Robust Rankings for College Football

Samuel Burer[*]

October 11, 2011
Revised: January 6, 2012

Abstract

We investigate the sensitivity of the Colley Matrix (CM) rankings—one of six computer rankings used by the Bowl Championship Series—to (hypothetical) changes in the outcomes of (actual) games. Specifically, we measure the shift in the rankings of the top 25 teams when the win-loss outcome of, say, a single game between two teams, each with winning percentages below 30%, is hypothetically switched. Using data from 2006–2011, we discover that the CM rankings are quite sensitive to such changes. To alleviate this sensitivity, we propose a robust variant of the rankings based on solving a mixed-integer nonlinear program, which requires about a minute of computation time. We then confirm empirically that our rankings are considerably more robust than the basic CM rankings. As far as we are aware, our study is the first explicit attempt to make football rankings robust. Furthermore, our methodology can be applied in other sports settings and can accommodate different concepts of robustness besides the specific one introduced here.

1 Introduction

College football has been played in the United States since the 1860s and enjoys enormous popularity today. Colleges and universities of all sizes across the country sponsor teams that play each year (or *season*) within numerous conferences and leagues. We focus our attention on teams in the Football Bowl Subdivision (FBS). Roughly speaking, the FBS includes the largest and most competitive collegiate football programs in the country. In 2011, there were 120 teams in the FBS, each of which typically played 12 games per season (not including post-season games).

For historical reasons, the FBS teams do not organize themselves into an elimination tournament at the end of a season to determine the best team (or *national champion*).

[*]Department of Management Sciences, University of Iowa, Iowa City, IA, 52242-1994, USA. Email: samuel-burer@uiowa.edu. The research of this author was supported in part by NSF Grant CCF-0545514.

Instead, the most successful teams from the regular season are paired for a group of extra games, called *bowl games*. In particular, every FBS team plays in at most one bowl game. Ostensibly, the bowl games serve to determine the national champion—especially if the bowl match ups are chosen well. However, there has always been considerable debate over how to choose the bowl match ups.

Prior to the year 1998, the bowl match ups were made in a less formal manner than today. One of the most important factors for determining the match ups were the human-poll rankings, such as the AP Poll provided by the Associated Press. As a result, the poll rankings have long had considerable influence in college football. (Although computer-generated rankings existed at the time, they were not used with any consequence.) In fact, prior to 1998, the national champion was generally considered to be the team ranked highest in the polls after completion of the bowl games. However, even this simple rule was problematic because the final polls could disagree on the top-ranked team. This occurred, for example, after the 1990 season.

Since 1998, the Bowl Championship Series (BCS) has been implemented to alleviate the ambiguity of determining the national champion in college football [1]. The BCS procedure is essentially as follows. At the end of the regular season, multiple human-poll and computer rankings are combined using a simple mathematical formula to determine a *BCS score* for each FBS team. The FBS teams are then sorted according to BCS score, which determines the *BCS rankings*. Then, following a set of pre-determined rules and policies, ten of the top teams according to the BCS rankings are matched into 5 bowl games. In particular, the top-2 teams are matched head-to-head in a single game, so that the winner of that game can reliably be called the national champion.

Even with the BCS now in place, there is still considerable reliance on rankings (human and now also computer). It is clear that quality rankings are necessary for the BCS to function properly, i.e., to reliably setup the game that will determine the national champion and to setup other quality games.

However, it can be quite challenging to determine accurate rankings, especially in college football. Intuitively, good sports rankings are easy to determine when one has data on many head-to-head matchups, which allow the direct comparison of many pairs of teams. This happens, for example, in Major League Baseball, where many pairs of teams play each other often, and thus a team's winning percentage is a good proxy for its ranking. In college football, the large number of teams and relatively short playing season makes such head-to-head information scarce. For example, for 120 FBS teams each playing 12 games against other FBS teams, only 720 games are played out of $7140 = \binom{120}{2}$ possible pairings. In reality, even less information is available because FBS teams often play non-FBS teams and because conference teams mainly play teams in the same conference (making it hard to compare across conferences).

Nevertheless, there are many ranking systems for college football that perform well in practice. One such method, which is one of six computer rankings used by the BCS and which we will investigate in this paper, is the Colley Matrix (CM) method [12]. For a given schedule of games involving n teams, this method sets up a system $Cr = b$ of n equations in n unknowns, where the $n \times n$ matrix C depends only on the schedule of games and the n-vector b depends only on the win-loss outcomes of those games. In particular, b does not depend in any way on the points scored in the games. The solution r is called the *ratings vector*, and it can be shown mathematically to be the unique solution of $Cr = b$. To determine the rankings of the n teams, the entries of r are sorted, with more positive entries of r indicating better ranks. The CM method shares similarities with other ranking systems; see, for example, [17, 18, 14].

In this paper, we investigate whether computer ranking systems can be improved, and we focus in particular on improving the CM method. We do not mean to presume or imply that the CM method needs improvement while the other five BCS computer rankings do not, but the CM method is the only one of the six that is published fully in the open literature and hence can be systematically investigated [6].

We are especially interested in the robustness of the CM rankings, and our research was in part motivated by a situation that arose at the end of the 2010 regular season (i.e., immediately before the bowl match ups were to be determined):

> The final BCS ratings show LSU ranked 10th and Boise State 11th. But . . . Wes Colley's final rankings, as submitted to the BCS, were incorrect. The Appalachian State-Western Illinois FCS playoff game was missing from his data set . . . the net result of that omission in Colley's rankings is that LSU, which he ranked ninth, and his No. 10, Boise State, should be switched. Alabama and Nebraska, which he had 17th and 18th, would also be swapped. . . . LSU and Boise State are so close in the overall BCS rankings (.0063) that this one error switches the order. Boise State should be 10th in the overall BCS rankings and LSU should be No. 11. [5]

In other words, the CM rankings—and hence the BCS rankings—proved quite sensitive to the outcome (or rather, the omission) of a single game. Moreover, this game was played between two FCS (Football Championship Subdivision) teams, and FCS teams are generally considered to be much less competitive than the top-ranked FBS teams and anyway play relatively few games against the FBS teams.

We are also motivated by a recent work of Chartier et al. [9] that investigates the sensitivity of the Colley Matrix rankings (and other types of rankings) under perturbations to a hypothetical "perfect" season in which all teams play one another and the correct rankings are clear (i.e., the top team wins all its games, the second team beats all other teams except the top team, etc). In this specialized setting, the authors conclude that the Colley Matrix rankings are stable but also present a real-world example where the rankings are unstable.

We propose that the top rankings provided by computer systems should be more robust against the outcomes of inconsequential games, that is, games between teams that should clearly not be top ranked. Of course, the top rankings should still be sensitive to important games played between top contenders or even to games played between one top contender and one non-contender.

