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

Itay Fainmesser, Chaim Fershtman and Neil Gandal

Journal of Sports Economics 2009; 10; 582 originally published online Jun 10, 2009;
DOI: 10.1177/1527002509336891

Journal of Sports Economics
Volume 10 Number 6
December 2009 582-600

10.1177/1527002509336891
http://jse.sagepub.com
hosted at
http://online.sagepub.com

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

Itay Fainmesser
Harvard University
Chaim Fershtman
Neil Gandal
Tel Aviv University

The National Collegiate Athletic Association (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 article 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 postseason 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 and the outcome of the 2004-2006 regular seasons to rank the teams each year for 2004-2006. We then calculate the number of bowl games whose outcomes were correctly predicted following the 2004-2006 season. None of the six ranking schemes used

Authors' Note: We are grateful to the Editor, Leo Kahane and two anonymous referees whose comments and suggestions significantly improved the paper. We thank Irit Galili and Tali Ziv for very helpful research assistance. We are grateful to Drew Fudenberg and participants at the Conference on "Tournaments, Contests and Relative Performance Evaluation" at North Carolina State University for helpful suggestions.
Please address correspondence to Neil Gandal, Department of Public Policy, Tel Aviv University, Tel Aviv, 69978, Israel; gandal@post.tau.ac.il.

by the Bowl Championship Series (BCS) predicted more bowl games correctly over the 2004-2006 period than our CWR scheme.

Keywords: *consistent ranking; empirical estimation; sports*

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: Louisiana State University (LSU) and University of Southern California (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 Entertainment Sports Programming Network (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.[1] 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 determined either 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 12−13 games and yet, there are 120 teams in (the premier) Division I-A NCAA college football.[2]

The teams form a network, where teams are nodes and there is a link between the teams if they play each other. 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 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 use 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 preseason 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. A consistency requirement was first used by Liebowitz and Palmer (1984) when they constructed their academic journal ranking. See also Palacios-Huerta and Volij (2004) for an axiomatic approach for determining intellectual influence and in particular academic journal ranking.[4] Their invariant ranking (which is also consistent) is at the core of the methodology that the Google search engine uses to rank Web pages.[5] "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'."[6,7]

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 are 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, because there is often a "home field" advantage. An analogy for wins and losses also exists for the case of academic articles. 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., not submitted).[8]

This article 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 no 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 postseason 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 evaluated by using the outcomes of the bowl games for those 5 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 much lower than 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 very important in determining the ranking. Further, our estimates are such that a team is penalized more for a home loss than a road loss.

The wealth of information and rankings available on the Internet suggests that the rating of college football teams attracts a great deal of attention.[9] There are, however, just six computer ranking schemes that are used by the BCS. Comparing the CWR ranking to these six rankings indicates that over a 5-year period, the CWR ranking did approximately $10-14\%$ better (in predicting correct outcomes) than the four BCS rating schemes for which we have data for the 1999-2003 period. 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.

Hence, we use the 2004-2006 seasons, which were not used in estimating the parameters of the ranking, and perform a simple test. Using the estimated parameters, we use the CWR and the outcome of the 2004-2006 regular seasons to determine the ranking of the teams for each of the seasons from 2004-2006. We then evaluate our ranking scheme by using it to predict the outcome of the 2004-2006 postseason (bowl) games. While one of the BCS schemes did as well as we did over this period, our CWR ranking scheme predicted more bowl game outcomes correctly than the other five computer rankings used in the BCS rankings for 2004-2006 period. While these results do not necessarily suggest any significant difference between our ranking schemes and those of the computer ranking schemes used by the BCS, it is important to point out that our rankings endogenously determine the "strength of schedule" for each team each season, are consistent, and obtained using a formal objective function. Obtaining results in the same ballpark as the best of these six BCS computer rankings suggest that our methodology (with consistency and a formal objective function) has merit.

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 that uses 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.[10] 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% of the overall BSC ratings. The 2004 BCS rankings were based on the following three components, each with equal weights:[11] (a) The ESPN/USA Today poll of coaches, (b) The Associated Press poll of writers, and (c) Six computer rankings. Hence, the weight placed on the computer rankings was reduced.[12]

Following the 2004 season, the BCS system again came under scrutiny. The complaint involved California (Cal) that appeared to be on the verge of its first Rose Bowl appearance since 1959. Despite Cal's victory in its final game, it fell from fourth to fifth in the final BCS standings and lost its place to Texas, which climbed to fourth, 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.[13]

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[14] on the road 26-16, 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 fifth, Cal ended up playing in a minor (non-BCS) bowl.[15] 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.

