Properties of sports ranking methods

Properties of Sports Ranking Methods

Baback Vaziri *[1], Shaunak Dabadghao[2], Yuehwern Yih[3], and Thomas L. Morin[3]

[1]College of Business, James Madison University, 421 Bluestone Dr., Harrisonburg VA 22807, USA
[2]Department of Industrial Engineering, Eindhoven University of Technology, 5600 MB Eindhoven, The Netherlands
[3]School of Industrial Engineering, Purdue University, 315 N Grant St., West Lafayette IN 47907, USA

November 10, 2016

Abstract

Ideally, the ranking of sports teams should incorporate information (comprehensiveness) obtained from the outcome of a match, such as the strength of the opponent and schedule. In addition, the ranking method should be fair and not reward teams for poor performance or factors beyond their control, such as the sequence of the matches. We state properties such that if followed, the ranking methods will be fair and comprehensive. We evaluate five popular sports ranking methods and whether or not they adhere to these properties. Further, we identify a ranking method that under reasonably sufficient conditions will satisfy all of the properties.

1 Introduction

It is often necessary to determine the importance of an alternative compared to others in a group. The process of ordering a list of alternatives based on their relative strength is referred to as ranking. In many cases, a ranking method develops this list by assigning a rating for each alternative, and then ordering the alternatives in decreasing order of rating. Others can be based on minimum violations [1], or linear ordering [6, 7]. Ranking methods are used for a wide array of applications, including but not limited to sports teams [2, 5, 12, 18, 19, 20, 25, 28, 30], web search engines [23, 26, 29], and recommender systems [8, 24].

Pairwise comparison methods are a subset of ranking methods. These methods [15, 21] are still being used in widespread applications today. In this paper, we focus on pairwise comparison ranking methods with applications to sports. We refer to the alternatives being ranked as teams, and the pairwise comparison data as matches or games.

*vaziribx@jmu.edu

A fair and accurate ranking is essential to properly determine the best team(s) in a league, as many sports leagues determine participants for tournaments or playoffs based on the ranking of its teams. There is considerable literature that examines different ranking methods and measures their predictive power and performance [3, 4, 9, 27, 36]. However, it is difficult to rank the ranking methods, because each method has different strengths and weaknesses. For example, many professional leagues (i.e., NFL, NBA, and MLB) consider only the total number of wins and losses when ranking its teams, which fails to take into account several factors such as the quality of a match victory, the strength of schedule, etc. However, some ranking methods that take the quality of a match opponent into account, fail to properly reward a team for winning a match. In turn, different methods consider a different subset of the available information obtained from a match result.

The objective of this study is to identify a set of properties that, when followed, result in a **fair** and **comprehensive** ranking method. We consider that a ranking method is fair when it only considers the factors that are directly in control of the teams. For example, a team has direct control over the result of a match, but no control over the sequence of matches. It would be unfair to penalize or reward a team based on the predetermined sequence of matches. Currently the conventional way to rank teams is by only using the match results. Therefore we define comprehensiveness as the ability to maximize the inference drawn from the win loss data. Therefore injuries, suspensions and other external factors do not count towards comprehensiveness of a ranking method and we do not consider them for the scope of this study.

We point the reader to several articles that focus on analyzing ranking methods, often discussed in the social choice literature. Particularly we want to note [11] and [17], they study many ranking methods and compare their performance over various properties and characterizations for preference aggregation. Other relevant articles in this domain include [10, 22, 32] and [35].

We will study five popular pairwise comparison ranking methods with applications primarily to sports, all of which were recently highlighted in Whos #1? The Science of Rating and Ranking [27]: the traditional Win-Loss method, the Massey method [28], the Colley method [12], the Markov method [18, 27], and the Elo method [16]. These methods will be evaluated on their ability to satisfy the properties developed in this paper. Later we introduce a recently proposed modification to the Markov method [36] and show that under certain parametric conditions, this method will satisfy the properties.

These ranking methods are primarily useful to rank in tournament setups similar to that of a round-robin tournament. For example, single-elimination style tournaments do not need rankings, because the winner will be decided by the structure of the tournament. The design of a tournament is, however, important to examine when electing which ranking method to use, because different designs have varying characteristics [33].

The scope of this study is limited to tournament or league setups in which the teams play an equal number of matches, but they do not necessarily need to play each team in the league. In some cases, such as the English Premier League (EPL), it is a pure round-robin in which each team in the league plays an equal number of matches against each other team in the league. However, another example is the National Football League (NFL), in which each team plays 16 total matches, but will not play every other team in the league. The National Basketball Association (NBA) is a hybrid of the previous two examples. Each team plays 82 matches, and each team will play every other team in the league, but they will not be an equal number of matches with each opponent. However,

all three of the above mentioned leagues setups are acceptable for our study. We do not include Swiss-style tournaments in our scope due to its unique nature of having a large number of players. We direct the reader to [13] for an overview of ranking for such tournaments. We also note that there are many different tools in sports analytics that can be used to improve the predictive power of a ranking method, many of which are highlighted in recent literature [37].

The remainder of this paper is organized as follows: Section 2 outlines the five ranking methods that we study, with a brief description of their strengths and weaknesses. Section 3 introduces the properties and the motivation behind them. Section 4 maps the five ranking methods from Section 2 to the ranking properties in Section 3, and determines which methods satisfy which properties. In Section 5, we conjecture that a recently proposed modification to the Markov method can indeed satisfy all three properties. In Section 6 we discuss our results and future research considerations.

2 Ranking Methods

We outline below five popular sports ranking methods and discuss their relative strengths and weaknesses: 1) the Win-Loss method, 2) the Massey method, 3) the Colley method, 4) the Markov method, and 5) the Elo method. The Win-Loss, Colley, Markov and Elo methods do not consider margin of victory, but the Massey method does. In Section 4, we revisit these methods and evaluate their ability to satisfy the set of ranking properties developed in Section 3.