To this end, we develop a modification of the CM method that protects against modest (hypothetical) changes in the win-loss outcomes of (actual) inconsequential games. We do not handle the case of omitted games (as exemplified in the quote above) since, in principle, accidental omission can be prevented by more careful data handling. Rather, our goal is to devise a ranking system whose top rankings are stable even if a "far away," inconsequential game happens to have a different outcome. This is our choice of what it means for rankings to be robust. While there certainly may be other valid definitions of robustness, we believe our approach addresses a limitation of computer rankings and could also be easily modified for other definitions. Our approach also depends on the definition of "inconsequential," but this can be adjusted easily to the preferences of the user, too. We also remark that, since our approach considers only win-loss outcomes, it naturally incorporates other notions of robustness that strive to produce similar rankings even when a game's point margin of victory is (hypothetically) perturbed.

We stress our point of view that one should protect against *modest* changes to the inconsequential games. As an entire collection, the inconsequential games are probably of great consequence to the top-ranked teams, and so we do not propose, say, simply deleting the inconsequential games from consideration before calculating the rankings. Rather, our approach asks, "Suppose the outcomes of *just a few* of the inconsequential games switched, but we do not necessarily know which ones. Can we devise a ranking that is robust to these hypothesized switches?"

Our approach is derived from the fields of robust optimization [7] and robust systems of equations [13]. Ultimately, this leads to a mixed-integer nonlinear programming (MINLP) model, which serves as the robust version of the system $Cr = b$. Solving this MINLP provides a robust ratings vector r, which is then sorted to obtain the final robust rankings just as in the CM method. We remark that there exist other ranking methods that utilize optimization; see, for example, [10, 11, 15].

Our method depends on a user-supplied integer $\Gamma \geq 0$, which is the number of switched inconsequential games to protect against. In this way, the parameter Γ signifies the conservatism of the user, mimicking the robust approach of [8]. For example, if the user is not worried about inconsequential games affecting the top rankings at all, then he/she can simply set $\Gamma = 0$ (protect against no games changing), and then the ratings vector r is simply the usual CM ratings. On the other hand, choosing $\Gamma = 10$ means the user wants robust rankings that take into account the possibility that up to 10 inconsequential games happen to switch. It should be pointed out that there is no best *a priori* choice of Γ; rather, it will usually depend on the user's experience and conservatism.

It is important to point out that our approach is not stochastic. For example, we do not make any assumptions about the distributions of switched inconsequential games, and we do not study average rankings. Rather, we calculate a single set of rankings that intelligently takes into account the possibility of Γ switched games—but without knowing anything else about the switched games. This is characteristic of robust optimization approaches, which differentiates them from stochastic ones.

This paper is organized as follows. Section 2 reviews the CM method and discusses the data we use in the paper. We also describe our focus on FBS rankings even though our data contains non-FBS data as well. Section 3 then empirically investigates the sensitivity of the top rankings in the CM method to modest changes in the win-loss outcomes of games between teams with losing records. In Section 4, we propose and study the MINLP, which we solve to make the CM method robust to modest changes in the data. In Section 5, we provide several examples and repeat the experiments of Section 3 except with our own robust rankings. We conclude that our rankings are significantly less sensitive than the CM rankings. Finally, we conclude the paper with a few final thoughts in Section 6.

2 The Colley Matrix Method and Our Data

Colley proposed the following method for ranking teams, called the Colley Matrix (CM) method [12]. The CM method uses only win-loss information (as required by the BCS system) and automatically adjusts for the quality of a team's opponent (also called the team's *strength of schedule*). We refer the reader to Colley's paper for a full description; we only summarize it here.

Let $[n] := \{1, \ldots, n\}$ be a set of teams, which have played a collection of games in pairs such that each game has resulted in a winner and a loser (i.e., no ties). Define the matrix

$W \in \Re^{n\times n}$ via

$$W_{ij} := \text{number of times team } i \text{ has beaten team } j.$$

In particular, $W_{ij} = W_{ji} = 0$ if i has not played j, and $W_{ii} = 0$ for all i. Note that ij-th entry of $W + W^T$ encodes the number of times that i and j have played each other, and letting e be the all-ones vector, the i-th entries of $(W + W^T)e$ and $(W - W^T)e$ give the total number of games played by i and its win-loss spread, respectively. With I the identity matrix, also define

$$C := 2I + \text{Diag}((W + W^T)e) - (W + W^T) \tag{1}$$

$$b := e + \tfrac{1}{2}(W - W^T)e, \tag{2}$$

where Diag($\cdot$) places its vector argument into a diagonal matrix. Colley shows that C is diagonally dominant and hence positive definite, which implies in particular that C^{-1} exists. He then defines the *ratings vector* r to be the unique solution of the linear system

$$Cr = b,$$

or equivalently, $r := C^{-1}b$. Then Colley sorts r in descending order, i.e., determines a permutation π of $[n]$ such that the vector $(r_{\pi_1}, \dots, r_{\pi_n})^T$ is sorted in descending order. Then the *rankings vector* is precisely π; that is, the ranking of team i is π_i. If any of r's entries are equal, one can easily adjust the rankings to exhibit ties, but this is unlikely to occur in practice. In the following section, we will provide a specific example of the CM rankings.

We now discuss the data used throughout the paper. We downloaded football data from the website [4] for the 2006–2011 regular seasons. (This website appears to be an archive of Wolfe's website [3].) In particular, no post-season data is included. For each regular season, the data contains the outcomes of all college football games played in the United States, but we limit our focus to just FBS teams. For example, consider the 2010 college football season, which included 3,960 games played between 730 teams around the country. Of the 730 teams, 120 were FBS teams, and of the 3,960 games, 772 involved at least one FBS team. We focus our attention on these 772 games since they contain all data directly related to FBS teams. In the case of 2010, these 772 games yield $n = 195$ because the FBS teams played 75 outside teams. Throughout this paper, ratings will be done for all n teams in a given season, but only ratings and rankings for FBS teams will be discussed since our interest is in ranking these teams. Specifically, we will rate all n teams using the vector r, but prior to computing the FBS rankings, we will delete the non-FBS teams from r before sorting and ranking. In this way, the FBS rankings are computed using all available FBS data, but we focus our rankings on just the FBS teams. (Colley handles non-FBS teams in a more involved pre-processing step, but he likewise maintains a focus on FBS teams [2].)

3 Sensitivity of the Colley Matrix Method

In this section, we empirically investigate the sensitivity of the Colley Matrix (CM) rankings to modest changes in the win-loss outcomes of "inconsequential" games. We specifically focus on the sensitivity of the rankings of the top teams.