Table 1
Changes in Ratings Between November 27 and December 4

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%)

Note: BCS Computer Ranking: No change in California's and Texas' rankings. Games: California 26 Southern Mississippi 16; Texas (idle).

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.[16] The teams form a network, where teams are nodes and there is a link between the teams if they play each other. 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 data set is hence uniquely defined by the matrix $A = |a_{ij}|$. We interpret each $a_{ij} = 1$ as a positive signal regarding object i. The objective is to define a rating function: $R : A \rightarrow R^n$ which generates a rating (and not just a ranking) for every agent that summarizes the information in A.

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 best sellers, 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:

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

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, that is 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 use a simple normalization requiring, for example, that $\Sigma z_i = 1$ and $\min\limits_{i=1,..,n} z_i = 0$. Specifically,

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

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

Equation 1 is related to Google's ranking of Web pages—see Brin and Page (1998) and the Wikipedia entry on PageRank (available at http://en.wikipedia.org/wiki/PageRank.) From Wikipedia, the "page rank" value of Web page i is $z_i = (1-\mathrm{d})/N + \mathrm{d}\sum_{j=1}^{N} l_{ij} z_j$, where N is the number of Web pages, d is an exogenous constant, and l_{ij}= (1/# of outgoing links from Web page j) if Web page j links to Web page i, and 0 otherwise.[17] The "Google" normalization is that the sum of the PageRanks equals 1, that is, $\sum\limits_{i=1}^{N} l_{ij} = 1$.

3.2 Incorporating Wins and Losses

Our discussion up to this point considered the case when $a_{ij} \in \{0, 1\}$. However, 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 $\overline{a}_{ij} \in Z^+$ indicates the number of times team i lost to team j, so the matrix $\overline{A} = |\overline{a}_{ij}|$ is added to the data set and identifies losses while the matrix A is defined as above and identifies the wins.[18] Returning to the analogy of ranking articles, if it would have been feasible to use both acceptance and rejection data, the $\overline{A}$ matrix would be the "rejection" matrix.

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

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

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 ensure 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. Because 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 used for the importance of "home games" vs. "away games."We split each matrix A $(\overline{A})$, into home wins (losses) and away wins (losses). Thus, for every pair of teams $i, j \in N$, there are four relevant values $a_{ij}^{\text{home}}, a_{ij}^{\text{away}}, \overline{a}_{ij}^{\text{home}}, \overline{a}_{ij}^{\text{away}} \in Z^+$ 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 $A^{\text{home}}, A^{\text{away}}, \overline{A}^{\text{home}}, \overline{A}^{\text{away}}$, and we modify the ranking function as follows:

$$z_i\left(A^{\text{home}}, A^{\text{away}}, \overline{A}^{\text{home}}, \overline{A}^{\text{away}}, z\right)$$
$$= \frac{\left[\sum_{j \neq i} a_{ij}^{\text{away}} z_j + h^w \sum_{j \neq i} a_{ij}^{\text{home}} z_j\right] - b\left[\sum_{j \neq i} \overline{a}_{ij}^{\text{away}}\left(\gamma - z_j\right) + h^l \sum_{j \neq i} \overline{a}_{ij}^{\text{home}}\left(\gamma - z_j\right)\right] + g}{\sum_i \left(\left[\sum_{j \neq i} a_{ij}^{\text{away}} z_j + h^w \sum_{j \neq i} a_{ij}^{\text{home}} z_j\right] - b\left[\sum_{j \neq i} \overline{a}_{ij}^{\text{away}}\left(\gamma - z_j\right) + h^l \sum_{j \neq i} \overline{a}_{ij}^{\text{home}}\left(\gamma - z_j\right)\right] + g \right)} \qquad (3)$$

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

Road wins and road losses are normalized to 1. 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 use 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. Because 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. However, that does not pose a problem, because 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.[19] 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 5-year period. In section 5, we discuss some alternative estimation methodologies and explain why we believe our methodology is more appropriate.

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.

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):

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

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 minimum distance 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, $G(b, \gamma, h^w, h^l)$, for every set of parameters.[20]

4.1 Estimation Algorithm

We now describe the algorithm for obtaining our estimates. (i) For each set of the four parameters (b, γ, h^w, h^l), we first need to find a fixed point in the continuum using equation 3. This makes the ranking consistent for the given set of the parameters. (ii) Once, we have the fixed point, we can then assign a rating to each team and rank the teams from the highest to the lowest team. (iii) Finally, we then need to go through all of the bowl games and assign a grade $G(b, \gamma, h^w, h^l)$ based on the number of bowl games correctly predicted by the ranking derived from this set of parameters.

