5-2016

Markov-based ranking methods

Baback Vaziri

Graduate School Form
30 Updated 12/26/2015

PURDUE UNIVERSITY
GRADUATE SCHOOL
Thesis/Dissertation Acceptance

This is to certify that the thesis/dissertation prepared

By Baback Vaziri

Entitled
MARKOV-BASED RANKING METHODS

For the degree of Doctor of Philosophy

Is approved by the final examining committee:

Yuehwern Yih
Co-chair

Tom Morin
Co-chair

Mark Lehto

Robert Plante

To the best of my knowledge and as understood by the student in the Thesis/Dissertation Agreement, Publication Delay, and Certification Disclaimer (Graduate School Form 32), this thesis/dissertation adheres to the provisions of Purdue University's "Policy of Integrity in Research" and the use of copyright material.

Approved by Major Professor(s): Yuehwern Yih

Approved by: Abhijit Deshmukh 2/3/2016

Head of the Departmental Graduate Program Date

MARKOV-BASED RANKING METHODS

A Dissertation

Submitted to the Faculty

of

Purdue University

by

Baback Vaziri

In Partial Fulfillment of the

Requirements for the Degree

of

Doctor of Philosophy

May 2016

Purdue University

West Lafayette, Indiana

ACKNOWLEDGMENTS

Many people have helped me along this journey. I'd like to thank my family for providing me with the emotional support to persevere. I'd like to thank my friends for making my experience at Purdue special. I'd also like to thank the faculty and staff in the School of Industrial Engineering for the opportunities to grow as a professional. Last, I'd like to thank my major advisors, Professor Yuehwern Yih and Professor Tom Morin, for believing in me as a researcher, educator, and an individual.

TABLE OF CONTENTS

Page

LIST OF TABLES

Table Page

ABSTRACT

Vaziri, Baback Ph.D., Purdue University, May 2016. Markov-based ranking methods. Major Professors: Yuehwern Yih, Tom Morin.

Ranking methods are an essential tool to help make decisions. This dissertation document examines different aspects of the theory and application of pairwise comparison ranking methods, specifically those that use Markov chains. First, a new method is developed to solve a traditional recruiting problem, and is shown to improve the predictive power of its ranking. Next, modifications are made to an existing method that theoretically improves the reliability, while maintaining the rank integrity. Last, a framework is developed that defines a fair and comprehensive ranking method, and several popular methods are evaluated in their ability to adhere to the said framework.

1. INTRODUCTION

The rank of an alternative in a set is the measure of its dominance in comparison to the other alternatives in that set. A ranking method consists of an algorithm that determines the rank of the alternatives. Generally, the ranking method will first develop a rating for each alternative, and then the sorted ratings are used to obtain a ranking. Ratings contain more information than rankings because they provide a cardinal value that measures the worth of each alternative, whereas a ranking only contains an ordinal value that measures the placing of that alternative in respect to others. Oftentimes, however, we are only concerned with making a decision or prediction, and thus the ranking is sufficient. Rankings are important in many applications, ranging from popular sports to web searches to recommender systems.

A common class of ranking methods involves observing a series of pairwise comparisons between the alternatives to develop a ranking. In a perfect scenario, where a better alternative is always preferred to a weaker alternative, a ranking is easy to develop. For example, if alternative A_1 is preferred to alternatives A_2 and A_3, and A_2 is preferred to A_3, we have a clear ranking relationship of $A_1 > A_2 > A_3$.

In practical scenarios, however, there will most likely be inconsistencies and imperfect results. For example, in the above case we could have A_3 be preferred to A_1, which introduces the need for a ranking method.

Many ranking methods use a series of pairwise comparisons between alternatives to produce a matrix of preference relationships. The rating vector will be either a solution to a system of linear equations or a dominant eigenvector, depending on the ranking method. Many popular methods have experienced considerable success using this framework. For example, the Massey [1] and Colley [2] methods are popular sports ranking methods that use a system of linear equations to develop ratings for teams. The Analytic Hierarchy Process [3] is a ranking method developed by Saaty to help users make decisions involving multiple criteria. The method takes a series of pairwise preferences from a user to develop a dominance matrix, and then solves the dominant eigenvector of that matrix to obtain its ratings and rankings.

Perhaps the most popular example of pairwise comparison ranking methods is Google and their use of Markov chains to rank web pages [4]. Others have used the theory behind this approach to represent the alternatives being ranked as nodes in a discrete-time Markov chain, with the transition probabilities being the pairwise comparisons [5–9]. A generalized form of this ranking method is called the Markov method [7, 10], and was recently highlighted in the text "Who's No. 1? The Science of Rating and Ranking" from Langville and Meyer. The strength of this method is that all of the alternatives are connected in a network of comparisons, so the quality of a victory (or preference) is weighted by the quality of the opponent.

The strength of the Markov method was the motivation behind the first study in this document. We use the traditional Markov method in a new application in a recruiting scenario to provide a ranking.

However, the Markov method has been to shown to be sensitive, especially in its tail, to small changes or outliers in data [11]. For example, in sports applications, an *upset* (where a weaker team beats a stronger team) can challenge the integrity of a ranking that was developed by the Markov method. In this document, we will propose a modification to the Markov method that will reduce the sensitivity of the ranking vector to these upsets.

The last section of this research work is to develop a framework that can evaluate ranking methods and their ability to be fair and comprehensive. It is important to note that we will not consider predictive power in this section, but only the satisficing criteria.

This dissertation is organized in the following manner. Chapter 2 will provide a literature review of the current work and relevant information regarding pairwise comparison ranking methods, specifically ones that use Markov chains. Chapter 3 is a case study that uses a basic form of the Markov method to develop a ranking of college football recruiting classes, and to compare the predictability of this method to that of a leading existing recruiting rankings service. Chapter 4 addresses the theoretical issue with the sensitivity of the Markov method by proposing a modification to the basic voting scheme, and provides experimental and theoretical results that show how to control the sensitivity while maintaining rank integrity. In Chapter 5, we develop a framework of axioms that constitute the fairness and comprehensiveness of a ranking method. Popular ranking methods are evaluated in their ability to adhere to the listed axioms. Finally, Chapter 6 will discuss a summary of conclusions and future research work beyond this dissertation.

2. LITERATURE REVIEW

2.1 Popular Pairwise Comparison Ranking Methods

In general, many ranking methods use a matrix of pairwise comparisons to solve either a system of linear equations or an eigenvector to obtain a rating vector, which in turn, provides a ranking vector. Although the initial intent of many of these methods were to rank sports teams, there are other applications, such as web search engines or recommender systems that use similar principles.

2.1.1 Massey method

Kenneth Massey developed the Massey method in 1997 to rank college football teams using the theory of least squares [1]. The Massey method was used by the NCAA FBS (Football Bowl Subdivision) in calculating the BCS (Bowl Championship Series) 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.

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

$$Mr = p \tag{2.1}$$

where 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 of the teams being ranked.

It is important to note that 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 point differential could be obtained from defeating weaker opponents by large amounts, which could introduce potential biases in the ranking.

2.1.2 Colley method

A similar ranking method is the Colley method, which was developed in 2002 by Wesley Colley [2]. 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} is the number of times teams i and j play each other. The equation for the ranking method is written as

$$Cr = b \tag{2.2}$$

where C is the Colley matrix, r is the unknown rating vector, and b is a vector of cumulative wins and losses. The following equations are appropriate for the matrix and win-loss vector:

$$C_{ij} = \begin{cases} 2 + t_i & i = j \\ -n_{ij} & i \neq j \end{cases} \tag{2.3}$$

$$b_i = 1 + \frac{1}{2}(w_i - l_i) \tag{2.4}$$

Again, solving the system of linear equations for the unknown rating vector will provide a ranking of the teams. A shortcoming, however, 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.

2.1.3 Keener's method

Next, we examine Keener's method [12], which was developed by James Keener in 1993 to rank college football teams. The fundamental equation for Keeners method is

$$Ar = \lambda r \tag{2.5}$$

where A is a matrix that satisfies the eigenvector r and the eigenvalue λ. If the matrix A is irreducible, the Perron-Frobenius theorem guarantees the existence and uniqueness of the ratings vector. There are various approaches to obtain the Keener matrix, but generally, it is formed by taking the ratio of scores between two teams. Let S_{ij} be the number of points scored by team i against team j, then the following equation holds:

$$A_{ij} = h\left(\frac{S_{ij} + 1}{S_{ij} + S_{ji} + 2}\right) \tag{2.6}$$

In general, h is a smoothing function that minimizes the effect of a team running up the score against its opponents and mitigates differences at the extremes, and can look similar to

$$h(x) = \frac{1}{2} + \left(\frac{1}{2}\right)sin\left(x - \frac{1}{2}\right)\sqrt{|2x - 1|} \tag{2.7}$$

The primary weakness of the Keener method is that the strength of schedule can oftentimes be too influential, and playing many weak opponents can cause more harm than good. Since many teams do not control their schedule, it seems counterintuitive to penalize a team for defeating teams that they were required to play.

2.1.4 Analytic Hierarchy Process (AHP)

Developed by Thomas Saaty, the Analytic Hierarchy Process (AHP) is yet another method that uses pairwise comparisons to populate a matrix in which the dominant eigenvector is the rating vector [3, 13, 14]. In this sense, there are many parallels between the AHP and Keeners method previously discussed. The AHP is a tool for decision makers that need to make complex, multi-criteria decisions. The AHP is not only versatile in application, but also a widely used tool in many developing countries.

The foundation of this method is a reciprocal pairwise comparison matrix, in which a user inputs the preference of one alternative to another. The preference value placed for one alternative to another will be the inverse of the reverse relationship. Simply put,

$$A_{ij} = \frac{1}{A_{ji}} \tag{2.8}$$

Once the entire matrix is obtained, the dominant eigenvector is the rating vector. Oftentimes in this application the user is interested in simply making a decision, so the rating vector is not necessary, just the ranking vector.

A drawback of AHP is that it relies fully on the user's preferences to obtain its rating vector. This also has advantages, however, in that a personalized ranking can be obtained for a specific user.

2.2 Popular Pairwise Comparison Ranking Methods using Markov Chains

In this section, we outline a subset of pairwise comparison rankings that use Markov chains to produce ratings and rankings. The fundamental input is to obtain a stochastic matrix that represents the transition probabilities between alternatives, and then to obtain the steady-state probability vector of that matrix, which will correspond to the desired rating vector.