Given the win matrix W, let π be the permutation vector encoding the CM rankings for W. Given an integer $t \in [n]$, define

$$T := \{i \in [n] : \pi_i \leq t\}$$

to be the index set of the top t teams (ranks between 1 and t). In contrast, let $\omega \in [0, 1]$ be given and define

$$B := \left\{ i \in [n] : \frac{\sum_{j=1}^{n} W_{ij}}{\sum_{j=1}^{n} (W_{ij} + W_{ji})} < \omega \right\}$$

to be the bottom teams (winning percentages less than ω). As long as t is relatively small and ω is relatively close to 0, it is highly likely that T and B are disjoint. For example, in all experiments, we take $t = 25$ and $\omega = 0.3$ and find that, for the years 2006–2011, T and B never intersect. Note that B does not depend on the rankings π, whereas T does. We call a game *inconsequential* if it has occurred between two bottom teams $i, j \in B$, and we define

$$\mathcal{I} := \{(i, j) \in B \times B : i < j, \ W_{ij} + W_{ji} > 0\}$$

to be the set of all pairs playing inconsequential games. Note that, to remove redundancy, $(i, j) \in \mathcal{I}$ implies $i < j$ by definition.

We wish to examine the sensitivity of the CM rankings of teams in T to modest changes in the win-loss outcomes of games between pairs $(i, j) \in \mathcal{I}$. For this, we define perturbations $W' := W + \Delta$ of the win matrix W that switch the outcomes of a few inconsequential games. Formally, define

$$\mathcal{D} := \left\{ \Delta \in \mathbb{Z}^{n \times n} : \begin{array}{ll} \Delta_{ij} = \Delta_{ji} = 0 & \forall \ (i, j) \notin \mathcal{I} \\ \Delta_{ij} + \Delta_{ji} = 0 & \forall \ (i, j) \in \mathcal{I} \\ -W_{ij} \leq \Delta_{ij} \leq W_{ji} & \forall \ (i, j) \in \mathcal{I} \end{array} \right\}.$$

The condition $\Delta_{ij} = \Delta_{ji} = 0$ for all $(i, j) \notin \mathcal{I}$ guarantees that only inconsequential games are switched, and the equations $\Delta_{ij} + \Delta_{ji} = 0$ for all $(i,j) \in \mathcal{I}$ ensure that any switch is mathematically consistent between (i,j) and (j,i). For example, if we wish to switch a game having $W_{ij} = 0$ and $W_{ji} = 1$, then we need to perturb W_{ij} by $+1$ and W_{ji} by -1. Finally, the inequalities $-W_{ij} \le \Delta_{ij} \le W_{ji}$ limit the number of switched games for $(i,j) \in \mathcal{I}$. For example, in case $W_{ij} = 1$ and $W_{ji} = 2$, it is clear that we logically need $-W_{ij} = -1 \le \Delta_{ij} \le 2 = W_{ji}$.

We also define a convenient restriction of $\mathcal{D}$. Given $\Delta \in \mathcal{D}$, the quantity $\sum_{i<j} |\Delta_{ij}|$ equals the number of games switched by Δ. For any integer limit $L \ge 0$, we define

$$\mathcal{D}(L) := \left\{ \Delta \in \mathcal{D} : \sum_{i<j} |\Delta_{ij}| \le L \right\}$$

to be those perturbations that switch no more than L inconsequential games. For example, $\mathcal{D}(0) = \{0\}$, and $\mathcal{D}(1)$ consists of all perturbations changing exactly 1 or 0 games. Letting $N := \sum_{(i,j)\in\mathcal{I}} (W_{ij} + W_{ji})$, one can see that the number of perturbations in $\mathcal{D}(L)$ equals $\sum_{\ell=0}^{L} \binom{N}{\ell}$.

For any $\Delta \in \mathcal{D}$, define $W' := W + \Delta$ and consider the CM rankings π' based on W'. We investigate the differences between the rankings π and π' of the teams in T via the measure

$$\delta(W, W') := \sum_{i \in T} |\pi_i - \pi'_i| \,.$$

Alternatively, $\delta(W, W')$ is the 1-norm of the sub-vector indexed by T of the difference $\pi - \pi'$. We call $\delta(W, W')$ the *switch measure*. For example, if the top 2 teams switch places but no other ranks change, then $\delta(W, W') = 2$; if the first and third teams switch places but no other ranks change, then the switch measure is 4; and if the top team drops to fourth place but otherwise the orderings remain the same, then $\delta(W, W') = 6$. If all teams in T remain in the top t of π', then $\delta(W, W')$ is an even number, but it can be odd if some team drops out of the top t.

For each football year $y = 2006, \ldots, 2011$ and each $L = 1, 2$, we examine the distribution of switch measures

$$\mathcal{H}(L, y) := \left\{ \delta(W, W') : \begin{array}{c} W' = W + \Delta \\ \Delta \in \mathcal{D}(L) \end{array} \right\},$$

where $t = 25$ and $\omega = 0.3$. This involves enumerating all $\Delta \in \mathcal{D}(L)$ and calculating $\delta(W, W')$ for each. Computationally, calculating $\delta(W, W')$ is quick, and enumeration of each $\Delta \in \mathcal{D}(L)$ is reasonable for $L \le 2$.

It turns out that, with L fixed, the distributions $\mathcal{H}(L, y)$ of the switch measure behave similarly irrespective of the year y, and so to save space, we merge $\mathcal{H}(L, 2006)$,. . . , $\mathcal{H}(L, 2011)$ into a single histogram for each $L = 1, 2$. The resulting two histograms are shown in Figure 1 with basic summary statistics.