The estimation process is computationally intensive, because we must go through steps (i), (ii), and (iii) for each set of parameters. Given the four dimensions and the fineness of the grid (see below), this process is computationally intensive. The computational cost is especially high because finding the fixed point itself (step (i)) is very computationally intensive.

We first chose relatively broad intervals for the parameters to find areas that provided the best grade. The values chosen for the initial grid (see Table 2 below) were as follows: b which accounts for the importance of losses relative to wins was allowed to vary between 0.1 and 4.0. This means that the importance of losses relative to wins could vary between 10% and 400%. γ was allowed to vary between from 0.01 to 0.32. A γ of 0.32 is roughly 20 times the rating of the most highly ranked

Table 2
Initial Grid and Intervals

	b	γ	h^w	h^l
Lower bound	0.1	0.01	0.1	0.1
Upper bound	4.0	0.32	3.2	3.2
Broad grid intervals	0.3	0.05	0.3	0.3
Narrow grid intervals	0.1	0.01	0.1	0.1

Table 3
Optimal Parameter Estimates

Parameters	b	γ	h^w	h^l
Estimates set 1	3.6	0.022	2.7	1.9
Estimates set 2	0.75	0.038	1.4	1.3

team; hence the range for γ is also very large. h^w and h^l were chosen to allow a large range as well.

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.[21] The best predictions were given by two sets of parameters in two areas of the grid; these two distinct areas yielded 81 and 80 correct predictions, respectively, over the 5-year period (out of a possible 130). The two sets of parameters shown in Table 3 are at the center of the two regions with the highest scores.

To interpret γ, we need to know that the highest rating each year (in the 2004-2006 period) was approximately 0.015. This means that other things being equal, the "loss penalty" for the first set of parameter estimates from losing to a very highly rated team is $\gamma - .015 = .007$, which is approximately 32% of the "loss penalty" of losing to a team with a very low rating ($\gamma - 0 = .022$). Hence, the relatively low γ suggests that it indeed matters to whom one loses. (A high value of γ implies that the ranking is more sensitive to the number of losses, rather than to whom one loses.) When b is close to 1, wins and losses affect the ratings symmetrically. Hence, in the case of the first set of parameters, $b = 3.6$ suggests that ratings are much more sensitive to losses than wins.

The estimated value of h^w (2.7), the value of a home win, is very high relative to the value of a road win (which is normalized to 1). Because nearly 60% of "wins" occur at home, a high value of h^w somewhat offsets the high value of b, and provides a reward for winning. The estimated value of h^l (1.9) means that a team is "punished" more for a home loss than a road loss, which is normalized to 1.

In the second set of parameters, both b and h^w are quite a bit lower than that in the first set of parameters, while γ is somewhat higher and h^l is somewhat lower. There is still a smaller "loss penalty" when losing to highly ranked teams: for the second set of parameter estimates, the loss penalty from losing to a very highly rated team is $\gamma - .015 = .023$, which is approximately 61% of the "loss penalty" of losing to a team with a very low rating. Hence in both sets of parameters, it indeed matters to whom one loses.

The two different sets of parameters give similar results because of the substitutability among the parameters. For example, as b falls from 3.6 to 0.75, much more weight is given to wins than losses. This effect is offset in part by a lower value of h^w (1.4 in the second set of parameters vs. 2.7 in the first set of parameters), which decreases the importance of wins, most of which occur at home. The effect is also offset by a larger loss penalty for losses to more highly ranked teams: 61% (vs. 32%) of the loss penalty from losing to a team with a very low rating.

In the appendix (Figure A1), we provide a sense as to the shape of the objective function as a function of b and γ. In constructing this graph, for a fixed value of b and γ, we let h^w and h^l each take on two values: 1.5 and 3.0, i.e., one low value and one high value.[22] We then took the greatest number of wins among these four possibilities. The graph makes it clear that relatively low values of γ are critical for maximizing the number of correctly predicted games and that as γ rises, b needs to fall to keep the number of correctly predicted outcomes high.

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 used 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 $\Pr\{a \text{ beats } b | z_a, z_b\} = \frac{z_a}{z_a + z_b}$. 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 because 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 downside: 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 closer to one half for the higher ranked teams than for teams lower in the ranking.[23] When we used the alterative estimation scheme, we obtained the parameter estimates shown in Table 4.