The advantage to these methods is that they take the quality of each individual opponent into consideration, meaning a victory over a strong team will help more than a victory over a weak team. Although previous methods, such as the Massey and Colley methods, take strength of schedule into account, they do not take individual results into account.

2.2.1 PageRank algorithm

Perhaps the most influential ranking method of recent memory is the PageRank algorithm developed by Sergey Brin and Lawrence Page [4]. Google uses this method to produce ratings for web pages, and then obtain a ranking of web pages. Each web page can be thought of as a node in a discrete finite Markov chain, and the edges are defined by hyperlinks between the web pages. An adjacency matrix is constructed that contains zeros and ones depending on whether or not there exists a connection between two nodes. Several adjustments are made to the adjacency matrix to ensure that it is stochastic, and finally the dominant eigenvector is the rating vector. It is also acceptable to use the steady-state probability vector as the rating vector, as we will see in later chapters. Oftentimes, when the matrix is large and difficult to solve, one can use the Power method to obtain the steady-state probability vector.

2.2.2 Random Walker method

The Random Walker ranking method was introduced by Callaghan, Mucha, and Porter to provide a ranking for Division I-A football teams [5]. The motivation was to find a methodology that factored in the strength of opponents played. This is especially important in college football where there are a dearth of games played and many teams with vastly different schedules.

The idea behind this method is that there are a collection of automated voters (random walkers) that declare preferences for a single team. Each voter will randomly select a game from that teams schedule and decide whether or not to change its preference to the opponent based on the outcome of the game. This process is repeated and eventually there is a steady-state distribution of voters for each team, which is equivalent to that team's rating. In this sense, the method strongly resembles the Markov method (which will be introduced later).

For team i, let w_i equal the number of wins, let l_i equal the number of losses, let N_{ij} equal the number of games played between team i and team j, and let A_{ij} equal the number of times team i beats team j minus the number of times team i loses to j. If team i beats team j, the average rate at which an automated voter changes its preference from team j to team i is proportional to p. The primary matrix, D, is defined by the following equations:

$$D_{ii} = -pl_i - (1-p)w_i \tag{2.9}$$

$$D_{ij} = \frac{1}{2}N_{ij} + \frac{(2p-1)}{2}A_{ij} \tag{2.10}$$

Finally, the steady-state vector v satisfies the following equation:

$$Dv = 0 \tag{2.11}$$

The vector v is the expected population of random walkers that will vote for each team, and that information is used directly to rank teams, in that the population of random walkers for each team is equivalent to the rating of that team.

2.2.3 Logistic regression / Markov chain method (LRMC)

The LRMC ranking method was developed by Kvam and Sokol, and is an extension of the Random Walker ranking method [6]. The primary purpose of this ranking method was to use the information from point scores and home court advantages to rank teams in NCAA Division I men's college basketball.

What separates this method from other methods that use Markov chains is that it uses logistic regression to estimate the transition probabilities. However, once those values are obtained, the remainder of the method is to simply calculate the stationary (steady-state) vector which will be equivalent to the rating vector. This method was very successful at predicting team performance in the NCAA tournament, especially in comparison to some of the competing ranking methods.

2.2.4 Park-Newman method

In 2005, the Park-Newman ranking method was developed to rank college football teams [9]. This is another ranking method that indirectly uses the concept of Markov chains to rank its alternatives. The main difference with this method is that it takes indirect wins into account. An indirect win is when team i defeats team j, and team j defeats team k, we say that team i had an indirect win of degree 2 over team k. Indirect wins of higher dimension are also considered but have less impact as the degree increases. A major task when using this method is to develop a value for the parameter that will determine the magnitude of the impact of indirect wins.

2.2.5 Markov method

Last, we examine the general form of the Markov method [7, 10], which was used by Govan in 2008 to rank sports teams, and was recently highlighted by Langville and Meyer in their text, "Who's No. 1? The Science of Rating and Ranking." The Markov method can be thought of a pairwise comparison ranking method that uses Markov chains to rate and rank its alternatives.

The main concept of the method is that each individual competition between two alternatives (or teams) results in the losing alternative voting for the winning alternative. These collection of votes will populate a square matrix that represents the head-to-head competitions between all of the alternatives. There are many ways to construct this voting matrix. For example, a voting matrix could contain information on just wins and losses, and another voting matrix could contain information score differentials. In this document, we will use the basic form of voting for wins and losses, and will refer to this as the (0, 1) voting scheme.

Next, we transform the voting matrix into a stochastic matrix. The stochastic matrix will ultimately provide the steady-state probability vector and thus, the rating vector.

We now introduce an example of the Markov method. Say that we are given four alternatives, {A, B, C, D}, and the following table that displays their head-to-head match results.

Table 2.1
Win-Loss Records, Markov method example

Win-Loss Records	A	B	C	D
A	-	3-2	4-1	4-1
B	2-3	-	2-3	5-0
C	1-4	3-2	-	3-2
D	1-4	0-5	2-3	-

Next, we develop the voting matrix. Each time an alternative loses to another alternative, they will place a vote for that alternative in the matrix. The voting matrix is as follows:

$$V = \begin{bmatrix} 0 & 2 & 1 & 1 \\ 3 & 0 & 3 & 0 \\ 4 & 2 & 0 & 2 \\ 4 & 5 & 3 & 0 \end{bmatrix}$$

We then normalize the rows of the voting matrix to develop a stochastic transition probability matrix, P.

$$P = \begin{bmatrix} 0 & \frac{1}{2} & \frac{1}{4} & \frac{1}{4} \\ \frac{1}{2} & 0 & \frac{1}{2} & 0 \\ \frac{1}{2} & \frac{1}{4} & 0 & \frac{1}{4} \\ \frac{1}{3} & \frac{5}{12} & \frac{1}{4} & 0 \end{bmatrix}$$

Finally, we solve for either the dominant eigenvector of this matrix, or use the Power method to obtain the steady-state probability vector. Both approaches will yield the following ratings and rankings:

The major advantage of the Markov method is that 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. In Chapter 3, we will introduce a new application for the Markov method to solve a traditional recruiting problem, and compare the results to a leading ranking service in terms of their predictive ability.

Table 2.2
Ratings and Rankings, Markov method example

Rank	Alternative	Rating
1	A	0.317
2	B	0.283
3	C	0.257
4	D	0.143

A major drawback of the Markov method, however, is that it is sensitive to small changes in data, especially in its tail, and can exhibit faulty behavior under these circumstances [11]. Intuitively, one can think of an *upset* in sports as a small change in data, and the Markov method has been shown to perform poorly under these conditions. In Chapter 4, we aim to address the issue of sensitivity to erratic data by modifying the traditional voting scheme.

3. CROWD-RANKING METHOD

3.1 Introduction

In general, for a set of n alternatives, the rank of an alternative is its relative importance to the other alternatives in the set. Often, a ranking method will first produce ratings of the alternatives. Next, sorting the alternatives in order of decreasing ratings will provide a ranking.

Many ranking algorithms use pairwise comparisons between the alternatives to determine ratings and rankings. Kendall and Smith [15] introduced the Method of Paired Comparisons, which developed rankings of alternatives based on a set of pairwise preferences between alternatives. Arpad Elo's Rating System [16] is another example of a rating method that uses the results of head-to-head competitive matches to provide ratings, and has been used primarily in the rating of chess players. The Analytic Hierarchy Process [3] is a popular ranking method that uses pairwise comparisons to populate a reciprocal dominance matrix, in which the dominant eigenvector of the matrix is the rating of the alternatives.

In addition to using pairwise comparisons, some ranking methods use Markov chains [17] to develop ratings. Google uses the PageRank method [4] to rank its webpages when returning search results, which contains a series of pairwise comparisons embedded in its algorithm. There also exist methods that use pairwise competitive matches to estimate conditional probabilities [5, 6] that ultimately provide ratings. There is also the Markov method [7] that directly uses Markov chains to rate its alternatives by connecting them through a voting process based on pairwise results.

The Crowd-Ranking method is an extension of the Markov method. It uses the basic theory of the method as the foundation of the ranking algorithm. The basic idea of the Markov method can be described by voting, in that the weaker alternative will place a vote for the stronger alternative. These votes populate a dominance matrix in which the steady-state probability vector is the rating of the alternatives. There are many ways to determine the voting scheme for the method. In Crowd-Ranking, we use a simple approach in that the losing alternative places one vote for the winning alternative.

The remainder of this chapter is organized as follows. Section 2 will provide the problem definition regarding the ranking method. Section 3 will define the ranking method and algorithm. Next, in Section 4, a perfect-case scenario is observed to validate the theoretical basis of the ranking method. Section 5 will provide a case study application of Crowd-Ranking to Big Ten football recruiting from $2002-2013$, and compare its predictability to a leading recruiting ranking service, Rivals. Last, Section 6 provides a conclusion and discussion points.

3.2 Problem Definition

This study is to focus on a specific scenario, in which a set of groups are recruiting a set of individuals to join the groups. The main objective is to rank the groups in order of the quality of individuals that they ultimately obtain. This problem can be thought of as a typical recruiting process, where groups recruit individuals to join them. The scenario is defined below:

1. There are n groups, $\{G_1, G_2, ..., G_n\}$, that extend offers to a set of m individuals, $\{I_1, I_2, ..., I_m\}$.

2. An individual group can extend offers up to the entire set of individuals. Each individual, however, may only select one group to join.

There are several cases in application where the above scenario holds true. For example, a university can be defined as a group that extends admittance to a set of individual students, and the students can only select one university to attend. Another example is to consider a company (group) that extends job offers to a set of individual prospective employees (individuals), and the prospective employees may only select one job offer. Last, consider the case of collegiate athletic recruiting, specifically, in NCAA football. The college football teams (groups) will extend scholarship offers to a set of individual recruits (individuals), and the recruits may only select one team from the set of offers.

It is important to note that we are ranking one specific crop of individuals for each group, so we won't necessarily have a comprehensive ranking of the groups as a whole. This is because a group is comprised of several crops of individuals over time. However, we could use these individual rankings to feed into an encompassing ranking. For example, in college football, the set of five recruiting classes that comprise a team could be combined to obtain one final team ranking (as will be seen in Section 5 when measuring predictability).

3.3 Methodology

The Crowd-Ranking method relies on a dual-level decision process between two parties: the *groups* and the *individuals*. The groups are the entities being ranked, and must decide on the set of individuals that will receive an offer. The individuals are a large set of decision-makers that select a group based on their available offers. The fundamental input of the ranking algorithm relies on two forms of data:

- An individual's set of offers.
- An individual's selection of a group from its set of offers.