There are many other ranking methods currently in use. The Glicko rating method is used for ranking Chess and Go players, and it is an improvement over the Elo rating system. Microsoft uses its TrueSkill ranking in order to determine match-ups for on-line gaming. Both these methods determine the rating of a player as well as a confidence in that rating, i.e. (μ, σ). Although both methods are powerful, there are several resulting issues because it is hard to rank teams using two parameters. We could use conservative estimates such as $\mu - k\sigma$, but that will create a different rank order for different values of k. Since we rank teams only based on performance of the current season, a starting value of σ has to be assigned - and different values result in different ranks. While both these methods are effective, they are based on the Elo method which is powerful enough and deemed sufficient for this study. We also looked at other methods like Google's PageRank and the Bradley-Terry pairwise comparison. PageRank is a Markov-based ranking method closely related to the Markov method that is considered here.

2.1 Win-Loss Method

The first method we examine is the traditional Win-Loss method, which is the most commonly used method in professional sports. The method is very intuitive, and requires no modeling to obtain its ratings. Simply sum up the total number of wins and losses for all teams, and assign rating values for each team equal to the total number of wins (draws are counted as half a win and thus count for $\frac{1}{2}$ of a rating point). The advantage of this ranking method is that it provides a clear and direct incentive to win each match. Also, the result of external matches will not affect a specific teams rating value. The disadvantage is that each win is treated the same, regardless of the strength of opponent or the margin of victory. For example, two teams could end up with an equal number of wins, but one team faced much stronger opponents.

2.2 Massey Method

Kenneth Massey developed the Massey method in 1997 to rank college football teams [28]. The concept in this ranking method is that the difference in the ratings of two teams should equal the difference in the score of their competition. The fundamental equation for the ranking method is written as follows:

$$\boldsymbol{M}r = p \tag{1}$$

In Eq. (1), $\boldsymbol{M}$ is the Massey matrix, r is the unknown rating vector, and p is a vector of cumulative point differentials. The Massey matrix is comprised of the diagonal element M_{ii} which is equal to the total number of games played by team i, and the element M_{ij} which is the negation of the number of games played between team i and j. Because the linear system does not have a unique solution, one of the rows of the Massey matrix must be replaced with all ones and the corresponding entry of the right-hand side vector with a zero. The solution to this revised system of linear equations above will give the rating vector.

The point differential vector does not take into account the scoring margins against specific teams, only the cumulative sum for each individual team. In turn, a large cumulative point differential can be obtained from defeating weaker opponents by large amounts, which is not necessarily a strong indicator of team quality. The Massey method was used by the NCAA Football Bowl Subdivision (FBS) in calculating the Bowl Championship Series (BCS) rankings. The BCS rankings were used from 1998−2013 to determine the two teams that would play for the National Championship, as well as several other major bowl games.

2.3 Colley Method

The Colley method was developed in 2002 by Wesley Colley [12]. This method also solves a system of linear equations, but has different definitions for its matrix and its right-hand side vector. Let w_i equal the number of wins for team i, l_i equal the number of losses for team i, t_i equal the total number of games played by team i, and n_{ij} equal the number of times teams i and j play each other. The equation for the ranking method is written as follows, with $\boldsymbol{C}$ as the Colley matrix, r as the unknown rating vector, and b as the vector of cumulative wins and losses:

$$\boldsymbol{C}r = b \tag{2}$$

$$C_{ij} = \begin{cases} 2 + t_i & i = j \\ -n_{ij} & i \neq j \end{cases} \tag{3}$$

$$b_i = 1 + \frac{1}{2}(w_i - l_i) \tag{4}$$

Solving the system of linear equations for the unknown rating vector r will provide a ranking of the teams. A shortcoming of the Colley method is that the strength of an individual opponent is not taken into consideration, only the total number of wins and losses. In fact, the strengths and weaknesses of the Colley method are similar to those of the Massey method, the only difference being that one accounts for total point differential and the other the total win differential.

2.4 Markov Method

The Markov method [18, 27], is a pairwise comparison ranking method that uses Markov chains to rank its teams. The main concept of the method is that each individual competition between two teams results in the losing team voting for the winning team. These collection of votes populate a matrix that represents the head-to-head competitions between all the teams. Transforming the voting matrix into a stochastic matrix will ultimately provide the steady-state probability vector, which is equivalent to the rating vector.

There are many ways to construct the final rating vector, which can be calculated from a linear combination of several stochastic matrices. For example, one voting matrix could contain information on just wins and losses, and another voting matrix could contain information on score differentials. In this study, we will use the basic form of voting only for wins and losses. We refer to this as the $(0, 1)$ Markov method. (The losing team receives a "0" vote from the winning team and the winning team receives "1" vote from the losing team.)

The major advantage of the Markov method is that it takes the quality of the victory into account, meaning a victory over a stronger opponent will be valued higher than a victory over a weaker opponent, as will be shown later. A major drawback of the Markov method is that it is sensitive to small changes in data, especially in its tail, and can exhibit faulty behavior under these circumstances [9, 36]. In fact, in some extreme cases, teams will have an incentive to lose a match to increase their rating.

2.5 Elo Method

Finally, we observe the Elo rating method [16], that was initially developed in 1960 to rate chess players. Since then, the method has become popular outside of the chess world, and other outlets have used the method to rank sports teams. Nate Silver's FiveThirtyEight blog [34] uses the method to rank teams in both the NFL and NBA. For the purpose of this paper, we use the Elo method as it was originally designed.

After each player (or team) participates in a match, their rating is modified by the following formula:

$$r_{\text{new}} = r_{\text{old}} + K(S - \mu) \tag{5}$$

In Eq.(5), K is a constant determined by the nature of the competition and the sport. The amount of change in your rating after a game depends on this K value. For leagues like the MLB, where teams play a lot of matches, the K value will be small, and for leagues like the NFL where teams play fewer games, K will be large. A higher K value results in higher variability in rating changes, and a lower K value results in sluggish behavior. We use values of $K = 5$ for MLB, $K = 20$ for NFL and $K = 10$ for NHL and NBA. S is an indicator variable that reflects the outcome of the match (it takes the value of 1 for a win and 0 for a loss), and $\mu \in (0, 1)$ is a logistic function of the difference in the ratings of the two opponents, given by:

$$\mu = \frac{1}{1 + 10^{(r_b - r_a)/400}} \tag{6}$$

The Elo method is strong because it gives a clear and direct incentive for a win, and external matches do not directly impact a teams performance. It also takes into account the quality of the opponent in the match. However, in its standard form, the Elo method

calculates a rating after every match, and thus the sequence of matches for a team can have a significant impact on their Elo rating.