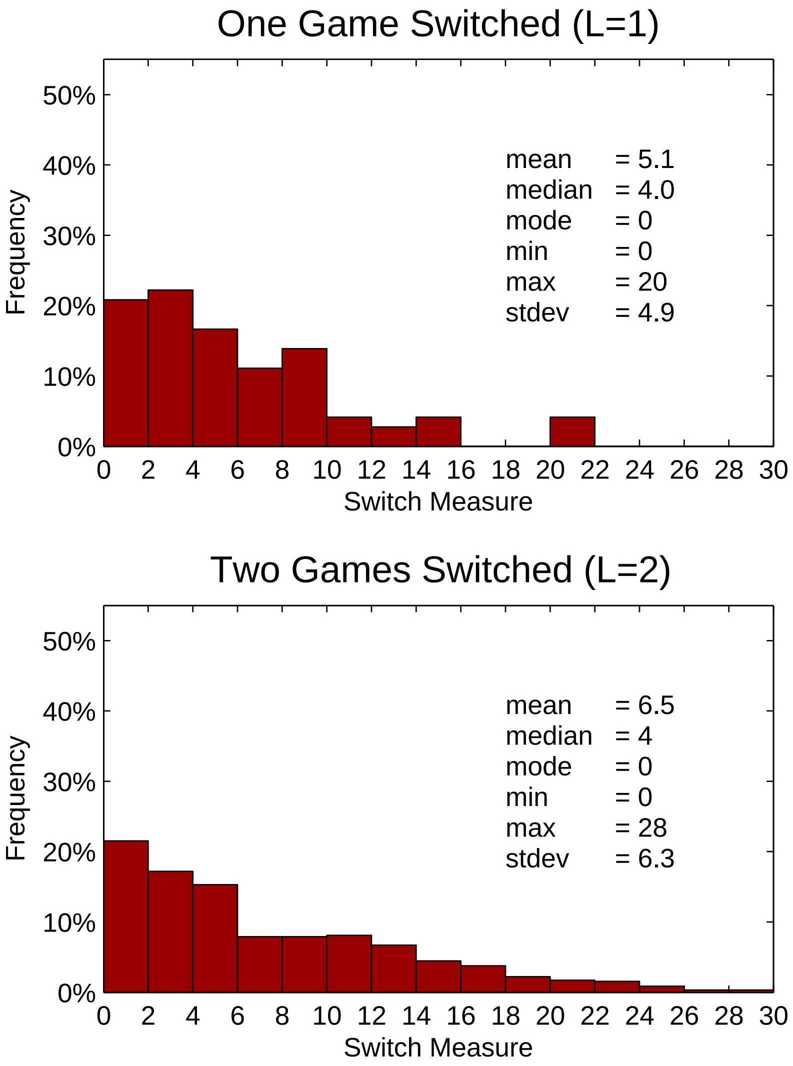


Figure 1: Histograms for the switch measure $\delta(W, W')$ for $L = 1, 2$ switched games, each over the years 2006–2011. This illustrates the sensitivity of the Colley Matrix rankings of the top teams to modest changes in the win-loss outcomes of inconsequential games.

One can see from Figure 1 that the CM rankings of the top $t = 25$ teams are quite sensitive to changes in just a few inconsequential games. For $L = 1$, the mean of $\delta(W, W')$ is 5.1, and the maximum (or worst case) is 20, and both of these statistics increase noticeably for $L = 2$. The standard deviation is also relatively large and increases between $L = 1$ and $L = 2$. In our opinion, such sensitivity is an undesirable property of the CM rankings, especially since rankings are relied upon so heavily in college football.

In Table 1, we provide an illustrative (albeit worst-case) example of how the CM rankings can change when the outcome of a single inconsequential game is switched. The first column of teams contains the top-25 CM rankings for 2007. These teams comprise T in our experiments. The second column shows the perturbed rankings when the result of the inconsequential game between Marshall and Rice is switched. Note that, in 2007, both Marshall and Rice had winning percentages below 30%, and Marshall beat Rice in real life. In both columns, a bold typeface indicates a ranking that changes. For this example, $\delta(W, W') = 20$, and one can see quite plainly that there is a significant amount of shuffling in the rankings. Some of the shuffling is logical. For example, West Virginia beat Marshall in real-life, and when Marshall loses to Rice hypothetically, West Virginia then becomes a weaker team with a lower ranking. What is unclear, however, is why West Virginia drops four spots, which in our opinion seems excessive.

1	**Virginia Tech**	**LSU**
2	**LSU**	**Virginia Tech**
3	Missouri	Missouri
4	**Ohio State**	**Georgia**
5	**Georgia**	**Ohio State**
6	Oklahoma	Oklahoma
7	**West Virginia**	**Florida**
8	**Florida**	**Hawaii**
9	**Hawaii**	**Kansas**
10	**Kansas**	**Arizona St**
11	**Arizona St**	**West Virginia**
12	Boston College	Boston College
13	Southern Cal	Southern Cal
14	South Florida	South Florida
15	Clemson	Clemson
16	Brigham Young	Brigham Young
17	Illinois	Illinois
18	Tennessee	Tennessee
19	**Cincinnati**	**Virginia**
20	**Virginia**	**Cincinnati**
21	**Connecticut**	**Auburn**
22	**Auburn**	**Connecticut**
23	**Wisconsin**	**Texas**
24	**Oregon**	**Wisconsin**
25	**Texas**	**Oregon**

Table 1: Comparison of the 2007 Colley Matrix rankings before (left) and after (right) the result of the inconsequential game between Marshall and Rice is switched.

4 The Robust Method

The basic idea of our robust method is to calculate a ratings vector r that works well even if the win matrix is modestly perturbed from its real-life value W. Robust rankings will then be determined by sorting r in descending order just as with the regular Colley Matrix (CM) method. We call our method the *Colley Matrix Plus* (CM+) method.

Recall the CM system of equations $Cr = b$ for the win matrix W. For a user-specified integer $\Gamma \geq 0$, we consider perturbations $W' := W + \Delta$ for $\Delta \in \mathcal{D}(\Gamma)$ as introduced in the preceding section, and we analyze the perturbed system $C'r = b'$ for W', where C' and b' are given by (1)–(2) except that W' takes the place of W. (Note that Γ plays essentially the same role as L in Section 3, but we will actually use the two parameters Γ and L in slightly different ways for our experiments in Section 5. To facilitate the discussion therein, we introduce and use Γ in this section.) Using properties of Δ, it holds that $\Delta + \Delta^T = 0$, which implies

$$\begin{aligned} C' &= 2I + \mathrm{Diag}((W' + W'^T)e) - (W + W'^T) \\ &= 2I + \mathrm{Diag}(((W + \Delta) + (W + \Delta)^T)e) - ((W + \Delta) + (W + \Delta)^T) \\ &= C + \mathrm{Diag}((\Delta + \Delta^T)e) - (\Delta + \Delta^T) \\ &= C, \end{aligned}$$

i.e., the perturbation Δ does not alter the matrix C. This makes sense because C depends only on the schedule of games, which is not changed by Δ. On the other hand, it holds that

$$b' = b + \tfrac{1}{2}(\Delta - \Delta^T)e, \tag{3}$$

and so b' changes linearly with Δ. In total, we are faced with perturbed systems $Cr = b'$, where Δ ranges over $\mathcal{D}(\Gamma)$.

Because C is invertible, there is clearly no single r that solves $Cr = b'$ for all $\Delta \in \mathcal{D}(\Gamma)$ except in the special case when Γ equals 0. A standard idea from robust optimization and the study of robust systems of equations is to search for an r that minimizes the worst-possible violation of $Cr = b'$ over all $\Delta \in \mathcal{D}(\Gamma)$, i.e., to solve the optimization problem

$$\min_{r} \max_{\Delta \in \mathcal{D}(\Gamma)} \|Cr - b'\|_p \tag{4}$$

where $\|\cdot\|_p$ is a user-specified vector p-norm. It is not immediately clear that (4) can be solved in a tractable manner (either practically or theoretically). We focus on the case $p = 2$[1] and argue next that, even though (4) is a mixed-integer nonlinear program that appears to be NP-hard, we can devise an exact solution procedure that works well in practice (at least for relatively small numbers of inconsequential games and relatively small values of Γ).