Once the data is obtained, the first step in the ranking algorithm is to develop the voting matrix. When an individual selects a group, which is said to be the "winning" group, the remaining groups that offered the individual and were not selected are said to be the "defeated" groups. The set of defeated groups will place a "vote" for the winning group, and this will be done for the set of all individuals. The voting matrix, V, will take the following form:

$$V = \begin{bmatrix} 0 & v_{12} & v_{13} & \dots & v_{1n} \\ v_{21} & 0 & v_{23} & \dots & v_{2n} \\ v_{31} & v_{32} & 0 & \dots & v_{3n} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ v_{n1} & v_{n2} & v_{n3} & \dots & 0 \end{bmatrix}$$

In the voting matrix, v_{ij} is the total number of votes from G_i to G_j. Each matrix entry indicates the total number of preferences from individuals regarding the groups. For example, the entry $v_{32} = 7$ indicates that there were seven individuals that selected group G_2 that had offers from group G_3. In turn, the entries of the voting matrix can be thought of as a series of pairwise comparisons that display the relative dominance among the groups.

One can also think of the voting matrix as a collection of wins and losses in terms of individuals between all of the groups in their respective contests, with a contest being a head-to-head matchup between two groups that offered an individual, and the winner of the matchup being the selected group. There is always one winner per individual, the selected group, and there can be any number of losers, depending on how many offers were received by that individual. Notice that a strong individual with many offers will have a larger impact on the rankings than an individual with only a few offers. Also, the quality of the offers are important, meaning offers from groups with higher rankings will carry more weight than from groups with lower rankings.

Based on the voting matrix, we can use the principles from the Markov method to rank the relative dominance of the groups [7, 10]. The Markov method assumes that the voting matrix can be represented as a Markov chain. If a random walk is taken along the Markov graph, the long-run proportion of time spent at each group will be the rating of that groups strength.

The next step is to normalize the rows of the voting matrix in order to obtain a transition probability matrix. The matrix, P, will take the following form:

$$P = \begin{bmatrix} 0 & p_{12} & p_{13} & \dots & p_{1n} \\ p_{21} & 0 & p_{23} & \dots & p_{2n} \\ p_{31} & p_{32} & 0 & \dots & p_{3n} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ p_{n1} & p_{n2} & p_{n3} & \dots & 0 \end{bmatrix}$$

Each matrix entry indicates the probability that the group will vote for the respective group. For example, the entry $p_{32} = 0.25$ indicates that group G_3 will vote for G_2 25% of the time.

The P matrix is now a transition probability matrix from a Markov chain. We can then find the dominant eigenvector of the matrix, which is equivalent to the steady-state probability vector of the matrix. A necessary condition to obtain this steady-state probability vector is to have an ergodic Markov chain. This steady-state probability value for each group becomes a rating of the relative dominance of that group. There exist other ranking methods, which use the dominant eigenvector of a non-negative matrix to develop a ranking [12, 14]. A matrix that gives pairwise dominances of its alternatives will yield an eigenvector solution that displays the relative dominance among its alternatives.

A critical element of this model is that the groups possess a collective wisdom on the quality of individuals that they are offering. In turn, the quality of an individual is weighted by the quality of its offers. For example, if a group obtains an individual with offers from other groups with high ratings, it will gain votes from each of those groups. Because those top groups will now give higher transition probabilities to the winning group, the steady-state probability value and thus rating for that group will increase. There have been recent studies, for example, Herm, Callsen-Bracker, and Kreis [18], which have shown the value of using crowds to predict in sports.

3.4 Perfect Season

To validate the ranking method, a *perfect season* scenario will be introduced and reviewed. The purpose of this is to show that the ranking method is performing and providing results as expected. This is similar to the approach from Chartier et al. [11] when reviewing the sensitivity and stability of various ranking methods.

In defining a perfect season, it is assumed that there is a specific ordering of both the groups and individuals, and the best groups offer and obtain the best individuals. An intuitive explanation of a perfect season is to have a series of matches with no "upsets," where an upset is defined as a weaker group defeating a stronger group in a match.

To start, assume the following preference relationships hold true for the set of all groups and individuals:

$$G_i > G_{i+1}, \forall i \in \{1, \dots, N-1\} \tag{3.1}$$

$$I_i > I_{i+1}, \forall i \in \{1, \dots, M-1\} \tag{3.2}$$

In this example, for simplicity, say that both the number of groups and individuals are equal to five, and that the following set of offers and selections were made based on the above preference relationships. An offer is indicated by a lower-case "x" and the selection of an offer is indicated by a capital and bold "Y."

Table 3.1
Offers and selections for perfect season

	G_1	G_2	G_3	G_4	G_5
I_1	**Y**	x	x	x	x
I_2		**Y**	x	x	x
I_3			**Y**	x	x
I_4				**Y**	x
I_5					**Y**

As discussed, there are no upsets in the perfect season. Group 1 defeated all of the other groups for the best individual, Group 2 defeated the lower three groups for the $2nd$ best individual, and so on. The next step in the algorithm is to compile the voting matrix, which in this case will simply be a lower triangular matrix with values of one.

$$V = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 \\ 1 & 1 & 1 & 0 & 0 \\ 1 & 1 & 1 & 1 & 0 \end{bmatrix}$$

There is an issue with the voting matrix in that an undefeated group creates an absorbing state, because there were no losses and hence no votes. If you were to normalize the rows of the voting matrix, the first row would still contain all zeroes. A popular strategy for handling an undefeated group is obtained from the PageRank method and its "dangling node" adjustment [19]. This is also the strategy used by Chartier et al. [11] when observing the Markov method for a perfect season. The adjustment is to add a value of $\frac{1}{n}$ to the row in the transitional probability matrix of the absorbing state. Thus, the resulting matrix will be stochastic and can be solved to obtain a steady-state probability vector. Intuitively, think of this step as restarting the random walk of the Markov chain.Below is the stochastic transitional probability matrix, P.

$$P = \begin{bmatrix} \frac{1}{5} & \frac{1}{5} & \frac{1}{5} & \frac{1}{5} & \frac{1}{5} \\ 1 & 0 & 0 & 0 & 0 \\ \frac{1}{2} & \frac{1}{2} & 0 & 0 & 0 \\ \frac{1}{3} & \frac{1}{3} & \frac{1}{3} & 0 & 0 \\ \frac{1}{4} & \frac{1}{4} & \frac{1}{4} & \frac{1}{4} & 0 \end{bmatrix}$$

To solve the steady-state probability vector, it is acceptable to either calculate the dominant eigenvector or use the Power Method as discussed previously. The resulting rating vector for this example is as follows:

$$\pi = \begin{bmatrix} \frac{60}{137} \\ \frac{30}{137} \\ \frac{20}{137} \\ \frac{15}{137} \\ \frac{12}{137} \end{bmatrix}$$

The ratings produce a coherent ranking of the five groups to what was expected from the initially assumed preferences. Generally, for n groups, the rating for a perfect season with the Markov method is given as follows [11], where $H(n)$ is the nth partial sum of the harmonic series:

$$\pi = \frac{1}{H(n)} \begin{bmatrix} 1 \\ \frac{1}{2} \\ \frac{1}{3} \\ \vdots \\ \frac{1}{n} \end{bmatrix}$$

In turn, for a perfect season, it can be seen that the Crowd-Ranking method produces a *correct* ranking in terms of the initial preference relationships for all of the groups. However, it should be noted that the Crowd-Ranking method is an extension of the Markov method, and that the Markov method can be extremely sensitive to upsets, especially in its tail [11].

3.5 Application of Crowd-Ranking to NCAA Football Recruiting

To demonstrate the Crowd-Ranking method in a case study application, it was applied to NCAA football recruiting in the Big Ten (B1G) Conference of the FBS (Football Bowl Subdivision). In this study, the groups were Big Ten football teams' recruiting classes (excluding Nebraska, since they recently joined the Big Ten in 2011), and the individuals were the prospective football recruits. The data for scholarship offers and team selection were available online through Rivals (www.rivals.com), a subscription-based online recruiting website. The data from 2002 – 2013 were used in this study.

The Rivals website provides the data in its individual player pages of all of the teams that have offered a football scholarship to that particular recruit. The recruit will select one team by the end of the recruiting process. The selected team will receive votes from all other teams that extended this recruit a scholarship offer. As explained in previously, each individual recruit can consist of up to $n-1$ competitive matches among the teams. Table 3.2 is an example of the voting matrix for Big Ten teams during the 2012 recruiting season. Each matrix entry indicates a vote from

Table 3.2
Voting matrix, B1G teams, 2012

	ILL	**IND**	**IA**	**UM**	**MSU**	**MN**	**NU**	**OSU**	**PSU**	**PU**	**WI**
ILL	0	5	6	16	8	4	4	6	4	5	4
IND	3	0	7	10	5	1	5	9	0	4	4
IA	3	1	0	10	4	3	4	6	5	1	4
UM	0	0	2	0	2	1	3	11	1	1	4
MSU	0	1	2	14	0	2	1	10	1	1	4
MN	2	3	5	3	3	0	1	7	0	2	3
NU	1	1	3	3	2	0	0	3	2	2	2
OSU	0	0	1	6	2	2	1	0	0	0	1
PSU	0	0	1	6	1	0	0	7	0	1	0
PU	0	3	4	4	5	0	1	8	0	0	1
WI	3	0	4	5	4	1	0	5	1	1	0

one team to another. For example, the entry (ILL, MSU) = 8 means that there were

8 recruits that selected MSU (Michigan State) that had scholarship offers from ILL (Illinois). Notice that the voting matrix is not a symmetrical matrix.