3 Ranking Properties

In this section, we construct a set of properties that a fair and comprehensive ranking method should follow. To be fair, a ranking method must provide the teams being ranked with consistent objectives. The objective for each team is simple: win the match. In turn, winning a match should always result in at least as good of a rating as before, and losing a match should never result in an increased rating. To be comprehensive, a ranking method must examine the information that can be obtained from each match, and adequately assess and rank the teams based on that information.

There is debate as to whether or not the score differential of a match is a good indicator of team performance. On one hand, Redmond [31] found that score differential can often be a misleading characteristic in determining the strength of a team, and more emphasis should be placed on gaining the victory. On the other hand, there are successful ranking methods, such as the Massey method [28], that have been used and primarily consider score differential. In the EPL, and many other international soccer leagues, score differential is used as a tiebreaker when two teams have an equal rating. In leagues such as the MLB, NFL, and NBA, score differential is not taken into account, and the tiebreakers are usually determined by head-to-head match results. For our study, score differential is optional information to use when ranking teams. It is advantageous to have the capability to use score differential, but it is not a requirement based on the properties we require.

3.1 Property I: Opponent Strength

Our first property is based on the idea that each match victory is not equivalent, and that some victories contain more information than others. For example, it would be misleading to give a similar award for beating the best team in the league as opposed to beating the worst team in the league. Thus, a comprehensive ranking method must take into account the quality of a victory when calculating the rating of a team.

In many ways, Property I can be thought of as the counterpoint to the Independence of Irrelevant Matches (IIM) property mentioned in [17]. As indicated in [17], a pure round-robin tournament may desire IIM and thus reduce the need for Property I. However, in general applications where not every team plays each other, IIM is not desired, and thus Property I adds information to the rating vector. There are other arguments in favor of taking opponent strength into account as well, as indicated in [13].

Property I. *The strength of an opponent from a specific match result should be a factor in calculating the rating of a team.*

As stated previously in Section 3, the score differential of a match can often be misleading information when calculating team ratings. Thus, the extension for Property I is a "soft" property, or otherwise an optional property.

Property Ia (Optional)**.** *The margin of victory over an opponent from a specific match result should be a factor in calculating the rating of a team.*

If point differential is used in calculating team ratings, it is strongly recommended that there be a smoothing function to delineate the impact, similar to Keener's approach [20].

3.2 Property II: Incentive to Win

The next property aims to unify the objective for each competitive match, which is simply to win the match. If a team has incentive to lose a match to increase its rating, that will dilute the information obtained from that match. The information used by the ranking method relies on the fact that in each individual match, both teams are trying to win.

Most ranking methods will satisfy this property. However, as we will see in Section 4, some methods rely too heavily on the strength of opponents to calculate ratings, and this can result in erratic cases where teams have incentive to lose a match. This property is analogous to the Nonnegative Responsiveness to the Beating relation property (NNRB) as seen in [17].

Property II. *A team should always have a clear incentive to win a match to increase its rating.*

The converse of this property is not strictly true, but only partially true. Obtaining a victory over a significantly inferior opponent may not improve the rating, but it should not harm it. Also, losing to a strong opponent may not decrease your rating, but it should not be preferred to winning.

Property II also indirectly implies that strong interdependence between teams' ratings can have a negative impact on the ranking vector. Chartier et al. [9] analyzed several ranking methods and their sensitivity, and found a specific case in the NFL where a high interdependence in ratings can lead to teams having an incentive to lose a match.

3.3 Property III: Sequence of Matches

Teams do not select the order of their match schedule. In some collegiate sports, like NCAA football and basketball, teams can dictate their out of conference schedule, but they have no control over their conference schedule. In major professional sports (NBA, NFL, MLB, EPL), teams do not select the sequence of their matches.

In turn, it would be unfair to award or penalize teams differently based on the sequence of their matches. So, if we were to reorder the matches of a season, the rating and ranking vector should not change. In most ranking methods, this is the case, because the results are tallied and tabulated in a static formula.

It is important to note that we are assuming that team strength is not variable during a specific season. There are so many factors that can effect performance both positively or negatively − including, but not limited to: injuries, player transfers or trades, coaching changes, weather conditions. As mentioned previously in this article, we will not be considering external factors when analyzing season results.

Property III. *The specific sequence of matches should not influence the rating and ranking of a team.*

We now have a list of three properties that we declare all ranking methods should satisfy to be both fair and comprehensive.

4 Ranking methods and properties

In this section, we analyze the five ranking methods from Section 2, and whether or not they follow the properties developed in Section 3.

4.1 Property I: Opponent Strength

Property I states that the strength of an opponent should have an impact on the team rating following a specific match. If a team rating changes an equal amount regardless of the opponent, then Property I is not satisfied.

The Win-Loss method, the Massey method, and the Colley method violate Property I. For the Win-Loss method, a team can win or lose against the strongest or weakest team in the league, and their rating will change by the same amount. For the Massey method, wins and losses are not considered, only total score differential is considered. In turn, a team can score many points against weak teams and have a higher rating than a team that defeated strong teams by a smaller margin of points.

For the Colley method, only the total number of wins is considered, not the individual match results. For example, consider a perfect season round robin tournament consisting of five teams, in which the stronger team wins each match. The ranking is shown in Table 1.

Now, lets assume that team E had beaten team A, and recalculate the Colley ratings. The ranking is shown in Table 2.

Table 1: Perfect season, Colley method

Team	Rank	Win-Loss Record	Colley Rating
A	1	$4-0$	0.786
B	2	$3-1$	0.643
C	3	$2-2$	0.5
D	4	$1-3$	0.357
E	5	$0-4$	0.214

As you can see, both teams A and B have an equal rating and ranking, but they each had beaten different teams. The same point can be made for teams D and E, which have the same rating and ranking but different quality of wins. If the Colley method considered the quality of a victory into account, both teams A and B and teams D and E would have different ratings and rankings.