So fix $p = 2$. We first transform (4) by minimizing the maximum squared norm and

[1] In the first version of this paper, we focused on the case $p = \infty$ for which (4) can be solved as a polynomial-time LP. However, in this case, there were many alternative optima r, which introduced considerable ambiguity in the resultant rankings. We thank the anonymous referees for suggesting and encouraging a switch to a different p-norm.

separating the objective function via (3):

$$\min_{r} \left(\|Cr - b\|^2 + \max_{\Delta \in \mathcal{D}(\Gamma)} (b - Cr)^T (\Delta - \Delta^T) e + \tfrac{1}{4} \|(\Delta - \Delta^T) e\|^2 \right). \tag{5}$$

By introducing an auxiliary variable t, we can rewrite the inner maximization using a set of explicit linear constraints:

$$\begin{aligned} \min_{r,t} \quad & \|Cr - b\|^2 + t \\ \text{s. t.} \quad & (b - Cr)^T (\bar{\Delta} - \bar{\Delta}^T) e + \tfrac{1}{4} \|(\bar{\Delta} - \bar{\Delta}^T) e\|^2 \le t \quad \forall\, \bar{\Delta} \in \mathcal{D}(\Gamma). \end{aligned} \tag{6}$$

It is important to note that Δ is no longer a variable. Rather, there is one linear constraint in (r, t) for each specific $\bar{\Delta} \in \mathcal{D}(\Gamma)$. As such, (6) is a strictly convex quadratic program with a unique optimal solution that can in principle be solved by CPLEX, for example.

There is still one challenge, however. Since $\mathcal{D}(\Gamma)$ contains $\sum_{\gamma=0}^{\Gamma} \binom{N}{\Gamma}$ elements, where N is the total number of inconsequential games, for most combinations of N and Γ we cannot simply list and solve over all linear constraints; the number of such constraints is simply too large. So instead we adopt the following strategy. First, we solve (6) over a limited subset of constraints to generate an approximate solution $(\bar{r}, \bar{t})$ of (6). Then we solve the following subproblem over the variable Δ:

$$\max_{\Delta \in \mathcal{D}(\Gamma)} (b - C\bar{r})^T (\Delta - \Delta^T) e + \tfrac{1}{4} \|(\Delta - \Delta^T) e\|^2 - \bar{t}$$

Let $\bar{\Delta}$ be an optimal solution. If the optimal value of the subproblem is positive, then we have determined a violated constraint of (6), and this constraint is added to the approximate model and the process is repeated. On the other hand, if the optimal value is nonnegative, then we have proved that the current $(\bar{r}, \bar{t})$ is optimal for (6), and hence $\bar{r}$ is the robust ratings vector.

Solving the subproblem for $\bar{\Delta}$ is actually a difficult problem in theory, but CPLEX is able to solve it quickly as long as N and Γ are not too large. In all instances of this paper, solving (4) via (6) and the procedure just outlined requires less than a minute using CPLEX 12.2 [16] within Matlab R2010b [19] on an Intel Core 2 Quad CPU running at 2.4 GHz with 4 GB RAM under the Linux operating system. However, larger values of N and Γ may lead to solve times that take a few minutes or even a few hours.

5 Behavior of the Robust Method

In this section, we examine the behavior of our Colley Matrix Plus (CM+) method in practice on the football data from 2006–2011.

5.1 Variation as Γ increases

Figure 2 presents the top-25 CM+ rankings for the 2008 football season for eleven choices of Γ: $\Gamma = 0, 1, \ldots, 10$. Note that $\Gamma = 0$ yields the regular CM rankings (though keep in mind that these do not necessarily match the rankings on Colley's website [12] due to our different handling of non-FBS data as mentioned at the end of Section 2). The figure includes both a text table and a graphical chart. Each line in the chart depicts the rank trend of a particular team. For example, Oklahoma is ranked 1 for all Γ, and this corresponds to the top-most, flat line. In contrast, the rank line for Virginia Tech starts at 18 and ends at 16.

When examining Figure 2 on its own, it is difficult to make and support claims such as: "The rankings for $\Gamma = 8$ are *better* than the rankings for $\Gamma = 3$." Of course, we would say that $\Gamma = 8$ is more robust than $\Gamma = 3$ by construction (and we investigate this empirically in the next subsection), but in the absence of further analysis, we believe it can be challenging to compare *any* two rankings objectively. So here we would simply like to point out some observations that we believe are relevant concerning the robust rankings as Γ changes.

First, as Γ increases, the rankings are sensible compared to $\Gamma = 0$. For example, we do not see teams making huge jumps in the rankings. In fact, the ranking of each team moves by at most two positions over all $\Gamma \leq 10$.

Second, the changes in the rankings appear to involve several separate groups of closely ranked teams, and each group switches ranks among itself only. For example, Utah and Texas Tech switch places, while Brigham Young, Missouri, and North Carolina adjust to accommodate a decline in the rank of Brigham Young. Two additional groups are Oklahoma St/Florida St/Virginia Tech and Michigan St/Ball St/Boston College.

Third, the rank trends are not necessarily monotonic, i.e., a team's rank can increase and then decrease (or decrease and then increase) as Γ increases. However, the ranks appear to stabilize for larger Γ. For Figure 2, in particular, all ranks are stable for $7 \leq \Gamma \leq 10$.

Finally, the Γ-rankings confirm the robustness of the top-3 teams since they each retain their rank as Γ increases. This could be interpreted as an affirmation of the CM rankings ($\Gamma = 0$) for these top teams in 2008. In a similar manner, the top-15 CM rankings are confirmed to be mostly robust (with the exception of Utah and Texas Tech).

In Figure 3, we show similar charts for the remaining years 2006–07 and 2009–2011. These depict very similar trends as 2008.

Rank	$\Gamma = 0$	$\Gamma = 1$	$\Gamma = 2$	$\Gamma = 3, 4, 5$	$\Gamma = 6$	$\Gamma = 7, 8, 9, 10$
1	Oklahoma	Oklahoma	Oklahoma	Oklahoma	Oklahoma	Oklahoma
2	Florida	Florida	Florida	Florida	Florida	Florida
3	Texas	Texas	Texas	Texas	Texas	Texas
4	Utah	Texas Tech	Texas Tech	Texas Tech	Texas Tech	Texas Tech
5	Texas Tech	Utah	Utah	Utah	Utah	Utah
6	Alabama	Alabama	Alabama	Alabama	Alabama	Alabama
7	Penn State	Penn State	Penn State	Penn State	Penn State	Penn State
8	Boise St	Boise St	Boise St	Boise St	Boise St	Boise St
9	Southern Cal	Southern Cal	Southern Cal	Southern Cal	Southern Cal	Southern Cal
10	Ohio State	Ohio State	Ohio State	Ohio State	Ohio State	Ohio State
11	Cincinnati	Cincinnati	Cincinnati	Cincinnati	Cincinnati	Cincinnati
12	Georgia Tech	Georgia Tech	Georgia Tech	Georgia Tech	Georgia Tech	Georgia Tech
13	Georgia	Georgia	Georgia	Georgia	Georgia	Georgia
14	TCU	TCU	TCU	TCU	TCU	TCU
15	Pittsburgh	Pittsburgh	Pittsburgh	Pittsburgh	Pittsburgh	Pittsburgh
16	Oklahoma St	Oklahoma St	Oklahoma St	Oklahoma St	Oklahoma St	Virginia Tech
17	Florida St	Florida St	Florida St	Florida St	Florida St	Florida St
18	Virginia Tech	Virginia Tech	Virginia Tech	Virginia Tech	Virginia Tech	Oklahoma St
19	Michigan St	Michigan St	Ball St	Michigan St	Ball St	Ball St
20	Ball St	Ball St	Michigan St	Ball St	Michigan St	Boston College
21	Boston College	Boston College	Boston College	Boston College	Boston College	Michigan St
22	Brigham Young	Brigham Young	Brigham Young	Missouri	Missouri	Missouri
23	Missouri	Missouri	Missouri	Brigham Young	Brigham Young	North Carolina
24	North Carolina	North Carolina	North Carolina	North Carolina	North Carolina	Brigham Young
25	Nebraska	Nebraska	Nebraska	Nebraska	Nebraska	Nebraska