The next step is to normalize the rows of the voting matrix so that the sum of each row in the voting matrix is equal to a value of one. Table 3.3 is an example of the row-normalized matrix for Big Ten teams during the 2012 recruiting season. The values in Table 3 are the probabilities that each team will vote for the other

Table 3.3
Transition probability matrix, B1G teams, 2012

	ILL	**IN**	**IA**	**UM**	**MSU**	**MN**	**NU**	**OSU**	**PSU**	**PU**	**WI**
ILL	0	0.08	0.10	0.26	0.13	0.07	0.07	0.10	0.07	0.08	0.07
IN	0.06	0	0.15	0.21	0.10	0.02	0.10	0.19	0	0.08	0.08
IA	0.07	0.02	0	0.24	0.10	0.07	0.10	0.15	0.12	0.02	0.10
UM	0	0	0.08	0	0.08	0.04	0.12	0.44	0.04	0.04	0.16
MSU	0	0.03	0.06	0.39	0	0.06	0.03	0.28	0.03	0.03	0.11
MN	0.07	0.10	0.17	0.10	0.10	0	0.03	0.24	0	0.07	0.10
NU	0.05	0.05	0.16	0.16	0.11	0	0	0.16	0.11	0.11	0.11
OSU	0	0	0.08	0.46	0.15	0.15	0.08	0	0	0	0.08
PSU	0	0	0.06	0.38	0.06	0	0	0.44	0	0.06	0
PU	0	0.12	0.15	0.15	0.19	0	0.04	0.31	0	0	0.04
WI	0.13	0	0.17	0.21	0.17	0.04	0	0.21	0.04	0.04	0

teams. For example, the entry (ILL, MSU) = 0.129 means that ILL will vote for MSU 12.9% of the time. The above row-normalized matrix is now analogous to the one-step transition probability matrix for a Markov chain, and can be used to obtain the rating vector of the teams. Table 3.4 is the steady-state probability vector and resulting ranking for the 2012 Big Ten recruiting classes.

Table 3.4
Ranking and ratings for B1G recruiting classes, 2012

Rank	**Team**	**Rating**
1	UM	0.2344
2	OSU	0.2246
3	MSU	0.1062
4	WI	0.0939
5	IA	0.0922
6	NU	0.0649
7	MN	0.0627
8	PSU	0.0361
9	PU	0.0359
10	ILL	0.0276
11	IND	0.0215

3.5.1 Comparison of Crowd-Ranking to Rivals

The Crowd-Ranking method will now be compared to Rivals, which is a leading provider of recruiting news, information, and rankings for NCAA football. To calculate rankings, Rivals first calculates the total points each team obtains from its recruiting class. The total points for a recruiting class is the summation of the individual points for the top 20 recruits. Rivals has a team of analysts that assign a point value to each recruit. Table 3.5 shows a comparison of the two methods in 2012. The ratings for the Rivals method have been normalized out of 100 total points for comparison.

It can be seen that both ranking methods perform similar to one another not only in rankings, but in the distribution of ratings. Table 3.6 shows the difference in ranking for each team between these two methods in 2012.

Table 3.5
Rivals vs. Crowd-Ranking, 2012

Rivals			**Crowd-Ranking**		
Rank	Team	Rating	Rank	Team	Rating
1	OSU	23.14	1	UM	23.44
2	UM	20.71	2	OSU	22.46
3	PU	9.76	3	MSU	10.62
4	MSU	8.51	4	WI	9.39
5	IA	8.28	5	IA	9.22
6	PSU	6.88	6	NU	6.49
7	WI	5.70	7	MN	6.27
8	NU	5.10	8	PSU	3.61
9	ILL	4.46	9	PU	3.59
10	IND	4.23	10	ILL	2.76
11	MN	3.23	11	IND	2.15

Notice that there were a few teams with significant differences in rankings. For example, in 2012, the Rivals method ranked Purdue (PU) $3rd$ and Wisconsin (WI) $7th$, while the Crowd-Ranking method ranked Wisconsin $4th$ and Purdue $9th$. The reason behind the difference is that in 2012 Purdue had 26 recruits and Wisconsin had only 12 recruits. This illustrates that although Purdue had a large number of recruits, they werent recruits that were offered by other Big Ten teams. Wisconsin, on the other hand, had a smaller number of recruits in 2012, but the recruits were offered by many other higher ranked Big Ten teams. Table 3.7 and Table 3.8 show the rankings of B1G teams from $2002 - 2013$ by both methods.

Table 3.6
Crowd-Ranking vs. Rivals rankings, 2012

Team	**Rivals**	**Crowd-Ranking**
ILL	9	10
IND	10	11
IA	5	5
UM	2	1
MSU	4	3
MN	11	7
NU	8	6
OSU	1	2
PSU	6	8
PU	3	9
WI	7	4

3.5.2 Predictive power of Crowd-Ranking and Rivals

In this section, we examine the predictive power of the two ranking methods. Since the Big Ten conference does not issue its own set of rankings, we will examine the results of all Big Ten games (excluding Nebraska from $2011-2013$) from $2006-2013$, and compare Crowd-Ranking to Rivals in their ability to accurately pick the winner. Also, we will look at the Big Ten champion(s) each year, and see which ranking method had the team(s) ranked higher.

First, we need to establish a composite ranking for each team in each season. Since college football players can remain on their team for up to five years, we used the sum of the previous five recruiting class rankings to obtain a composite team ranking. For example, in 2009, the composite ranking for Illinois will be the sum of their recruiting class rankings from 2005-2009, since those are the classes that directly impact the team.

Table 3.7
Rivals B1G recruiting class rankings, 2002 – 2013

Team	**’02**	**’03**	**’04**	**’05**	**’06**	**’07**	**’08**	**’09**	**’10**	**’11**	**’12**	**’13**
ILL	6	2	8	8	4	3	4	5	8	7	9	6
IND	11	8	10	11	11	11	11	9	11	9	10	3
IA	8	7	7	2	6	5	8	10	5	3	5	8
UM	2	1	1	1	3	1	2	2	2	2	2	2
MSU	4	9	2	7	5	7	7	3	4	4	4	4
MN	9	4	9	10	9	9	3	6	6	8	11	11
NU	10	10	11	9	10	8	10	8	9	10	8	7
OSU	1	6	3	3	2	2	1	1	3	1	1	1
PSU	3	11	4	4	1	4	6	4	1	5	6	5
PU	5	3	5	5	8	10	9	11	7	11	3	9
WI	7	5	6	6	7	6	5	7	10	6	7	10

The first metric we will examine is the total number of games accurately predicted by each ranking method. Next, we will see how many games that Crowd-Ranking predicted correctly and Rivals predicted did not predict correctly, and how many games that Rivals predicted correctly and Crowd-Ranking did not predict correctly. This is important because many of the games will involve both ranking methods predicting either correctly or incorrectly, but we are concerned with when one method out-performs the other method. This approach is similar to Kvam and Sokol [6] when comparing the predictability of several ranking methods. Also, when a ranking method has the same ranking for both teams (the event of a tie), we will not consider this to be a correct prediction.

Table 3.9 shows the results for the Rivals and Crowd-Ranking (CR) methods for all Big Ten games played from 2006 – 2013.

Table 3.8
Crowd-Ranking B1G recruiting class rankings, 2002 – 2013

Team	**'02**	**'03**	**'04**	**'05**	**'06**	**'07**	**'08**	**'09**	**'10**	**'11**	**'12**	**'13**
ILL	3	4	7	9	8	5	1	5	9	7	10	6
IND	11	11	9	11	11	11	11	9	11	8	11	7
IA	7	8	8	2	4	2	8	8	4	4	5	9
UM	1	1	2	5	3	6	3	2	1	6	1	1
MSU	5	5	4	6	5	7	7	4	5	3	3	3
MN	10	10	11	7	9	10	5	7	8	10	7	11
NU	9	9	10	10	10	9	9	10	10	11	6	8
OSU	2	3	1	3	2	1	2	1	2	1	2	2
PSU	4	6	3	8	1	4	4	3	3	5	8	4
PU	8	2	5	1	7	8	10	11	6	9	9	10
WI	6	7	6	4	6	3	6	6	7	2	4	5

Crowd-Ranking was more successful at predicting the winner over the eight seasons, and had significantly more cases in which it predicted the correct winner and Rivals did not. Also, there was not a single season in which the Rivals method predicted more correct games than the Crowd-Ranking method.

Next, we look at the one-tailed significance results when comparing Crowd-Ranking to Rivals. A one-tailed version of McNemars test was used, similar to the approach by Kvam and Sokol [6]. This test is used because we are concerned with the differences in the ranking methods, not necessarily when both accurately predict a winner, or when both do not predict the winner. Using McNemars test, we find a Chi-squared value equal to 6.62 and a p-value equal to 0.005, indicating that Crowd-Ranking performed significantly better than Rivals.

Table 3.10 shows the Big Ten champions from 2006-2013, and the rankings given to those teams by both methods. Notice that the Crowd-Ranking method ranked the

Table 3.9
Crowd-Ranking vs. Rivals in B1G games, 2006 – 2013

Year	**Rivals**	**CR**	**Rivals >CR**	**CR >Rivals**	**Total Games**
2006	29	32	0	3	44
2007	30	32	2	4	44
2008	27	27	2	2	44
2009	24	27	1	4	44
2010	26	27	2	3	44
2011	26	27	2	3	41
2012	28	30	0	2	40
2013	28	32	0	4	41
Total	218	234	9	25	342

Table 3.10
Crowd-Ranking vs. Rivals to predict champion, 2006 – 2013

Year	**Big Ten Champs**	**Rivals Ranking**	**CR Ranking**
2006	OSU	2	1
2007	OSU	2	1
2008	OSU, PSU	2, 3	1, 3
2009	OSU	1	1
2010	OSU, MSU, WI	1, 5, 8	1, 5, 5
2011	Wisc.	8	4
2012	Wisc.	8	5
2013	MSU	3	3

Big Ten champion higher in four out of the eight season. In two seasons, both methods ranked the eventual champion equally. In 2008 and 2010, where there were conference co-champions, both methods ranked some of the teams equally, but Crowd-Ranking ranked the remaining co-champion higher than Rivals.

3.5.3 Potential data biases

In this study, the accuracy of the reported scholarship offers of a specific recruit is a primary input of the model. In turn, the results of the comparison above rely on the quality of data posted on the Rivals website regarding scholarship offers received by the recruits.

Also, occasionally a high-quality player accepts an offer early in the recruitment process. In this case, the player may not receive additional offers. This is not to say the player was not coveted by other teams, but just that the recruitment process ended before they could obtain more offers.

Last, teams that recruit nationally and out of the Big Ten region will be subject to bias because not as many Big Ten teams will recruit those players. For example, since Penn State is on the eastern footprint of the region, they will recruit heavily along the east coast and often against Big East and ACC schools. Some quality players from that region might not receive Big Ten offers due to their proximity from the conference footprint, and in turn, Penn State will not receive as many votes.

3.6 Conclusion

In conclusion, a new approach, the Crowd-Ranking method, was proposed for a special recruiting problem. The method reflects the decisions of two stakeholders, the groups and the individuals. This Markov-based method considers both quantity and quality when producing a ranking.

Based on our application in Big Ten football recruiting, the Crowd-Ranking method performed better than Rivals, one of the popular ranking methods, in both predicting future performance in Big Ten football games, and predicting the eventual Big Ten champion(s). The ranking method can be used in any application of a recruiting problem.