Table 2: Perfect season with upset, Colley method

Team	Rank	Win-Loss Record	Colley Rating
A	1 (tie)	$3-1$	0.643
B	1 (tie)	$3-1$	0.643
C	3	$2-2$	0.5
D	4 (tie)	$1-3$	0.357
E	4 (tie)	$1-3$	0.357

The Elo method and the Markov method both adhere to Property I. For the Elo method, it is clear that the quality of the opponent will affect the rating and beating a stronger team will improve your rating more than beating a weaker team.

For the Markov method, as we showed in Section 2, the rating vector directly comes from the transition probability matrix, which directly comes from the voting matrix. The voting matrix consists of all head-to-head results between all of the teams, and obtaining votes from a specific team will impact your rating based on the rating of that specific

team. Mathematically, given the transition probability matrix P, the rating of team j can be written as:

$$\pi_j = \sum_{i=1}^{n} p_{ij}\pi_i \tag{7}$$

From Eq (7), it can be seen that wins over stronger teams will increase your rating more than wins over weaker teams.

4.2 Property II: Incentive to Win

Property II states that teams should always have an incentive to win to improve their rating. If a team rating increases more from losing a match, as opposed to having won that match, then Property II is not satisfied.

The Win-Loss method, Massey method, Colley method, and Elo method all follow Property II, and there is always a clear incentive for teams to win the next match to improve their rating. It is not possible to improve your rating with a loss in any of these four methods, and in most cases, the rating will decrease as a result of a loss.

The Markov method, on the other hand, can have cases where teams have an incentive to lose to improve their rating, thus violating Property II. There is a strong inter-dependency in the team ratings when using the Markov method, and this can cause erratic behavior in the rating vector. Lets look at two examples, one theoretical and one case study, to illustrate this point. (For a complete example encompassing all of the ranking methods, we direct the reader to [27].)

Again, consider a perfect season round robin tournament consisting of five teams, in which the stronger team wins each match. The ranking is shown in Table 3.

Table 3: Perfect season, Markov method

Team	Rank	Win-Loss Record	Markov Rating
A	1	$4-0$	0.438
B	2	$3-1$	0.219
C	3	$2-2$	0.146
D	4	$1-3$	0.109
E	5	$0-4$	0.088

Next, we will add an upset in which team E instead had defeated team A. The ranking can be seen in Table 4.

Table 4: Perfect season with upset, Markov method

Team	Rank	Win-Loss Record	Markov Rating
A	1 (tie)	$3-1$	0.29
E	1 (tie)	$1-3$	0.29
B	3	$3-1$	0.194
C	4	$2-2$	0.129
D	5	$1-3$	0.097

Notice that the worst team E is now rated and ranked equally with the best team A, which shows how sensitive the Markov method can be to upsets. To see what is meant by having an incentive to lose, lets add another upset. Imagine that the last match is

still to be played between team A and team D. If team A beats team D, we are left with the ranking from Table 4. However, lets see what happens if team A intentionally loses the match to team D.

Table 5: Perfect season with two upsets, Markov method

Team	Rank	Win-Loss Record	Markov Rating
A	1	$2-2$	0.293
B	2	$3-1$	0.22
D	3	$2-2$	0.195
C	4 (tie)	$2-2$	0.146
E	4 (tie)	$1-3$	0.146

From Table 5, not only did losing the match improve team A's rating, but it put them alone in first place. Both the rating and ranking for team A improved with losing that match. Although this theoretical example proves our point, lets also take a look at a real-world case study where this can take place. During the 2011 NFL season, the Green Bay Packers (GB) were 15−1 and had the best record in the league. Their only loss was to the Kansas City Chiefs (KC), who merely went 7−9, but had obtained an upset win over GB. When used to rank the 2011 season, the Markov method ranks KC as the first place team in the league. (Clearly, with a 7−9 record, it should not have been ranked as the best team in the league.) GB, on the other hand, was ranked 3rd even though they had the best record in the league. If GB had lost a second match, it would have changed the rating vector completely. We select the match-up between GB and the Chicago Bears (CHI) (two bitter rivals, which makes the potential of an upset more likely) as the test match. If GB had decided to lose this match, we observe that not only does it improve its rating, but it also improves its ranking to the first place team in the league. Table 6 shows an excerpt of both the actual 2011 NFL season Markov ratings, and the modified season with the incentive to lose case.

Table 6: 2011 NFL season with modifications, Markov method

2011 Season				2011 Season, modified			
Rank	Team	Record	Markov Rating	Rank	Team	Record	Markov Rating
1	KC	$7-9$	7.24	1	GB	$14-2$	6.04
2	BAL	$12-4$	6.14	2	BAL	$12-4$	5.93
3	GB	$15-1$	5.61	3	CHI	$9-7$	5.15
4	PIT	$12-4$	4.72	4	KC	$7-9$	5.11
5	SF	$13-3$	4.6	5	SF	$13-3$	4.63

It is clear that the incentive to lose a match to improve a rating exists in both theoretical examples and case studies. By losing an additional match to CHI, GB significantly improved their rating from 5.61 to 6.04 ($\sim 8\%$ increase) and also improved their ranking from 3rd to 1st place. Again, the sensitivity of the Markov method is displayed by the over inflated ratings for CHI and KC because of their upset victories over GB.

4.3 Property III: Sequence of Matches

Property III states that the sequence of matches on a teams schedule should not have an impact on their rating. At the end of the season, if a team rating changes based on the order of matches, then Property III is violated.

The Win-Loss method, Massey method, Colley method, and Markov method all satisfy Property III, and no team rating will change based on the sequence of matches. The Win-Loss method purely sums up the total number of wins, which will not change based on the order of matches. The Massey, Colley, and Markov methods all use matrices and/or vectors as inputs, and these are the sums of wins or points scored over the course of the season. Thus, the sequence of matches will not affect the entries of the matrices or vectors.

The standard Elo method, however, does depend on the sequence of matches, and thus violates Property III. We applied the Elo method to several NFL seasons and notice that changing the order of matches changes the final rating of the teams. We considered 1) the actual order, 2) the reverse order, and 3) a random order. In fact, in examining the NFL 2012 season, we notice that the order of matches would actually change which teams were selected to the playoffs. (The NFL selects the four division champions, and then the next two highest rated teams from each conference for the playoffs.)