Table 4
Alternative Methodology: Parameter Estimates

Parameters	b	γ	h^w	h^l
Estimates	3.1	0.02	3.0	3.0

It is reassuring that the parameter estimates, with the exception of h^l, are quite similar to our first set of preferred estimates. The higher values of h^w and h^l mean that home games are much more important than "road" games. The parameter estimates are intuitive, because 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—and the important information comes from the home games.

6. Evaluating the Performance of the CWR Ranking Methodology

We now compare our ranking methodology with the rankings of the experts. The six computer rankings included in the BCS rankings are[24]

- 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 5, we report the number of correct predictions of the 1999-2003 bowl games for the CWR as well as the four BCS ranking schemes for which we have data for the 1999-2003 period. Table 5 shows that over a 5-year period, the CWR rankings do $10-14\%$ 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.

Finally, we use the 2004-2006 seasons, which were 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 2004-2006 regular seasons, we

Table 5
Bowl Games Predicted Correctly for the 1999-2003 Seasons[25]

Ranking	CWR 1	CWR 2	AH	CM	KM
1999	17	18	14	12	14
2000	16	16	15	13	12
2001	16	15	14	14	15
2002	16	17	15	16	14
2003	16	14	14	14	19
Total 1999-2003	81	80	72	69	74

Note: AH = Anderson & Hester ratings; CM = Colley Matrix ratings; CWR = consistent weighted ranking; KM = Kenneth Massey ratings.

ranked the teams for all 4 years at the end of the regular season. The only information we used was the estimated parameters and the outcome of the relevant regular season. Information from one season to the other was not used in ranking the teams.

We then calculated the number of bowl games whose outcomes were correctly predicted following the 2004-2006 seasons and we compared our result with the number of correct predictions from the six computer ranking schemes used in the BCS ranking.

Table 6 shows that none of the six BCS ranking schemes predicted more games correctly than "CWR 2" for the 2004-2006 period. While these do not necessarily suggest any significant difference between our ranking schemes and those of the computer ranking schemes used by the BCS, it is important to point out that our rankings endogenously determine the "strength of schedule" for each team each season. That is, we do not include any exogenous information about the strength of the teams or the conferences. Further, because our methodology includes parameters for home and away games, we cannot use the results of conference championships held at neutral sites at the end of the regular season. While there are only a few such games, they usually involve two high-ranked teams playing each other. Hence, they include potentially important information that we cannot use. (The number of conference championship games has increased over time—currently five conferences have championship games.)

Finally, our estimation method should perform best when using "one-step ahead" forecasts. Table 6 indeed shows that our estimators (using data from 1999-2003) achieve their best performance (relative to the other ranking schemes) in 2004; in this case, none of the other ranking schemes outperform "CWR 1" or "CWR 2." This is intuitive, because the prediction for 2004 is indeed the "one-step ahead" forecast. Table 6 shows that over time, the relative performance of our estimator declines. In the case of 2005-2006, Table 6 shows that "CWR 2" (with 36 correct predictions) falls exactly in the middle of the pack: during this 2-year period, all of the other ranking schemes predicted between 35 and 37 games correctly. Finally, Table 6 shows (not surprisingly) that our CWR estimators perform relatively poorly in 2007.

Table 6
Bowl Games Predicted Correctly

Ranking	CWR 1	CWR 2	AH	CM	KM	RB	PW	JS
2004	18	19	16	18	15	17	12	15
2005	14	15	14	16	15	17	16	15
2006	19	21	21	21	22	20	21	20
Total 2004-2006	51	55	51	55	52	54	49	50
2007	17	16	21	21	18	16	18	21
Total 2004-2007	68	71	72	76	70	70	67	71

Note: AH = Anderson & Hester ratings; CM = Colley Matrix ratings; CWR = consistent weighted ranking; JS = Jeff Saragin ratings; KM = Kenneth Massey ratings; PW = Peter Wolfe ratings; RB = Richard Billingsley ratings.

The reason for the decline in relative performance is likely due to key institutional changes over time in college football. For example, beginning in 2006, teams were able to play an additional game (12 rather than 11). This added significantly to the number of nonconference games being played and hence provided important additional information that was not available when using parameters based on the 1999-2003 data. Our algorithm, however, is such that the parameters can be re-evaluated every year using the latest data and one can thus calculate "one-step ahead" forecasts every year.