4. SENSITIVITY OF THE MARKOV METHOD

4.1 Introduction

The objective of a ranking is to develop an ordered list of a set of alternatives based on their relative importance. A class of ranking methods use paired comparisons between the alternatives to produce ratings and rankings. These paired comparison ranking methods are heavily examined due to their widespread applications, spanning from sports teams [12], to chess players [16], to web search engines [4], to even movie recommendations [20]. The analytic hierarchy process (AHP) is a popular paired comparison ranking method [3] with many applications for multicriteria decision making, such as ranking decision making units [21]. There is also interest in understanding the underlying mechanics of ranking methods, as seen by Chartier and Peachey [22] and their approach to reverse engineer the annual college rankings of the U.S. News and World Report.

Of the pairwise comparison ranking methods, there are several methods that use Markov chains to rank a set of alternatives [7–9, 17]. Google is a recent example of having experienced significant success with using the theory of Markov chains to rank webpages [4]. Some methods use intricate mathematics to obtain the transition probabilities that will populate the transition matrix [5,6]. Kyriacos et al. [23] found that there is additional information regarding preference relationships when using Markov chains to rank alternatives.

The Markov method [7] is a popular ranking method that uses a series of pairwise comparisons to develop its rating and ranking vector. Many recent applications use the principles of the method, and it was recently highlighted by Langville and Meyer [10] in "Who's No. 1?: The Science of Rating and Ranking."

Recently, however, it has been shown that the Markov method has a sensitive ranking vector [11]. Although it can be difficult to evaluate the quality of a ranking, a poor ranking can be defined as one that is not robust and displays extreme sensitivity to small changes in data.

There are other studies that investigate the sensitivity of a ranking method. Burer [24] found that the Colley method [2] can be sensitive under special conditions when ranking college football teams. Ramanathan and Ramanathan [25] investigated the sensitivity of an extension of the AHP method and whether or not it possesses desirable rank reversal properties. Zahir [26] studied if acceptable rank reversal properties exist in the presence of imperfect human behavior.

In 2011, Chartier et al. conducted a study on the sensitivity and stability of various ranking vectors. In this study, they reviewed three popular methods: the Massey method [1], the Colley method [2], and the Markov method [7]. To determine the sensitivity of the methods, the authors used an input rating vector to build a perfect season in which a higher rated alternative defeated a lower rated alternative in each competition. Next, they examined the sensitivity of the ranking method by inflicting a small perturbation to the system. Intuitively, this was the sensitivity of a given ranking method to upsets (cases of a lower-rated alternative defeating a higher-rated alternative). It was found that the Massey and Colley methods generally exhibited insensitivity to upsets. The Markov method, on the other hand, displayed sensitivity to upsets, particularly in its tail. In turn, the Massey and Colley methods were more robust to upset events than the Markov method.

An *upset* can have similar effects outside of just sports ranking applications. For example, recommender systems use ranking methods to provide an ordered list of suggestions for its users. The primary inputs for theses ranking methods rely on past user behavior. A rare or one-time search topic by the user is seen as an upset event, and may cause unpredictable ranking behavior. Thus, insensitive ranking methods are essential for providing robust recommender systems.

The purpose of this study is to propose a modification to the voting scheme of the Markov method that will yield a more robust ranking. The two major issues with the current scheme are 1) the potential of a periodic Markov chain, and 2) the sensitivity of the ranking vector. We will show that the modified voting scheme produces both 1) an aperiodic Markov chain, and 2) a ranking vector with less sensitivity to upsets, particularly in its tail.

The remainder of this chapter is organized as follows. Section 2 will introduce the Markov method in its voting scheme, as described by Langville and Meyer [10], compared to a proposed modified voting scheme. Section 3 will show an example of the current voting scheme producing a periodic Markov chain, while the modified voting scheme is seen to always obtain a Markov chain that is aperiodic. In Section 4, we introduce an example of a perfect season, and compare both voting schemes and how they perform when an upset occurs. Last, in Section 5, we generalize the sensitivity of the two methods by observing the tailing effect of the ratings as the number of alternatives increases.

4.2 The Markov Method

The Markov method can be viewed as a voting process, where each competition allows a weaker alternative to "vote" for a stronger alternative. A collection of votes populates a voting matrix that is the fundamental input of the ranking vector. Next, the rows of the voting matrix are normalized, making it a stochastic matrix. The stochastic matrix can be thought of as a transitional probability matrix of a Markov chain, and the steady-state probability vector of that matrix is the rating vector of the alternatives. The scope of our modification is the compilation of the voting matrix, and all steps in the Markov method following the development of the voting matrix will remain unchanged. A generic voting matrix, V, takes the following form:

$$V = \begin{bmatrix} 0 & v_{12} & v_{13} & \dots & v_{1n} \\ v_{21} & 0 & v_{23} & \dots & v_{2n} \\ v_{31} & v_{32} & 0 & \dots & v_{3n} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ v_{n1} & v_{n2} & v_{n3} & \dots & 0 \end{bmatrix}$$

In the voting matrix, V, v_{ij} is the total number of votes from G_i to G_j.

4.2.1 Voting schemes

In its current form [10], the voting scheme is simple, in that the losing alternative will place a vote of value 1 to the winning alternative. In the case where A_j defeats A_i in a competition, v_{ij} will increase by a value of 1, indicating that the ith alternative voted for the jth alternative. Notice that v_{ji} remains unchanged, indicating that the jth alternative did not vote for the ith alternative. For the remainder of this paper, we will refer to this as the **(0,1) voting scheme**.

In this paper, we will introduce a new voting scheme, which we will refer to as the **(1,α) voting scheme**. In this scheme, both alternatives will vote for each other following a competition. In the case where A_j defeats A_i in a competition, v_{ij} will increase by a designated constant value of $\alpha > 1$. However, the winning alternative will also vote for the losing alternative, meaning that v_{ji} will increase by a value of 1.

4.3 Periodicity of the Markov Method

When developing a ranking of multiple alternatives, it is important that the rank reflects the order of all n alternatives, not just a subset. For example, a ranking that identifies the best alternative, but gives no information regarding the remaining $n1$ alternatives is not a complete ranking.

In this section, we show that when using the (0,1) voting scheme, if the associated Markov chain is periodic, it is possible to obtain an incomplete ranking. However, when using the (1,α) voting scheme, all associated Markov chains will be aperiodic and will give a complete ranking. A chain is said to aperiodic if it has at least one state that can return to itself at an irregular rate.

Consider an example where the number of alternatives $n = 5$. We assume a round robin tournament in which each alternative will have two competitions with each other alternative. Last, assume that alternative A_i defeats alternative A_{i+1} in every competition, except for when A_1 competes with A_2, which results in a split (one win and one loss for both). The tournament would result in the following standings: We

Table 4.1
Win-Loss record for round robin tournament

Rank	**Alternative**	**Win-Loss Record**	**Win Pct.**
T-1st	A_1	$7 - 1$	0.875
T-1st	A_2	$7 - 1$	0.875
3rd	A_3	$4 - 4$	0.50
4th	A_4	$2 - 6$	0.25
5th	A_5	$0 - 8$	0.00

see that although the first two alternatives are tied for first, the remaining alternatives have a clear ordering: $A_3 > A_4 > A_5$.

4.3.1 Periodicity of (0,1) voting scheme

We will now use the Markov method to develop a ranking for the above scenario. Below is the voting matrix obtained from the (0,1) voting scheme.

$$V = \begin{bmatrix} 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 \\ 2 & 2 & 0 & 0 & 0 \\ 2 & 2 & 2 & 0 & 0 \\ 2 & 2 & 2 & 2 & 0 \end{bmatrix}$$

Next, we transform the voting matrix into a transition probability matrix, P, by normalizing the rows of the matrix.

$$P = \begin{bmatrix} 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 \\ \frac{1}{2} & \frac{1}{2} & 0 & 0 & 0 \\ \frac{1}{3} & \frac{1}{3} & \frac{1}{3} & 0 & 0 \\ \frac{1}{4} & \frac{1}{4} & \frac{1}{4} & \frac{1}{4} & 0 \end{bmatrix}$$

Last, we use the matrix P to obtain the steady-state probability vector, which is equivalent to the rating vector of the Markov method.

$$\pi = \begin{bmatrix} \frac{1}{2} \\ \frac{1}{2} \\ 0 \\ 0 \\ 0 \end{bmatrix}$$

One can see from the matrix P that the Markov chain is periodic between the first two states. This is an issue because the steady-state probability vector will have a value of zero for the states 3, 4, and 5. Although the desired relationship is that $A_3 > A_4 > A_5$, our rating vector gives us the relationship that $A_3 = A_4 = A_5$.

4.3.2 Periodicity of (1,α) voting scheme

We will now use the (1,α) scheme for the Markov method to develop a ranking for the tournament example. The voting matrix obtained from the modified voting scheme is below.

$$V = \begin{bmatrix} 0 & \alpha+1 & 2 & 2 & 2 \\ \alpha+1 & 0 & 2 & 2 & 2 \\ 2\alpha & 2\alpha & 0 & 2 & 2 \\ 2\alpha & 2\alpha & 2\alpha & 0 & 2 \\ 2\alpha & 2\alpha & 2\alpha & 2\alpha & 0 \end{bmatrix}$$

Next, we transform the voting matrix into a transition probability matrix, P, by normalizing the rows of the matrix.

$$P = \begin{bmatrix} 0 & \frac{\alpha+1}{\alpha+7} & \frac{2}{\alpha+7} & \frac{2}{\alpha+7} & \frac{2}{\alpha+7} \\ \frac{\alpha+1}{\alpha+7} & 0 & \frac{2}{\alpha+7} & \frac{2}{\alpha+7} & \frac{2}{\alpha+7} \\ \frac{\alpha}{2\alpha+2} & \frac{\alpha}{2\alpha+2} & 0 & \frac{1}{2\alpha+2} & \frac{1}{2\alpha+2} \\ \frac{\alpha}{3\alpha+1} & \frac{\alpha}{3\alpha+1} & \frac{\alpha}{3\alpha+1} & 0 & \frac{1}{3\alpha+1} \\ \frac{1}{4} & \frac{1}{4} & \frac{1}{4} & \frac{1}{4} & 0 \end{bmatrix}$$

Immediately, we can see by inspection that this is not a periodic Markov chain. In fact, since the (1,α) method can have zero entries only in the diagonal, it is not possible to obtain a periodic Markov chain from that voting scheme. In other words, it will always be possible to travel back to a given state in an irregular pattern. To fully complete the example, we will solve the above matrix for say, $\alpha = 3$. Plugging into the matrix P, we obtain the following matrix.

$$P = \begin{bmatrix} 0 & \frac{2}{5} & \frac{1}{5} & \frac{1}{5} & \frac{1}{5} \\ \frac{2}{5} & 0 & \frac{1}{5} & \frac{1}{5} & \frac{1}{5} \\ \frac{3}{8} & \frac{3}{8} & 0 & \frac{1}{8} & \frac{1}{8} \\ \frac{3}{10} & \frac{3}{10} & \frac{3}{10} & 0 & \frac{1}{10} \\ \frac{1}{4} & \frac{1}{4} & \frac{1}{4} & \frac{1}{4} & 0 \end{bmatrix}$$

Last, we use the matrix P to obtain the steady-state probability vector, which is equivalent to the rating vector of the Markov method.

$$\pi = \begin{bmatrix} 0.2557 \\ 0.2557 \\ 0.1860 \\ 0.1609 \\ 0.1417 \end{bmatrix}$$

This rating vector is an accurate representation of the dominance relationship among the alternatives. Unlike the (0,1) voting scheme that did not differentiate between the last three alternatives, the (1,α) voting scheme gives the desired relationship $A_3 > A_4 > A_5$.