Table 7 shows Elo ratings for the National Football Conference (NFC) in the NFL 2012 season with matches in the actual order. The teams in the gray shaded cells are the teams that would be selected for the playoffs.

Table 7: Elo ratings for NFC in NFL 2012 season, actual order of matches

NFC East		NFC North		NFC South		NFC West	
WAS	1541	GB	1551	ATL	1572	SF	1558
NYG	1515	MIN	1540	CAR	1491	SEA	1555
DAL	1500	CHI	1528	NO	1490	STL	1496
PHI	1428	DET	1434	TB	1483	ARI	1446

Now, lets observe what happens if we simply reverse the order of matches when calculating Elo rating values.

Table 8: Elo ratings for NFC in NFL 2012 season, reverse order of matches

NFC East		NFC North		NFC South		NFC West	
WAS	1525	GB	1548	ATL	1584	SF	1558
NYG	1520	CHI	1539	TB	1484	SEA	1545
DAL	1503	MIN	1529	CAR	1478	STL	1492
PHI	1442	DET	1441	NO	1478	ARI	1463

Of the 16 teams in the NFC, only San Francisco (SF) had the same rating based on a different order of matches. In addition, many teams changed their rank in their division as well. Most noticeably, in the NFC North, the Chicago Bears (CHI) and the Minnesota Vikings (MIN) swapped rank. Because they were fighting for sixth and final Wild Card spot in the playoffs, the order of matches actually affected which team would be selected for the playoffs.

Figures 1 and 2 show the Elo rating for both CHI and MIN throughout the 2012 season based on the different order of matches.

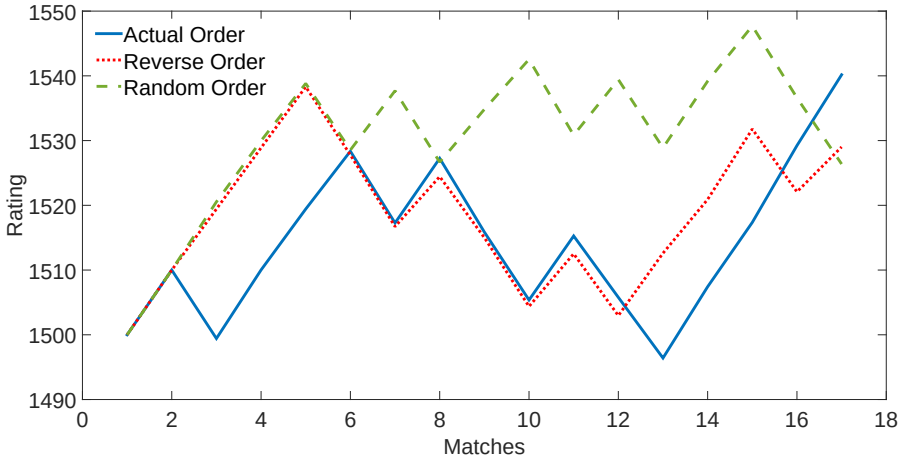

Figure 1: Rating of Minnesota Vikings during 2012 NFL season under different match orders

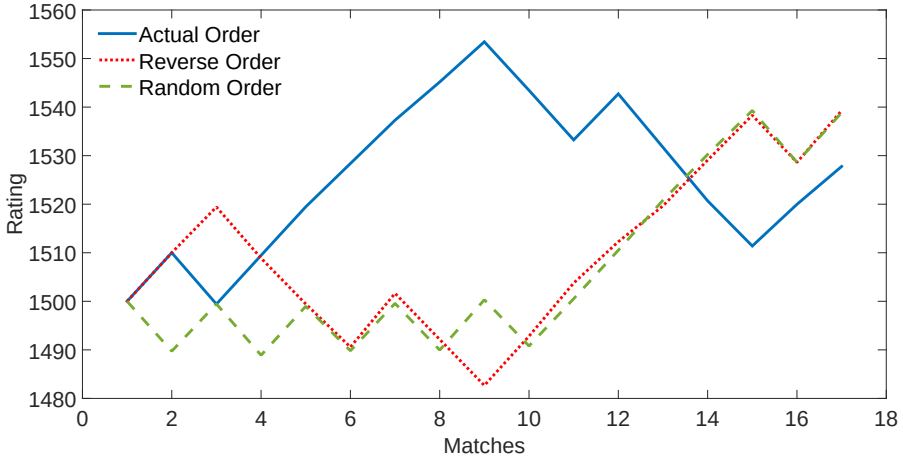

Figure 2: Rating of Chicago Bears during 2012 NFL season under different match orders

Clearly, the order of matches will affect the final team rating when using the Elo method, thus Property III is violated. In summary, all five of the ranking methods violated exactly one of the ranking axioms. Table 9 provides a summary of our findings.

5 A proposed method to satisfy all properties

In this section, we identify a recently developed method, by an author of this paper, in Vaziri et al [36] that is an extension of the Markov method, but with a modified voting scheme, referred to as the $(1, \alpha)$ method. Applied to the NFL seasons $2002 - 2011$ and under specific parametric conditions for α, we observe that this method adheres to all three properties. We discuss a heuristic to choose α in Section 5.1.

Before we examine this method, it is important to discuss the possibility of tweaking the other methods to satisfy the properties. For the Win-Loss, Massey, and Colley methods, it is not possible to modify the method to take Property I into account. The nature of the methods rely on aggregation of wins and losses (or score differential in Masseys case), and having uniform reward for winning a match. This is consistent with

Table 9: Summary of ranking methods and axioms

Method	Property I	Property II	Property III
Win-Loss	×	✓	✓
Massey	×	✓	✓
Colley	×	✓	✓
Markov	✓	×	✓
Elo	✓	✓	×

the findings from Chartier et al. [9] when they showed that the Massey and Colley methods had a uniformly spaced rating vector for a perfect season.