We hence went ahead and calculated the one-step ahead forecasts for 2005, 2006, and 2007, where, for example, we used data from 1999-2004 to calculate the one-step ahead forecast for 2005.[26] The correct predictions using one-step ahead forecasts for these three years are, respectively, 16, 21, and 17. Hence, the total number of correct predictions (73) for 2004-2007 using one-step ahead forecasts exceeds the number of correct predictions for both CWR1 and CWR2 (and all of the methods used by the BCS except CM.) A comparison with Table 6 shows that one-step ahead forecasts always do as well as (or better than) the maximum of CWR1 and CWR2 in each year for 2004-2007.

7. Concluding Remark

The article presents a consistent weighted rating scheme and showed how the results could be applied in developing useful rankings in sports settings. While the focus of this article 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.

In closing, we want to emphasize that we do not claim that our methodology is better than the six computer rankings used by the NCAA. Although these six rating methods are not necessarily transparent, and not necessarily based on any formal objective function, these computer rankings are not simple. Clearly, they are considered by the NCAA to be the best computer ratings available.

Obtaining results in the same ballpark as the best of these methods suggests that our methodology (with consistency and a formal objective function) has merit. In particular, a transparent rating system would likely reduce the number of controversies and allow for a discussion of substance. Additionally, it would provide a benchmark for future work to improve ratings. Finally, it would allow for integration of the knowledge of football experts into formal methods by opening a channel of communication with scholars.

Notes

1. 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.

2. There were 117 Division I-A teams through the 2004 season, 119 Division I-A teams in 2005-2006, and 120 Division I-A teams in 2007.

3. 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. See also Slutzki and Volij (2006).

5. The *consistency* property in Palacios-Huerta and Volij (2004) differs from our definition of consistency.

6. Quote appears at http://www.google.com/technology/

7. The consistent weighted ranking 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 (1987).Ballester, Calvo-Armengol, and Zenou (2006) have shown that the Bonacich centrality measure has significant impact on equilibrium actions in games involving networks.

8. 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. However, this is, of course, a hypothetical example because such data are not publicly available.

9. 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.

10. There are now five BCS bowl games.

11. See http://www.bcsfootball.org/news.cfm?headline=40 for details.

12. If the new system had been used during the 2003 season, the LSU and USC would have played in the 2003 BCS championship game.

13. 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.

14. Southern Mississippi finished the regular season 6-5 and later won its bowl game.

15. 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.

16. In some sports settings, there is the possibility of a tie. In NCAA college football, a game tied at the end of regulation goes into overtime and the overtime continues until there is a winner.

17. By definition, $l_{ii} =$ *(1/# of outgoing links from Web page i)*. The "damping factor," d, is typically set equal to 0.85.

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

19. 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. Because our methodology includes parameters for home and away games, we cannot use the results of conference championships held at neutral sites at the end of the regular season. There are three−four such games each year.

20. Because the grades are built from zeros and ones, each set of parameters is a point in a small region that gives the same result.

21. The search algorithm was written in Matlab. The data, the algorithm (including the code), and the complete set of results for the whole broad and narrow grids are available upon request.

22. We do this because the objective function is less sensitive to h^w and h^l.

23. The "downside" would be more severe if we would use the ranking (rather than the rating) to form $z_a/(z_a + z_b)$. Such a method places much 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 $z_a/(z_a + z_b) = 2/(2 + 1) = 2/3$. However, in a game between teams ranked #15 and #16, the "expected" probability that team number #15 will win is $16/(16 + 15) = 0.52$.