4.4 Perfect Season

To compare the sensitivity of the (0,1) voting scheme to the (1,α) voting scheme, we will employ the use of a perfect season scenario, similar to the approach of Chartier et al. [11]. In their case, a perfect season was a round robin tournament in which each alternative had one competition, with the stronger alternative winning each competition. However, we will employ a slight modification for this section, in that a perfect season will be defined as a scenario where all of the alternatives have two competitions between each other, and the stronger alternative wins each individual competition. In other words, there are no upsets that occur (later, we will measure the sensitivity of the schemes by inserting upsets into the season). In the next section, when we generalize the sensitivity findings, we will employ the perfect season used by Chartier et al., in which there is only one competition between each set of alternatives.

In our example, we set the number of alternatives $n = 5$ and assume the following relationship regarding the strength of the alternatives:

$$A_i > A_{i+1}, \forall i \in \{1, \ldots, n-1\} \tag{4.1}$$

Since we assume that no upsets occur, the relationship dictates that the weaker alternative will be defeated by the stronger alternative in each competition.

Given a perfect season for $n = 5$ alternatives, Table 4.2 shows the results of the season in terms of wins and losses.

Table 4.2
Win-Loss record for perfect season

Rank	**Alternative**	**Win-Loss Record**	**Win Pct.**
1st	A_1	$8-0$	1.00
2nd	A_2	$6-2$	0.75
3rd	A_3	$4-4$	0.50
4th	A_4	$2-6$	0.25
5th	A_5	$0-8$	0.00

4.4.1 (0,1) voting scheme

To use the Markov method to rank the alternatives, we begin by obtaining the following voting matrix, V.

$$V = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 \\ 2 & 0 & 0 & 0 & 0 \\ 2 & 2 & 0 & 0 & 0 \\ 2 & 2 & 2 & 0 & 0 \\ 2 & 2 & 2 & 2 & 0 \end{bmatrix}$$

Since the perfect season contains an undefeated alternative, there will be no outgoing votes from that alternatives state, leaving a sub-stochastic transition probability matrix. A strategy used by many [19] to resolve this issue is to replace the row of zeros with a row of values equal to $1/n$.

$$P = \begin{bmatrix} \frac{1}{5} & \frac{1}{5} & \frac{1}{5} & \frac{1}{5} & \frac{1}{5} \\ 1 & 0 & 0 & 0 & 0 \\ \frac{1}{2} & \frac{1}{2} & 0 & 0 & 0 \\ \frac{1}{3} & \frac{1}{3} & \frac{1}{3} & 0 & 0 \\ \frac{1}{4} & \frac{1}{4} & \frac{1}{4} & \frac{1}{4} & 0 \end{bmatrix}$$

The resulting matrix is stochastic and can be solved for a steady-state probability vector.

$$\pi = \begin{bmatrix} \frac{60}{137} \\ \frac{30}{137} \\ \frac{20}{137} \\ \frac{15}{137} \\ \frac{12}{137} \end{bmatrix}$$

The rating vector is consistent with the original ranking of the alternatives in terms of wins and losses. Generally, for a perfect season of n alternatives, the (0,1) voting scheme of the Markov method will yield the following rating vector, where $H(n)$ is the nth partial sum of the harmonic series [11].

$$\pi = \frac{1}{H(n)} \begin{bmatrix} 1 \\ \frac{1}{2} \\ \frac{1}{3} \\ \vdots \\ \frac{1}{n} \end{bmatrix}$$

Although the ranking is consistent and correct, the ratings can become very small as n becomes large, especially in its tail. Later, we will see how this effects the sensitivity of this voting scheme.

4.4.2 (0,1) voting scheme with an upset

Assume that there was an upset in the perfect season. To illustrate an extreme case, assume that the worst alternative, A_5, defeated the best alternative, A_1, in one of their two meetings. The resulting rankings in terms of wins and losses would not change, and the ratings only slightly change. Next, we examine how this upset effects

Table 4.3
Win-Loss record for perfect season with upset

Rank	Alternative	Win-Loss Record	Win Pct.
1st	A_1	$7-1$	0.875
2nd	A_2	$6-2$	0.75
3rd	A_3	$4-4$	0.50
4th	A_4	$2-6$	0.25
5th	A_5	$1-7$	0.125

the Markov method and its rating and ranking of the alternatives. The rating vector is shown below.

$$\pi = \begin{bmatrix} \frac{21}{68} \\ \frac{12}{68} \\ \frac{8}{68} \\ \frac{6}{68} \\ \frac{21}{68} \end{bmatrix}$$

We notice from Table 4.4 that A_5 has jumped from last place to a tie for first place with A_1. Using the Markov method, our ranking would be significantly different than just looking at the wins and losses of the alternatives. Although simply looking at the Win-Loss record does not take into account the quality of a victory, it seems counterintuitive to rank A_5 better than A_2, considering that A_2 had significantly more wins and beat A_5 in both of their competitions.

Table 4.4
Markov method (MM) rating for perfect season with upset

Rank	Alternative	Win-Loss Record	Rating
T-1st	A_1	$7-1$	0.309
T-1st	A_5	$1-7$	0.309
3rd	A_2	$6-2$	0.176
4th	A_3	$4-4$	0.118
5th	A_4	$2-6$	0.088

Nonetheless, adding a single upset to the season changed the ratings for all of the alternatives, and changed the rankings for all of the alternatives. Thus, the sensitivity of the Markov method with a (0,1) voting scheme can be seen in this example. Later in this chapter, we will show mathematically why the voting scheme is sensitive and can exhibit irrational behavior when upsets occur.

4.4.3 (1,α) voting scheme

We will now use the modified (1,α) method to rank the alternatives. In this section, we will use a numerical value for $\alpha = 2$. The rating vector is shown below.

$$\pi = \begin{bmatrix} 0.244 \\ 0.218 \\ 0.196 \\ 0.178 \\ 0.163 \end{bmatrix}$$

Similar to the (0,1) voting scheme, the (1,α) voting scheme produces a correct ranking for the perfect season. Notice that although the range of ratings is smaller, the ranking is identical to the ranking obtained from the (0,1) voting scheme, as well as the ranking obtained from observing only wins and losses.

The value of α will affect the ratings, but not the ranking for a perfect season, as seen in Table 4.5. One can see, however, that as α increases, the rating vector increases its range of values. This should be expected because the reward for winning a competition increases as α increases. In fact, as α becomes substantially large, the

Table 4.5
(1,α) method rankings for perfect season

	$\alpha =$							
	2	**3**	**4**	**5**	**10**	**20**	**50**	**1000**
π_1	0.244	0.271	0.289	0.302	0.335	0.357	0.372	0.383
π_2	0.218	0.226	0.230	0.232	0.237	0.239	0.240	0.240
π_3	0.196	0.192	0.188	0.184	0.175	0.169	0.164	0.160
π_4	0.178	0.166	0.158	0.152	0.139	0.130	0.124	0.120
π_5	0.163	0.146	0.136	0.129	0.114	0.106	0.100	0.096

rating vector will begin to approach a similar distribution as the rating vector for the (0,1) voting scheme.

4.4.4 (1,α) voting scheme with an upset

As we did previously, we now introduce an extreme case upset in the perfect season, in which the worst alternative, A_5, defeated the best alternative, A_1 in one of their two meetings.

First, lets see how this upset effects the (1,α) scheme of the Markov method and its rating and ranking of the alternatives. The rating vector is shown below.

$$\pi = \begin{bmatrix} 0.237 \\ 0.215 \\ 0.193 \\ 0.176 \\ 0.179 \end{bmatrix}$$

From Table 4.6, we notice that A_5 has only jumped up one place in this case, as opposed to jumping up to a tie for first place with A_1 in the (0,1) voting scheme. With the (1,α) voting scheme, our ranking is relatively similar to looking at the wins and losses of the alternatives. Also, notice that the ratings of the alternatives have not changed significantly in this case based on a single upset.

Table 4.6
(1,α) method rating for perfect season with upset

Rank	**Alternative**	**Win-Loss Record**	**Rating**
1st	A_1	$7-1$	0.237
2nd	A_2	$6-2$	0.215
3rd	A_3	$4-4$	0.193
4th	A_5	$1-7$	0.179
5th	A_4	$2-6$	0.176

In general, if we were looking at a traditional perfect season, in which each alternative plays only one match with the other alternatives, we notice that the value of α affects the impact of an upset. Table 7 shows that as the value of α increases, the significance of the upset also increases. In turn, depending on the application, the value of α will determine the tradeoff between the sensitivity of the method and the reward for winning a competition. In this particular example, A_5 jumps up to $3rd$ place when $\alpha = 3$, and then doesnt jump up to $2nd$ place until $\alpha = 6$. A_5 doesnt become tied with A_1 for $1st$ place until α is very large, whereas with the (0,1) voting scheme, it is tied for $1st$ place immediately following the upset.