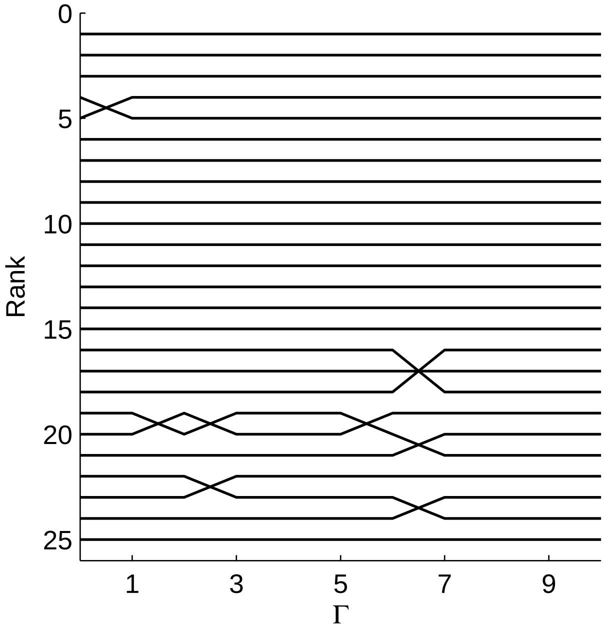
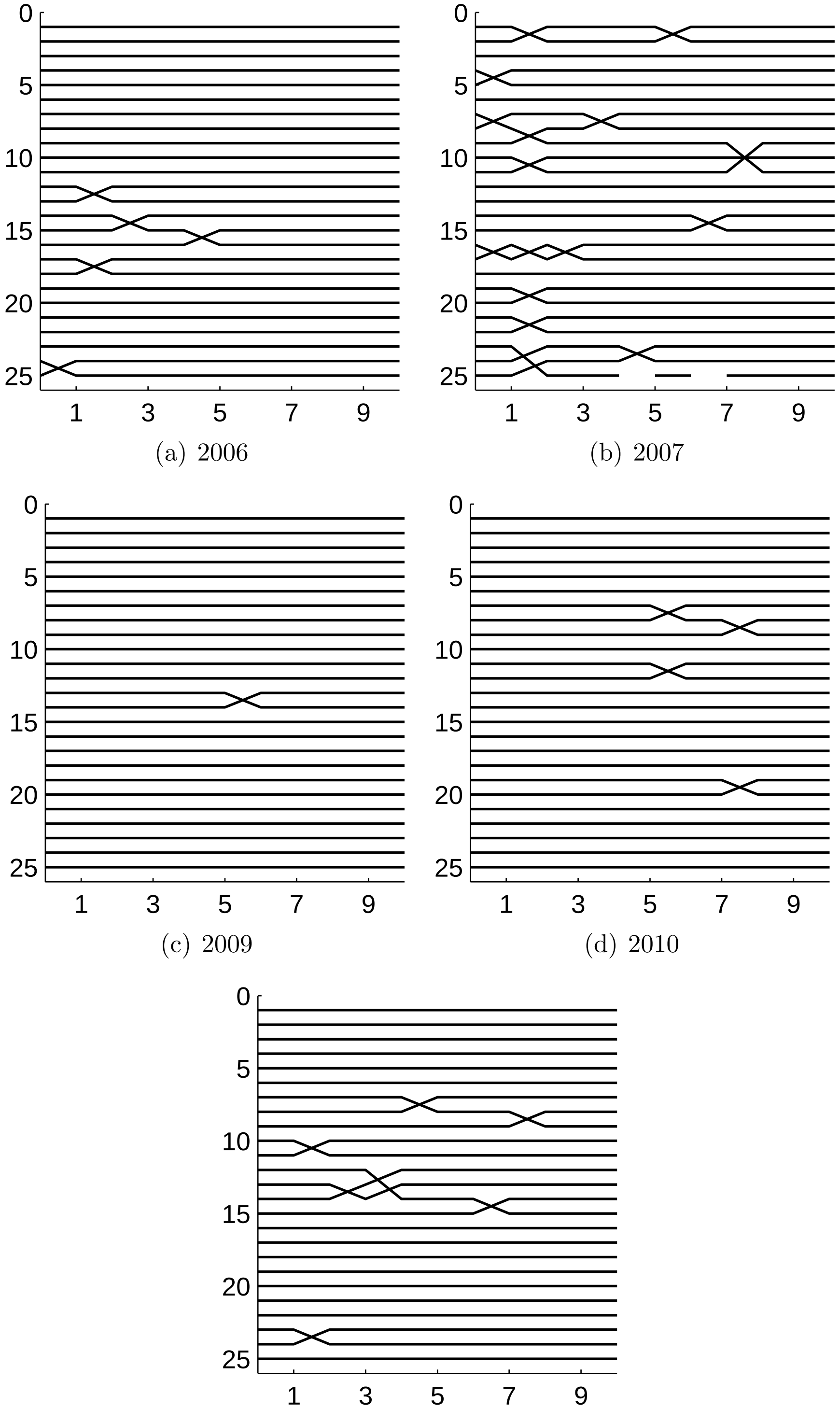


Figure 2: Colley Matrix Plus rankings with $0 \leq \Gamma \leq 10$ for the 2008 football season. Note that $\Gamma = 0$ yields the regular CM rankings (though these do not necessarily match the rankings on Colley's website [12] due to the different handling of non-FBS data as described at the end of Section 2).



(a) 2006

(b) 2007

(c) 2009

(d) 2010

(e) 2011

Figure 3: Colley Matrx Plus rankings versus Γ for all years except 2008.

5.2 Sensitivity of the robust rankings

In Section 3, we investigated the sensitivity of the CM method when the rankings are recalculated after L games are manually switched. We now conduct the same experiment except this time with our robust rankings. Our goal is to verify that our rankings are indeed less sensitive than the CM rankings, at least for certain values of Γ.