The $(1, \alpha)$ method uses a voting scheme that is a modification to the traditional $(0, 1)$ voting scheme of the Markov method. In the $(1, \alpha)$ method voting scheme, the winning team will vote a value of 1 to the losing team, and the losing team will vote a value of $\alpha > 1$ to the winning team. Another way to view this voting scheme is that when two teams play each other, they are always connected by two arcs. The weight of the arcs is dependent on who wins the match. The winner will have a higher weight, or more "flow" or "votes" coming in from the loser. The remainder of the method is the same algorithm as the traditional Markov method. The parameter α is selected by the user, and represents the confidence that the winning team is indeed the better team. An advantage of the $(1, \alpha)$ method is that it significantly reduces the sensitivity of the Markov method, as shown in [36], while maintaining the integrity of the rank order.

Since the $(1, \alpha)$ method is a modification of the Markov method, it will follow Properties I and III for the same reasons of the traditional method. However, since the $(1, \alpha)$ method also reduces the sensitivity of the Markov method, upsets have a much smaller impact than in the traditional scheme. Thus, in many cases, the $(1, \alpha)$ method will also adhere to Property II and not provide incentive to lose.

The $(1, \alpha)$ method will not satisfy Property II for all values of α, because as α grows very large, the method converges to the $(0, 1)$ Markov method and will have the same properties. However, for smaller values of α, we observe that the incentive to lose no longer exists, and Property II will be satisfied.

We revisit the example from Section 4.2, in which we demonstrated the incentive to lose for team A using the Markov method. This time we use the $(1, \alpha)$ method for $\alpha = 2$, and observe the behavior of the ranking. The ranking for the $(1, \alpha)$ method for a perfect season round robin tournament of five teams is shown in Table 10. Again, we add an

Table 10: Perfect season, $(1, \alpha)$ method, $\alpha = 2$

Team	Rank	Win-Loss Record	$(1, \alpha)$ Rating
A	1	$4 - 0$	0.244
B	2	$3 - 1$	0.218
C	3	$2 - 2$	0.196
D	4	$1 - 3$	0.178
E	5	$0 - 4$	0.163

upset in which team E instead had defeated team A. The ranking can be seen in Table 11.

Notice that the worst team E only improved its ranking by one spot, as opposed to in the traditional scheme in which it became rated and ranked equally with the best team

Table 11: Perfect season with upset, $(1, \alpha)$ method, $\alpha = 2$

Team	Rank	Win-Loss Record	$(1, \alpha)$ Rating
A	1	$3 - 1$	0.227
B	2	$3 - 1$	0.214
C	3	$2 - 2$	0.193
E	4	$1 - 3$	0.191
D	5	$1 - 3$	0.175

A. Also, team E defeated a stronger team than team D defeated, which is shown by the fact that it is rated and ranked ahead of team D. The reduced sensitivity to upsets and the maintained integrity to opponent strength is well demonstrated here.

Finally, we add the third upset to see if the incentive to lose is available for team A, by assuming that they intentionally lose the match to team D. From Table 12, losing the

Table 12: Perfect season with two upsets, $(1, \alpha)$ method, $\alpha = 2$

Team	Rank	Win-Loss Record	$(1, \alpha)$ Rating
B	1	$3 - 1$	0.217
A	2	$2 - 2$	0.211
D	3	$2 - 2$	0.197
C	4	$2 - 2$	0.196
E	5	$1 - 3$	0.179

match decreased both team A's rating and ranking. Also, notice that the ranking is more intuitive than before, in that the rankings closely follow the number of wins and losses for all teams, regardless of the number of upsets.

Next, as we did in Section 4.2, we observe the 2011 NFL season using the $(1, \alpha)$ method, and whether or not there is incentive for GB to lose a match to improve its rating and ranking. First, we show an excerpt of the season ranking based on different values of α, as seen in Table 13.

Table 13: $(1, \alpha)$ method ratings for 2011 NFL season

	$\alpha = 2$		$\alpha = 10$		$\alpha = 20$		$\alpha = 100$	
Rank	Team	Rating	Team	Rating	Team	Rating	Team	Rating
1	GB	4.055	GB	5.594	GB	5.76	KC	6.656
2	NO	3.658	BAL	5.057	BAL	5.489	BAL	5.984
3	BAL	3.639	NO	4.387	KC	5.212	GB	5.698
4	SF	3.602	SF	4.378	SF	4.505	PIT	4.612
5	PIT	3.553	KC	4.342	PIT	4.345	SF	4.589

Note that as α grows large, the rating and ranking vector converges to that of the traditional $(0, 1)$ voting scheme of the Markov method. Next, we add the same upset as we did in Section 4.2 (CHI beats GB in one match), and notice the effect it has on the final season rankings to see if GB had incentive to lose an additional match.

For any value of $\alpha \leq 5$, there was no incentive to lose, and thus, Property II is satisfied. However, once $\alpha \geq 10$, the incentive to lose exists. On analyzing data from

Table 14: $(1, \alpha)$ method ratings for 2011 NFL season with modification

	$\alpha = 2$		$\alpha = 10$		$\alpha = 20$		$\alpha = 100$	
Rank	Team	Rating	Team	Rating	Team	Rating	Team	Rating
1	GB	3.953	GB	5.48	GB	5.771	GB	5.994
2	NO	3.66	BAL	5.011	BAL	5.4	BAL	5.808
3	BAL	3.637	NO	4.468	CHI	4.645	CHI	5.026
4	SF	3.6	SF	4.374	NO	4.52	KC	4.912
5	PIT	3.551	CHI	4.33	SF	4.507	SF	4.61

the NFL seasons from $2002 - 2011$, we found that for α values less than 5, there is never an incentive to lose a match! For values 10 or greater, there were instances where losing a match was beneficial for a team. We visit the criteria for choosing α in the next subsection. The table below shows the number of matches in the season when a team had an incentive to lose for different values of α. One can also think of the values in this table as the number of times Property II was violated. The last row shows the number of matches where a team had an incentive to lose in the $(0, 1)$ Markov method. It is mentioned in Vaziri et al [36], that the rating vector obtained from the $(1, \alpha)$ method should converge to the rating vector obtained from the (0,1) Markov method for large values of α. The last two rows of Table 15 provide evidence of this convergence. Note that there is an exception in 2007, when the New England Patriots had an undefeated season. When using the (0,1) Markov method and having an undefeated team, the dangling node adjustment must be performed (see [26]), which modifies the rating vector. Since the same adjustment is not applied to the $(1, \alpha)$ method, the methods' rating vectors do not converge to equal values.