4.5 Sensitivity Analysis of (0,1) and (1,α) Voting Schemes

In this section, we will show a general relationship that represents the sensitivity of the two voting schemes, particularly in their tail. Again, we begin with a perfect

Table 4.7
(1,α) method rating for perfect season with upset, varying α

	$\alpha =$							
	2	**3**	**4**	**5**	**6**	**10**	**50**	**1000**
π_1	0.227	0.244	0.255	0.262	0.267	0.277	0.288	0.290
π_2	0.214	0.216	0.216	0.214	0.213	0.208	0.197	0.194
π_3	0.193	0.184	0.176	0.170	0.165	0.154	0.135	0.129
π_4	0.175	0.159	0.148	0.140	0.135	0.122	0.102	0.097
π_5	0.191	0.197	0.205	0.213	0.220	0.240	0.277	0.290

season of n alternatives. This time, however, our perfect season will be a round robin tournament with one competitive match between each set of alternatives.

It is difficult to quantify a unit of measurement for the sensitivity of a ranking method to upset events. However, we can indirectly measure the sensitivity by observing how easily an alternative can change their ranking with a small change in performance. For example, if the difference in rating between two consecutive alternatives is very small, it will take only a slight increase in rating value to move up a ranking spot.

The general rating vector for the (0,1) voting scheme will be the same as in the previous section.

$$\pi_{(0,1)} = \frac{1}{H(n)} \begin{bmatrix} 1 \\ \frac{1}{2} \\ \frac{1}{3} \\ \vdots \\ \frac{1}{n} \end{bmatrix}$$

The rating vector for the (1,α) voting scheme is difficult to generalize, but we can find an iterative relationship between π_j and π_{j+1} that will help us with our sensitivity analysis. We begin with developing the general voting matrix, and the associated transition probability matrix.

$$V = \begin{bmatrix} 0 & 1 & 1 & \dots & 1 \\ \alpha & 0 & 1 & \dots & 1 \\ \alpha & \alpha & 0 & \dots & 1 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \alpha & \alpha & \alpha & \dots & 0 \end{bmatrix}$$

$$P = \begin{bmatrix} 0 & \frac{1}{n-1} & \frac{1}{n-1} & \dots & \frac{1}{n-1} \\ \frac{\alpha}{\alpha+n-2} & 0 & \frac{1}{\alpha+n-2} & \dots & \frac{1}{\alpha+n-2} \\ \frac{\alpha}{2\alpha+n-3} & \frac{\alpha}{2\alpha+n-3} & 0 & \dots & \frac{1}{2\alpha+n-3} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \frac{1}{n-1} & \frac{1}{n-1} & \frac{1}{n-1} & \dots & 0 \end{bmatrix}$$

From the matrix, P, we see that two consecutive alternatives can be represented by an iterative relationship, and we obtain the following relationship regarding the final rating vector:

$$\pi_j = \frac{[(j+1)\alpha + n - (j+1)][(j-1)\alpha + n - j]}{[j\alpha + n - (j+1)][(j-1)\alpha + n - (j-1)]}\pi_{j+1} \tag{4.2}$$

This expression is derived from the fact that for any two consecutive alternatives, their associated columns are identical except for the $2x2$ matrix at the intersection of i and j.

Another important relationship we obtain from the transition probability matrix is the ratio of π_1 to π_n. When simplified, we see the following equation for the (1,α) voting scheme in a perfect season.

$$\frac{\pi_1}{\pi_n} = \frac{(n-1)\alpha + 1}{n + \alpha - 1} \tag{4.3}$$

This relationship will be used later in this chapter when we derive the sensitivity ratio expression.

4.5.1 Range of ratings based on voting scheme

The range of ratings is significantly different depending on the voting scheme. In the (0,1) voting scheme, the ratings of the first and last ranked alternatives are related by a ratio of n. In turn, as the number of alternatives increases, the ratio of π_1 to π_n increases as well.

$$\lim_{n\to\infty} \frac{\pi_1}{\pi_n} = n \tag{4.4}$$

For the (0, 1) voting scheme, as the number of alternatives approaches infinity, so does the ratio of the ratings of the first and last ranked alternatives.

Now, lets observe what happens with the same ratio for the (1,α) voting scheme. In this case, the number of alternatives does not affect the range of ratings, as α is a constant value. To better understand the sensitivity, in the next section we obtain a a ratio of the differences of ratings of consecutive alternatives.

$$\lim_{n\to\infty} \frac{\pi_1}{\pi_n} = \lim_{n\to\infty} \frac{(n-1)\alpha + 1}{n + \alpha - 1} = \alpha \tag{4.5}$$

4.5.2 Sensitivity based on incremental rating analysis

We propose to find a ratio that can serve as an indicator of how sensitive the two voting schemes are in general, particularly to a tailing effect. One could look at the difference in ratings between consecutive alternatives to see how much is needed to pass that alternative in the ranking. However, that value alone is not enough; we need to compare the largest and smallest increments in the ratings to see the difference. For a perfect season scenario, the difference in rating values for consecutive alternatives, $\pi_i - \pi_{i+1}$, will decrease as i increases. The difference in rating values of the first and second ranked alternatives will be the largest, and the difference in rating values of the last two alternatives will be the smallest. Thus, we define the following ratio to measure the sensitivity for both voting schemes:

$$R_s = \frac{\pi_1 - \pi_2}{\pi_{n-1} - \pi_n} \tag{4.6}$$

We will find the limit of this ratio as $n \to \infty$. The larger the ratio becomes, the more sensitive the ranking will be, especially in its tail, because the necessary rating points to move up a ranking spot become very small.

(0,1) voting scheme

First, we examine the (0,1) voting scheme and its sensitivity ratio of rating increments, R_s. We start with the following relationship:

$$\pi_{j+1} = (\frac{j}{j+1})\pi_j \tag{4.7}$$

Next, we simplify our ratio by plugging in for values of j, and using the previously derived relationships.

$$R_s = \frac{\pi_1 - \pi_2}{\pi_{n-1} - \pi_n} = \frac{n(n-1)}{2} \tag{4.8}$$

$$\lim_{n\to\infty} R_s = \lim_{n\to\infty} \frac{n(n-1)}{2} = \infty \tag{4.9}$$

The expression shows that the ratio grows polynomially with respect to n, which indicates that increments become extremely small in the tail of the rating vector as n grows large. This indicates extreme sensitivity, because only a small increment is needed to jump up a ranking spot. Also, there is no way to control the sensitivity, in that the number of alternatives is not determined by the user when conducting a complete ranking.

(1,α) voting scheme

Next, we examine the (1,α) voting scheme and its ratio of rating increments, R_s. Again, we start with our previously derived iterative relationship and obtain:

$$R_s = \frac{\pi_1 - \pi_2}{\pi_{n-1} - \pi_n} = \frac{\pi_1[(n-1)\alpha][(n-2)\alpha + 2]}{\pi_n\alpha[(2\alpha + n - 2)(n-1)]} \tag{4.10}$$

We have a relationship for the ratio of π_1 to π_n, so we simplify the above expression to obtain:

$$R_s = \frac{n^3\alpha^2 - 4n^2\alpha^2 + 3n^2\alpha + 5n\alpha^2 - 7n\alpha - 2\alpha^2 + 4\alpha + 2n - 2}{n^3 + 3n^2\alpha + 2n\alpha^2 - 7n\alpha - 4n^2 - 2\alpha^2 + 4\alpha + 5n - 2} \tag{4.11}$$

Again, we are concerned with what the ratio converges to as n grows very large.

$$\lim_{n\to\infty} R_s = \alpha^2 \tag{4.12}$$

In this case, our ratio value does not depend on the number of alternatives, n. Regardless of the number of alternatives, the ratio of increments will be a constant value, indicating that we can control the level of sensitivity. This result explains the example in the previous section, and will an even greater impact to other scenarios that use the Markov method to obtain rankings, especially those that have a large number of alternatives.

It is important to note that this does not indicate that the (0,1) voting scheme is *worse* than the (1,α) voting scheme, only that it is less robust and can exhibit erratic behavior in the case of upsets, especially as n grows large. Table 4.8 summarizes the theoretical findings of the sensitivity of the two voting schemes.

Table 4.8
Sensitivity of voting schemes

	(0,1) voting scheme	**(1,α) voting scheme**
R_s	∞	α^2
$\lim_{n\to\infty}(\frac{\pi_1}{\pi_n})$	n	α

The primary advantage of the (1,α) voting scheme is that the degree of sensitivity is controlled by the user. Again, a higher value of α will lead to more sensitivity but also more dispersion in ratings, whereas a lower value of α will lead to less sensitivity but a condensed range of ratings.

In the (0,1) voting scheme, the sensitivity is dictated by the number of alternatives, and in many applications, such as web-based recommender systems, this number can grow very large.

4.6 Conclusion

In conclusion, we found that the new (1,α) voting scheme eliminates the potential of a periodic Markov chain, which is possible with the (0,1) voting scheme. Periodicity can harm the integrity of a ranking because it ignores potential dominance relationships among alternatives.

We found through both an example and theory that the (1,α) voting scheme exhibits a more robust behavior towards upset events than the (0,1) voting scheme, especially in the tail of its rating vector as the number of alternatives increases.

In everyday context, this improvement can help with all applications of ranking methods that use Markov chains. Since all applications have upset events, whether they occur in sports team rankings or recommender systems, it is imperative to have a ranking method that is less sensitive to those events.

In this particular study, we have focused solely on the rankings of alternatives, and not the rating values. Further investigation is needed to examine the interpretation of the rating values when using the (0,1) voting scheme or the (1,α) voting scheme.

Additional work can be done on finding the optimal value of α based on the application. Again, a larger value of α value provides more confidence in selections but will also be more sensitive to upset events.

5. AXIOMS OF RANKING

5.1 Introduction

It is often necessary to determine the importance of an alternative compared to others in a similar group. The process of ordering a list of alternatives based on their relative strength is referred to as ranking. Generally, a ranking method will develop this list by assigning a rating for each alternative in a set, and then ordering the alternatives in decreasing order of rating. Ranking methods are used for a wide array of applications, including but not limited to sports teams [1, 2, 5–7, 9, 12, 27, 28], web search engines [19, 29, 30], and recommender systems [20, 31].

Pairwise comparison methods are a subset of ranking methods. These methods have been used for many years [17, 32], and are still being used in widespread applications today. In this paper, we focus on pairwise comparison ranking methods with applications primarily to sports. We refer to the alternatives being ranked as teams, and the individual pairwise comparison data as matches or games.

In sports ranking applications, the consequences can be significant. Many sports leagues determine participants for tournaments or playoffs based on the ranking of its teams. Thus, a fair and accurate ranking is essential to properly determine the best team(s) in a league, and the best or most fitting ranking method should be used. There is considerable literature that examines different ranking methods and measures their predictive power and performance [10, 11, 24, 33, 34].