In this section, it is important to keep in mind the different roles played by L and Γ. The parameter L determines the number of games that switch before recalculating the robust rankings, whereas Γ is the user-supplied parameter that controls the conservatism of the rankings. In particular, the two parameters are set independently.

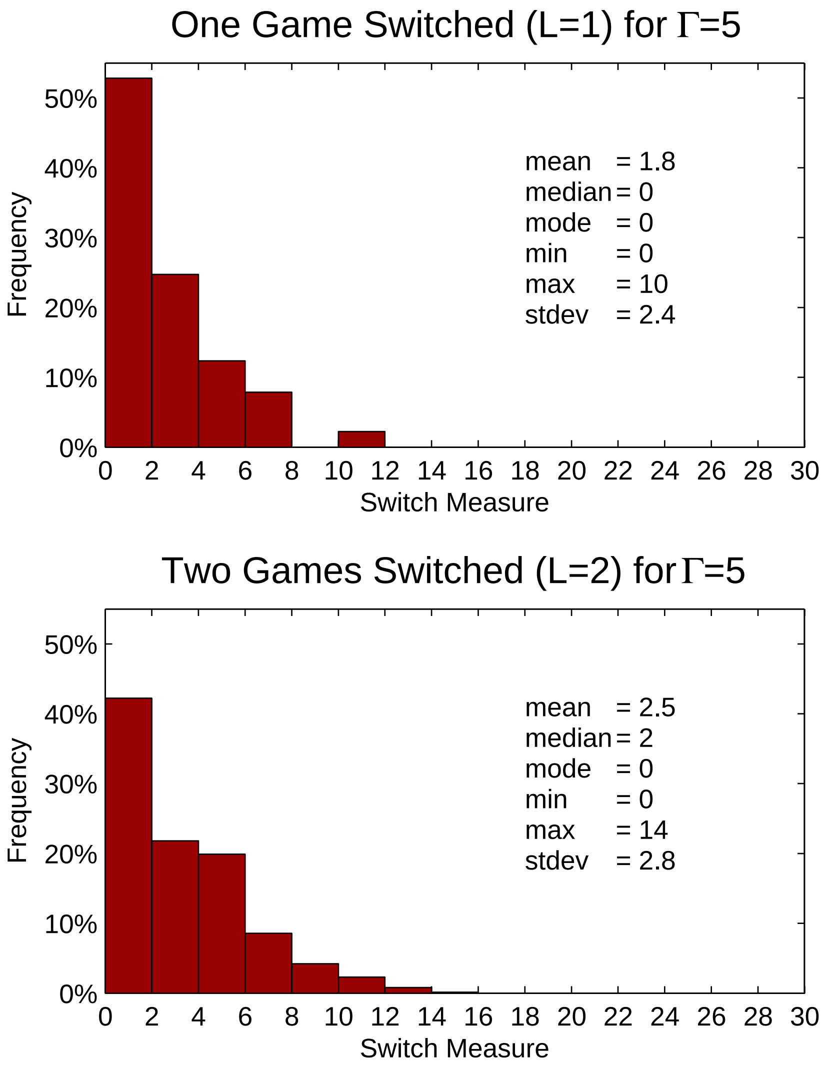
We first describe two important properties of the robust rankings that typify extreme cases. First, when $\Gamma = 0$, the robust rankings are clearly as sensitive as the CM rankings since they are exactly the CM rankings. Second, we claim that, when Γ is sufficiently large, the robust rankings are completely insensitive to L switches. Said differently, for very large Γ, the Γ-robust rankings cannot change upon recalculation after any number of manual switches. To see this, let the win matrix W be given, and suppose $\Gamma \geq N$, where $N := \sum_{(i,j)\in\mathcal{I}}(W_{ij} + W_{ji})$ is the total number of inconsequential games. Then the Γ-robust rankings π based on W take into account the possibility that *all* inconsequential games might switch. Next, let $W' = W + \Delta$ be any perturbed win matrix with $\Delta \in \mathcal{D}(L)$, and calculate the robust Γ-rankings π' based on W'. Because π' is also calculated allowing that all games might switch, it must hold that $\pi = \pi'$. More precisely, the set of scenarios b' optimized over in problem (4) is the same for both W and W' because Γ is so large, and so $\pi = \pi'$.

Between the two extremes $\Gamma = 0$ (as sensitive as CM) and $\Gamma \geq N$ (completely insensitive), it is reasonable to expect the Γ-rankings will become less sensitive as Γ increases, and we now exhibit this at the intermediate, fixed value of $\Gamma = 5$. So let π be the Γ-robust rankings determined by the original W, and let T be the index of the top $t = 25$ teams under π. Also let π' be the Γ-robust rankings determined by $W' := W + \Delta$, where $\Delta \in \mathcal{D}(L)$ for some L. As in Section 3, we investigate the distributions of the top-team switch measures

$$\mathcal{H}(L, y) := \left\{ \delta(W, W') : \begin{array}{c} W' = W + \Delta \\ \Delta \in \mathcal{D}(L) \end{array} \right\},$$

for each football year $y = 2006, \ldots, 2011$ and each $L = 1, 2$. As in Section 3, we then actually combine $\mathcal{H}(L, 2006)$,..., $\mathcal{H}(L, 2011)$ into a single histogram for each L, the results of which are shown in Figure 4.

We can compare Figure 4 directly with Figure 1 of Section 3. Note in particular that all histograms are plotted on the same scale. Looking at both the plots and summary statistics,



One Game Switched (L=1) for Γ=5

Two Games Switched (L=2) for Γ=5

Figure 4: Histograms illustrating the sensitivity of the CM+ rankings of the top teams to modest changes in the win-loss outcomes of inconsequential games.

we see very clearly that the distributions in Figure 4 are considerably lower than those in Figure 1. This demonstrates that, indeed, our CM+ rankings are less sensitive than the CM rankings under the same number of switches ($L = 1$ or $L = 2$), and higher values of Γ will further stabilize the robust rankings.

We conduct one last set of experiments to compare directly the sensitivity of the CM and CM+ rankings. Again, we fix $\Gamma = 5$ and take $L = 1, 2$. For each L, the histogram corresponding to L in Figure 4 is based on all possible switches of L games. For each of these same switches, we also calculate the switch measure for the regular CM rankings just as in Section 3. In Figure 5, we then plot the point (x, y), where x is the CM switch measure for that instance and y is the CM+ switch measure for the same instance. Then, over all switches, to show the frequency for various (x, y) pairs, we use a bubble chart, where the area of a bubble is proportional to the frequency of its (x, y) center. Please also note that the line "$x = y$" is plotted for reference.