Table 15: Matches when the victor had incentive to lose - NFL Seasons $2002-2011$

$(1, \alpha)$	NFL Seasons									
Method	'02	'03	'04	'05	'06	'07	'08	'09	'10	'11
2	0	0	0	0	0	0	0	0	0	0
3	0	0	0	0	0	0	0	0	0	0
4	0	0	0	0	0	0	0	0	0	0
5	0	0	0	0	0	0	0	0	0	0
10	0	0	0	2	0	0	0	1	0	2
20	1	0	1	8	0	0	0	4	1	7
100	1	2	5	13	4	3	5	9	2	11
1000000	2	2	5	14	6	3	5	16	2	14
(0,1) method	2	2	5	14	6	0	5	16	2	14

5.1 Choosing α

It is important to test and verify a value for α depending on the league size and the number of matches played by each team. In the MLB or NBA, for example, different values of α could satisfy Property II. Analysis, such as the one done in Table 15, is required to find the relationship and/or threshold of α based on league parameters to guarantee satisfaction of Property II. Further, we can determine α based on predictive

power. For example, in the case for NFL an α value of 2 has the highest predictive power when used to predict the outcome of the playoffs. There are other ways to choose α − we can incorporate home and away advantages, score differentials, etc. The value of α represents the confidence in victory, i.e. a higher value of α means that when team A beats team B, the ranking method will have more confidence that team A is better than team B. But α also affects the sensitivity of the ranking method. A higher value of α corresponds to less control over the sensitivity to upsets. Therefore, it is important to choose an alpha that strikes a balance between confidence and sensitivity. However, the focus in this article is to only consider the mentioned properties − that is to make sure that the value of α used is able to satisfy the requirements for a ranking method.

Table 16: Alpha vs Sensitivity and Confidence

	Sensitivity	Confidence in Victory
Large Alpha	High	High
Small Alpha	Low	Low

Since the $(1, \alpha)$ method is a parametric ranking method, we examine the conditions under which the chosen value of α is able to satisfy our requirements. Mathematically proving or obtaining limits on this parameter will be impossible unless stylized settings are adopted for tournament results (every season or tournament is different and therefore the Markov matrix will be different every time). This can limit us to wait until the end of the season to use this ranking method.

However, we propose a trade-off. In order to find such a limit[1], we examine the shortest path between two nodes (teams) of a graph. Physically, the length of the shortest path represents the least number of votes you have to give to another team to reach them.

For example, if Team A beats Team B, the shortest path from Team A to Team B is 1 using the $(1, \alpha)$ method. The shortest path from Team B to Team A will have a maximum value of α. However, Team B could reach Team A in a shorter distance if they have beat Team C which in turn has beat Team A, thus bringing the distance between B and A from α to 2. These shortest distances depend on the number of teams and the number of matches played by each team in the season. In the NFL, 32 teams play only 16 matches each whereas in the NBA 30 teams play 82 matches each season. It is easier to *reach* other teams in the NBA or NHL than it is in the NFL.

The link between this shortest distance and the violation of Property II is the value of α. We observe that in the seasons $2002 - 2011$ for the NFL, it can take up to 5 or 6 edges to reach another team. But since the teams play so few matches, there are numerous shortest distances of length equal to α. As this value of α grows higher than the 5 or 6, we start observing violations of Property II. If a team deliberately loses a match, *all* of the shortest distances between teams are affected, thus making it possible to have a positive impact on your rating! This behavior is natural to the Markov method, since it calculates the rating vector by looking at all possible ways nodes can reach other nodes. This corroborates our test results shown in Table 15 where we show that for $\alpha < 5$, the $(1, \alpha)$ ranking method will adhere to Property II. We do not observe this in the NBA and NHL (seasons 2002 through 2011) essentially due to the fact that each team plays many more matches than the NFL.

[1]We thank an anonymous referee for suggesting the link between the parameter α and the tournament setting such as teams and number of matches played.

The $(1, \alpha)$ method, for all values of α, needs to be characterized and tested against many of the well known properties as seen in [11, 17] for preference aggregation – and most importantly against the Nonnegative responsiveness to the beating relation (NNRB). We rely on our empirical findings on the methods' relationship with Property II for the purpose of this article, but recognize the need for a formal proof which we will address in future work.

In this section we aimed to and have successfully shown that the $(1, \alpha)$ method satisfies all three properties under certain parametric conditions when applied to the NFL, and thus can be a fair and comprehensive ranking method.

6 Conclusion

In summary, we have outlined a set of ranking properties that all fair and comprehensive pairwise comparison ranking methods should follow. The opponent strength in a match result should impact your rating, there should never be an incentive to lose a match to improve your rating, and the order of matches should not influence the final rating vector.

In future work we propose to study a few other properties that sports ranking methods should satisfy. One such property stated in many articles (see [11, 14, 17]) is the Inversion property. This states that if all the results of the tournament are reversed, the ranking should also be reversed. Whether or not it may be suitable for a round-robin tournament setting is a matter of debate. However, it is easy to see that the Inversion property is not compatible with Property I because it implies symmetric treatment between victories and losses. Both the $(0, 1)$ and the $(1, \alpha)$ methods will fail to satisfy the Inversion property because of their inherent nature that accounts for opponent strength.

Another direction for future work is to show the relationship of our Property II with the Nonnegative Responsiveness to the Beating Relation (NNRB) and the Positive responsiveness to the beating relation (PRB) as discussed in [17] and mathematically prove how the Markov ranking methods discussed in this paper fare against these properties.

We reviewed five popular sports ranking methods and found that none of the five adhered to all three of the properties, although all of them satisfied exactly two of the properties. The Win-Loss, Massey, and Colley methods did not take the opponent strength into account when rewarding a team for a victory. The Markov method is extremely sensitive, and thus has cases where a team has incentive to lose a match to improve its rating and ranking. The Elo method provides different team ratings based on the sequence of matches, which is oftentimes (and always, in major professional sports) not in the teams' control. In future work we propose to study various other ranking methods such as maximum likelihood, fair bets, least square and generalized row sum to see how they can be adapted to ranking sports tournaments and how they fare against the properties stated in this paper.

