NBER Working Paper W7741.

Liebowitz, S. and J. Palmer (1984), "Assessing the relative impacts of economic journals" *Journal of Economic Literature*, 22, 77-88.

Palacios-Huerta, I., and O. Volij, 2004 “The Measure of Intellectual Influence,” *Econometrica*, 72: 963-977.

Posner, R. A. (2000) An economic analysis of the use of citation in the law" *American Law and Economic Review* 2(2), 381-406.

Appendix: Shape of the Objective Function

In this appendix, we choose two relatively high and two relatively low values for each parameter in order to provide a sense as to the shape of the objective function. The "low" parameters employed in the table below are b=0.7, γ=0.03, h^w=0.7, h^l=0.7, while the high parameters are b =2.2, γ=0.27, h^w=2.8, h^l=2.8. Grade refers to the number of correct predictions for the 1999-2003 seasons.

Parameters				**Grade**
b	γ	h^w	h^l	
High	Low	High	High	81
High	Low	Low	High	77
Low	Low	High	High	77
Low	Low	Low	High	77
High	Low	High	Low	73
Low	Low	High	Low	73
Low	Low	Low	Low	72
High	Low	Low	Low	71
Low	High	High	Low	70
Low	High	Low	Low	69
High	High	Low	Low	59
High	High	High	Low	58
High	High	Low	High	52
High	High	High	High	49
Low	High	High	High	49
Low	High	Low	High	49

Table A1: Shape of the Objective Function