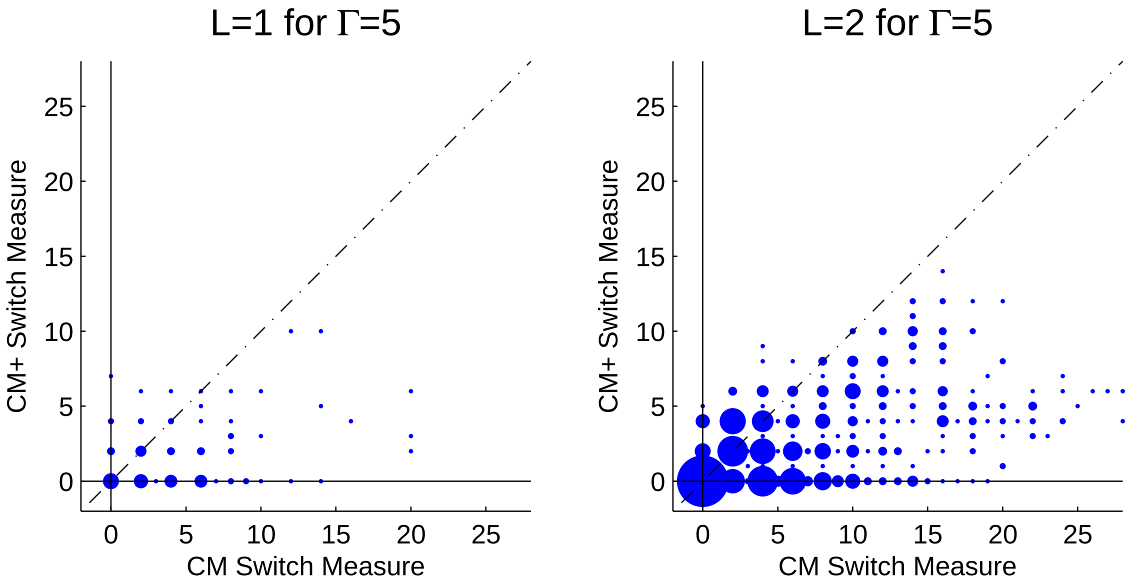


Figure 5: Bubble plots of CM+ versus CM switch measures. CM+ has $\Gamma = 5$ in both plots. Also, $L = 1, 2$, and the instances are all switches of L inconsequential games. The line "$x = y$" is plotted for reference.

Figure 5 shows clearly that the CM+ rankings are much less sensitive than the CM rankings on the same instances. Specifically, the fact that most bubbles are below the $x = y$ line illustrates that, on any given instance, the switch measure for CM+ is less than that of CM. We also note that the CM+ rankings are particularly successful at lessening the sensitivity of the worst-case switch measures for CM (approximately 15–20 for $L = 1$ and 20–25 for $L = 2$).

6 Conclusion

In recent years, the desire to develop robust analytical models has emerged in many fields, including finance, medicine, and transportation, and we believe that computer sports rankings can also benefit from increased robustness. This paper has introduced a particular concept of robustness for college football rankings via the Colley Matrix method. Through experimentation, we have shown that our concept of robustness is consistent and more robust to modest changes in the data. In addition, the time needed to compute the robust rankings (typically less than a minute) is not an obstacle since rankings would be recalculated about once per week in practice.

Our approach can be extended in a number of ways. First, just as the Colley Matrix method can be applied to many sports beyond football, so can ours. This opens the way to robust basketball rankings, chess rankings, etc. In addition, our notion of robustness can be modified to the user's liking. For example, simple changes could be to alter the parameter $\omega \in [0, 1]$ (the winning percentage defining the bottom teams) or to include FCS games in the inconsequential set $\mathcal{I}$. One could also manually choose a completely different inconsequential set $\mathcal{I}$; the analysis and the methodology of the paper will go through unchanged. For example, one may wish to protect the rankings against games that were very close, e.g., where the winner was determined by less than 3 points. $\mathcal{I}$ could then be constructed to contain just pairs of close games.

Acknowledgments

The author wishes to thank two anonymous referees for extremely valuable suggestions that led to a significant improvement in the paper.

References

[1] Bowl Championship Series Official Website. `http://www.bcsfootball.org/`. Accessed October 4, 2011.

[2] FCS Grouping System. `http://colleyrankings.com/iaagroups.html`. Accessed October 4, 2011.

[3] Peter wolfe's college football website. `http://prwolfe.bol.ucla.edu/cfootball/`. Accessed October 4, 2011.

[4] Wilson performance ratings. `http://homepages.cae.wisc.edu/~dwilson/`. Accessed October 4, 2011.

[5] Glitch in the (Colley) Matrix puts Boise State at #10, LSU #11 in BCS standings. `http://sportsratings.typepad.com/college_football/2010/12/glitch-in-the-colley-matrix-puts-boise-state-at-10-lsu-11-in-bcs-standings.html`, December 2010. Accessed October 4, 2011.

[6] Wes Colley, Alabama-Huntsville researcher, talks about his BCS error. `http://www.al.com/sports/index.ssf/2010/12/wes_colley_alabama-huntsville.html`, December 2010. Accessed October 4, 2011.

[7] A. Ben-Tal, L. El Ghaoui, and A. Nemirovski. *Robust Optimization*. Princeton Series in Applied Mathematics. Princeton University Press, Princeton, NJ, 2009.

[8] D. Bertsimas and M. Sim. The price of robustness. *Operations Research*, 52(1):35–53, 2004.

[9] T. P. Chartier, E. Kreutzer, A. N. Langville, and K. E. Pedings. Sensitivity and stability of ranking vectors. *SIAM J. Sci. Comput.*, 33(3):1077–1102, 2011.

[10] B. J. Coleman. Minimizing game score violations in college football rankings. *Interfaces*, 35(6):483–497, 2005.

[11] B. J. Coleman. Ranking sports teams: A customizable quadratic assignment approach. *Interfaces*, 35(6):497–510, 2005.

[12] W. N. Colley. Colley's bias free college football ranking method: The Colley Matrix explained. Manuscript, 2002. Available at `http://www.colleyrankings.com/`.

[13] L. El Ghaoui and H. Lebret. Robust solutions to least-squares problems with uncertain data. *SIAM J. Matrix Anal. Appl.*, 18(4):1035–1064, 1997.

[14] A. Y. Govan, A. N. Langville, and C. D. Meyer. Offense-defense approach to ranking team sports. *J. Quant. Anal. Sports*, 5(1):Art. 4, 19, 2009.

[15] D. S. Hochbaum. Ranking sports teams and the inverse equal paths problem. In *Proceedings of the Second International Workshop on Internet and Network Economics (WINE-2006), Lecture Notes in Computer Sciences*, volume 4286, pages 307–318, 2006.

[16] ILOG, Inc. *ILOG CPLEX 12.2, User Manual*, 2011.

[17] J. P. Keener. The Perron-Frobenius theorem and the ranking of football teams. *SIAM Rev.*, 35(1):80–93, 1993.

[18] K. Massey. Statistical models applied to the rating of sports teams. Master's thesis, Bluefield College.

[19] MATLAB. *version 7.11.0 (R2010b)*. The MathWorks Inc., Natick, Massachusetts, 2010.