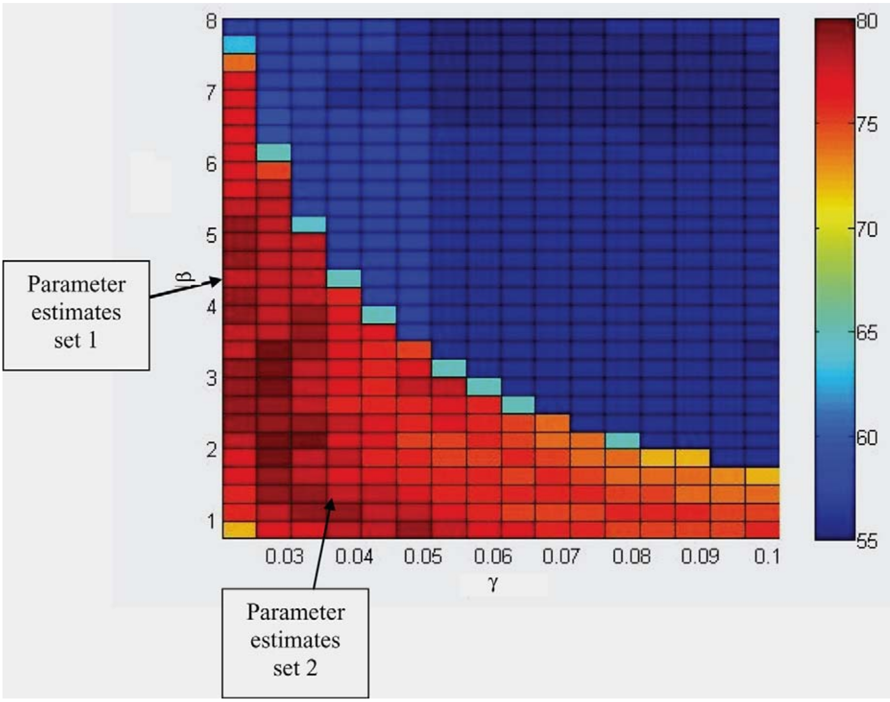
24. There are many other computer rankings in addition to the six used by the BCS. Massey, for example, includes more than 100 rankings on his comparison page. See, for example, the ratings comparison page at the end of the regular season in 2007, available at http://www.masseyratings.com/cf/compare2007-14.htm

25. 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. In the case of the "RB" ranking, we have data for the 2000-2003 period. During that period, the RB ranking predicted 62 games correctly, while CWR1 (CWR2) predicted 65 (66) games correctly. In the case of PW and JS, we only have data for the 2002-2003 period.

26. The number of correction predictions using the one-step ahead forecast for 2004 is, of course, the same (19 correct predictions) as reported for CWR2 in Table 6, because we used data from 1999-2003 in calculating the predictions for Table 6.

Appendix

Figure A1



Shape of the Objective Function

This figure (which is available in color on our website) illustrates how the shape of the objective function depends on b and γ. In constructing this graph, for a fixed value of b and γ, we let h^w and h^l each take on two values: 1.5 and 3.0, i.e., one low value and one high value. We then took the greatest number of wins among these four possibilities. (The numbers refer to the number of correct predictions for the 1999-2003 seasons)

References

Ballester, C., Calvo-Armengol, A., & Zenou, Y. (2006). Who's who in networks. Wanted: the key player. *Econometrica*, *74*, 1403-1417.

Bonacich, P. (1987). Power and centrality: A family of measures. *The American Journal of Sociology*, *92*, 1170-1182.

Brin, S., & Page, L. (1998). *The anatomy of a large-scale hypertextual Web search engine*. Standford University, mimeo.

Fair, R., & Oster, J. (2002). *Comparing the predictive information content of college football ratings*. Mimeo, Retrieved, from 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. Retrieved on May 6, 2009, from http://www.knightfoundation.org/dotAsset/131763.pdf

Hall, B., Jaffe, A., & Trajtenberg, M. (2000). *Market value and patent citations: A first look* (NBER Working Paper W7741).

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

Palacios-Huerta, I., & Volij, O. (2004). The measurement 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*, 381-406.

Slutzki, G., & Volij, O. (2006). Scoring of Web pages and tournaments—Axiomatizations. *Social Choice and Welfare*, *26*, 75-92.

Prof. Neil Gandal is Professor of Economics and Head of the School of Government and Policy at Tel Aviv University. He received his B.A. and B.S. degrees from Miami University (Ohio) in 1979, his M.S. degree from the University of Wisconsin in 1981, and his Ph.D. from the University of California-Berkeley in 1989. Professor Gandal has published numerous papers in industrial organization, the economics of information technology, and the economics of the software & Internet industries. He is an editor at the International Journal of Industrial Organization and a research fellow at the Centre for Economic Policy Research.

Itay Fainmesser is a PhD candidate in Economics at Harvard University. He received his B.Sc. and M.A. degrees from Tel Aviv University, and a second M.A. from Harvard University. Itay is working in the field of Economic Theory. His current work focuses on the interactions between social networks and market activity with implication to market design and include his most recent papers: ''Community Structure and Market Outcomes: Towards a Theory of Repeated Games in Networks'' and ''Social Networks and Unraveling in Labor Markets''.

Prof. Chaim Fershtman is Professor of Economics at the Eitan Berglas School of Economics at Tel Aviv University. He received his B.A., M.A. and Ph.D. degrees from The Hebrew University of Jerusalemhis. Professor Fershtman has published numerous papers in industrial organization and economic theory. He is an editor at the Rand Journal of Economics and a research fellow at the Centre for Economic Policy Research.