Last, we conjectured that a newly proposed modification to the Markov method, known as the $(1, \alpha)$ method, will satisfy all three properties under certain parametric conditions. We showed both a generic and case study example where the $(1, \alpha)$ method satisfied all three properties and removed the previous case of having incentive to lose. We note that for large values of α, the method's rating vector converges to the traditional Markov method rating vector, and Property II will be violated. To avoid this, we also provide guidelines on how α should be chosen.

References

[1] I. Ali, W. D. Cook, and M. Kress. On the minimum violations ranking of a tournament. *Management Science*, 32(6):660–672, 1986.

[2] R. D. Baker and I. G. McHale. A dynamic paired comparisons model: Who is the greatest tennis player? *European Journal of Operational Research*, 236(2):677–684, 2014.

[3] D. Barrow, I. Drayer, P. Elliott, G. Gaut, and B. Osting. Ranking rankings: an empirical comparison of the predictive power of sports ranking methods. *Journal of Quantitative Analysis in Sports*, 9(2):187–202, 2013.

[4] S. Burer. Robust rankings for college football. *Journal of Quantitative Analysis in Sports*, 8(2), 2012.

[5] T. Callaghan, P. J. Mucha, and M. A. Porter. Random walker ranking for NCAA division IA football. *American Mathematical Monthly*, 114(9):761–777, 2007.

[6] I. Charon and O. Hudry. A survey on the linear ordering problem for weighted or unweighted tournaments. *4OR*, 5(1):5–60, 2007.

[7] I. Charon and O. Hudry. An updated survey on the linear ordering problem for weighted or unweighted tournaments. *Annals of Operations Research*, 175(1):107–158, 2010.

[8] T. Chartier, A. Langville, and P. Simov. March madness to movies. *Math Horizons*, 17(4):16–19, 2010.

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

[10] P. Y. Chebotarev and E. Shamis. Constructing an objective function for aggregating incomplete preferences. In *Constructing Scalar-Valued Objective Functions*, pages 100–124. Springer, 1997.

[11] P. Y. Chebotarev and E. Shamis. Characterizations of scoring methods for preference aggregation. *Annals of Operations Research*, 80:299–332, 1998.

[12] W. N. Colley. Colley's bias free college football ranking method: The Colley matrix explained. *Princeton University*, 2002.

[13] L. Csató. Ranking by pairwise comparisons for swiss-system tournaments. *Central European Journal of Operations Research*, 21(4):783–803, 2013.

[14] L. Csató. Additive and multiplicative properties of scoring methods for preference aggregation. *Corvinus Economics Working Papers 3/2014, Corvinus University of Budapest, Budapest*, 2014.

[15] H. David. Ranking the players in a round robin tournament. *Revue de l'Institut International de Statistique*, pages 137–147, 1971.

[16] A. E. Elo. *The rating of chessplayers, past and present*. Arco Pub., 1978.

[17] J. González-Díaz, R. Hendrickx, and E. Lohmann. Paired comparisons analysis: an axiomatic approach to ranking methods. *Social Choice and Welfare*, 42(1):139–169, 2014.

[18] A. Y. Govan. *Ranking theory with application to popular sports*. ProQuest, 2008.

[19] A. Y. Govan, A. N. Langville, and C. D. Meyer. Offense-Defense approach to ranking team sports. *Journal of Quantitative Analysis in Sports*, 5(1), 2009.

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

[21] M. G. Kendall. A new measure of rank correlation. *Biometrika*, pages 81–93, 1938.

[22] M. Kitti. Axioms for centrality scoring with principal eigenvectors. *Social Choice and Welfare*, 46(3):639–653, 2016.

[23] J. M. Kleinberg. Authoritative sources in a hyperlinked environment. *Journal of the ACM (JACM)*, 46(5):604–632, 1999.

[24] Y. Koren, R. Bell, and C. Volinsky. Matrix factorization techniques for recommender systems. *Computer*, (8):30–37, 2009.

[25] P. Kvam and J. S. Sokol. A logistic regression/Markov chain model for NCAA basketball. *Naval Research Logistics (NrL)*, 53(8):788–803, 2006.

[26] A. N. Langville and C. D. Meyer. *Google's PageRank and beyond: The science of search engine rankings*. Princeton University Press, 2011.

[27] A. N. Langville and C. D. Meyer. *Who's # 1?: the science of rating and ranking*. Princeton University Press, 2012.

[28] K. Massey. Statistical models applied to the rating of sports teams. *Bluefield College*, 1997.

[29] L. Page, S. Brin, R. Motwani, and T. Winograd. The PageRank citation ranking: bringing order to the web. 1999.

[30] J. Park and M. E. Newman. A network-based ranking system for US college football. *Journal of Statistical Mechanics: Theory and Experiment*, 2005(10):P10014, 2005.

[31] C. Redmond. A natural generalization of the win-loss rating system. *Mathematics magazine*, pages 119–126, 2003.

[32] A. Rubinstein. Ranking the participants in a tournament. *SIAM Journal on Applied Mathematics*, 38(1):108–111, 1980.

[33] P. Scarf, M. M. Yusof, and M. Bilbao. A numerical study of designs for sporting contests. *European Journal of Operational Research*, 198(1):190–198, 2009.

[34] N. Silver. Introducing NFL Elo Ratings — FiveThirtyEight. `http://fivethirtyeight.com/datalab/introducing-nfl-elo-ratings/` (visited: 2015-07-20), 2014.

[35] G. Slutzki and O. Volij. Ranking participants in generalized tournaments. *International Journal of Game Theory*, 33(2):255–270, 2005.

[36] B. Vaziri, Y. Yih, and T. L. Morin. A proposed voting scheme to reduce the sensitivity of the Markov method. *International Journal of Operational Research*, Accepted, In Press, 2015.

[37] W. L. Winston. *Mathletics: How gamblers, managers, and sports enthusiasts use mathematics in baseball, basketball, and football.* Princeton University Press, 2012.