However, it is difficult to rank the ranking methods themselves, 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 develop a list of satisficing axioms that, when followed, guarantee a fair and comprehensive ranking method. In other words, if the axioms are satisfied, the ranking method will not dismiss any of the available information directly obtained from a match result.

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 [10]: the traditional Win-Loss method, the Massey method [1], the Colley method [2], the Markov method [7, 10], and the Elo method [16]. These methods will be evaluated on their ability to satisfy the axioms developed in this paper. Later we introduce a recently proposed modification to the Markov method [34] and show that under certain parametric conditions, the method will satisfy the axioms.

It is important to note that 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 (however, in some cases, a ranking is useful to help develop the initial seeding and placement of teams in 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 [35].

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) in soccer, 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 somewhat 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 between each team. However, all three of the above mentioned leagues setups are acceptable for our study. 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 [36].

The remainder of this chapter is organized as follows: Section 2 will outline the five ranking methods that we will study, and give a brief description of the strengths and weaknesses of each method. Section 3 will introduce the axioms and the motivation behind them. Section 4 will map the five ranking methods from Section 2 to the ranking axioms in Section 3, and determine which methods satisfy which axioms. In Section 5, we conjecture that a recently proposed modification to the Markov method can indeed satisfy all three axioms. Section 6 will discuss our results and future research considerations.

5.2 Ranking Methods

In this section, we outline 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. In Section 4, we will revisit these methods and evaluate their ability to satisfy the set of ranking axioms developed in Section 3.

5.2.1 Win-Loss method

The first method we examine is the traditional Win-Loss method, which is the most commonly used method, especially in professional sports. The method is very intuitive, and requires little to no modeling to obtain its ratings. Simply sum up the total number of wins and losses in competitive matches for all teams, and assign rating values for each team equal to the total number of wins. Some leagues look at win percentage, but in leagues with equal number of matches, this will result in an identical ranking vector.

The advantage to 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. However, 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 than the other team faced. The team that faced stronger opponents is more likely a stronger team and should have a greater rating value, but the Win-Loss method will not identify that rating difference.

5.2.2 Massey method

Kenneth Massey developed the Massey method in 1997 to rank college football teams using the theory of least squares [1]. 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:

$$Mr = p \tag{5.1}$$

In the above equation, 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.

It is important to note that 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 isnt 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.

5.2.3 Colley method

Next, we examine the Colley method, which was developed in 2002 by Wesley Colley [2]. 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:

$$Cr = b \tag{5.2}$$

In this equation, C is the Colley matrix, r is the unknown rating vector, and b is a vector of cumulative wins and losses. The following equations are appropriate for the matrix and vector:

$$C_{ij} = \begin{cases} 2 + t_i & i = j \\ -n_{ij} & i \neq j \end{cases} \tag{5.3}$$

$$b_i = 1 + \frac{1}{2}(w_i - l_i) \tag{5.4}$$

Again, solving the system of linear equations for the unknown rating vector will provide a ranking of the teams. A shortcoming, however, 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 one accounts for total point differential and the other total win differential.

5.2.4 Markov method

The Markov method [7, 10], in its general form, can be thought of as a pairwise comparison ranking method that uses Markov chains to rate and 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 will populate a matrix that represents the head-to-head competitions between all of the teams. Next, transform the voting matrix into a stochastic matrix by normalizing its rows. The 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 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, however, is that it is sensitive to small changes in data, especially in its tail, and can exhibit faulty behavior under these circumstances [11, 34]. In fact, in some extreme cases, teams will have incentive to lose a match to increase their rating.

5.2.5 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. Recently, Nate Silver's FiveThirtyEight blog used the method to successfully rank teams in both the NFL and NBA.

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

$$r_{\text{new}} = r_{\text{old}} + K(S - \mu) \tag{5.5}$$

In this equation, K is a constant determined by the nature of the competition, S is an indicator variable that reflects the outcome of the match, and μ is a logistic function of the difference in the ratings of the two opponents, given by:

$$\mu = \frac{1}{1 + 10^{(r_b - r_a)/400}} \tag{5.6}$$

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, one drawback of the Elo method is that it continuously changes over time, and thus the order of matches for a team can have a significant impact on their Elo rating.

5.3 Ranking Axioms

In this section, we construct a set of axioms that a fair and (data) 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 (i.e., a team should never have incentive to lose a match). To be complete, 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. When two teams compete, we consider the outcome of the match as the information obtained. This includes both the match winner (or in some cases, there is a tie) and the match score. Additional information (such as strategic formations, or roster assignments) could be obtained from a match result based on the nature of the sport, but that level of analytics is beyond our scope of evaluating the ranking methods.

There is some debate as to whether or not the score differential of a match is a good indicator of team performance. On one hand, Redmond [37] 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 [1], 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 axioms we construct.

5.3.1 Axiom I: opponent strength

Our first axiom 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. Clearly, beating the best team indicates that you are a stronger team than if you had beaten the worst team. Thus, a comprehensive ranking method must take into account the quality of a victory when calculating the rating of a team.

On the other hand, not all losses are equivalent. For example, it would be misleading to give a similar penalty for losing to the best team in the league as opposed to losing to the worst team in the league, for the same reasoning as listed above. So, a comprehensive ranking method must also take into account the quality of a loss when calculating the rating of a team.

Axiom 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, the score differential of a match can often be misleading information when calculating team ratings. Thus, the extension for Axiom I is a *soft* axiom, or otherwise an optional axiom.

Axiom 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 of sorts to delineate the impact, similar to Keener's approach [12]. For example, defeating a team 21-0 should not have the same impact as defeating a team 49-28, as the first score shows stronger dominance from the winning team.

5.3.2 Axiom II: incentive to win

The next axiom aims to unify the objective for each competitive match, which is simply to win the match. This axiom is rather self-evident, but is extremely important to the integrity of any ranking method. For example, 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.

In most ranking methods, this axiom will hold true. However, as we will see later in this chapter, 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.

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

It is important to note that the converse of this axiom 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 defeating that same opponent.

Axiom II also indirectly implies that strong interdependence between teams' ratings can have a negative impact on the ranking vector. Chartier et al. [11] 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. We will see in Section 4 how Axiom I and Axiom II will often be in direct conflict with each other due to the nature of their objectives. Basically, the more a ranking method factors in the strength of opponent, the stronger interdependence exists between rating values. This can ultimately lead to cases where there exists an incentive to lose.

5.3.3 Axiom III: sequence of matches

Teams do not select the sequence 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 order 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. However, if a ranking method did not adhere to this idea, it can be viewed as an unfair ranking method. Again, this is not to call attention to the quality of opponents on the schedule, only the order in which the matches are played.

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

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

5.4 Ranking Methods and Axioms

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

5.4.1 Axiom I review

Axiom 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 Axiom I is not satisfied.

The Win-Loss method, the Massey method, and the Colley method violate Axiom 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 5.1.

Table 5.1
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

Now, lets assume that team E had beaten team A, and recalculate the Colley ratings. The ranking is shown in Table 5.2.

Table 5.2
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

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.

The Elo method and the Markov method both adhere to Axiom 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 previously showed, 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:

$$\pi_j = \sum_{i=1}^{n} p_{ij}\pi_i \tag{5.7}$$

From this equation, it can be seen that wins over stronger teams will increase your rating more than wins over weaker teams.

5.4.2 Axiom II review

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

The Win-Loss method, Massey method, Colley method, and Elo method all follow Axiom 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. There are cases in which a team can win a match and their ranking will not improve because the teams ranked ahead also win their matches. However, this does not violate Axiom II, because we are only concerned with teams having incentive to lose to improve their rating.

The Markov method, on the other hand, can have cases where teams have an incentive to lose to improve their rating, thus violating Axiom II. There is a strong interdependency 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.

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 5.3.

Table 5.3
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 5.4.

Table 5.4
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	0.144
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 5.4. However, lets see what happens if team A intentionally loses the match to team D.

Table 5.5
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 5.5, 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.

For a real-world example, consider the 2011 NFL season, in which 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 matchup between GB and the Chicago Bears (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 5.6 shows an excerpt of both the actual 2011 NFL season Markov ratings, and the modified season with the incentive to lose case. 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

Table 5.6
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

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 overinflated ratings for CHI and KC because of their upset victories over GB.

5.4.3 Axiom III review

Axiom III states that the sequence of matches on a team's 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 Axiom III is violated.

The Win-Loss method, Massey method, Colley method, and Markov method all satisfy Axiom III, and no team rating will change based on the order 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 order of matches will not affect the entries of the matrices or vectors.

The Elo method, however, does depend on the order of matches, and thus violates Axiom 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 5.7 shows Elo ratings for the National Football Conference (NFC) in the NFL 2012 season with matches in the actual order. The teams in bold font are the teams that would be selected for the playoffs.

Table 5.7
Elo ratings for NFC in NFL 2012 season, actual 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

Now, lets observe what happens if we simply reverse the sequence of matches when calculating Elo rating values. Of the 16 teams in the NFC, only San Francisco (SF)

Table 5.8
Elo ratings for NFC in NFL 2012 season, reverse 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

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. Clearly, the order of matches will affect the final team rating when using the Elo method, thus Axiom III is violated.

In summary, all five of the ranking methods violated exactly one of the ranking axioms. Table 5.9 provides a summary of our findings.

Table 5.9
Summary of ranking methods and axioms

Method	Axiom I	Axiom II	Axiom III
Win-Loss	×	**YES**	**YES**
Massey	×	**YES**	**YES**
Colley	×	**YES**	**YES**
Markov	**YES**	×	**YES**
Elo	**YES**	**YES**	×

5.5 Proposed Method to Satisfy Axioms

In this section, we propose a recently developed method by Vaziri et al. [34] 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 axioms.

Before we examine this method, it is important to discuss the possibility of tweaking the other methods to satisfy the axioms. For the Win-Loss, Massey, and Colley methods, it is not possible to modify the method to take Axiom 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. Also, there is no individual mapping of a match result to a specific opponent. This is consistent with the findings from Chartier et al. [11] when they showed that the Massey and Colley methods had a uniformly spaced rating vector for a perfect season.

Next, the Elo method cannot be modified to fit Axiom III. The nature of the method depends on the timing of a win or a loss, because the ratings of teams are continuously changing. A potential modification could be to generate all possible orders of matches, and compute an average rating for each team based on every possible order. However, even for leagues with short seasons such as the NFL, there would be $17! \sim 10^{14}$ combinations of schedules.

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" 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 [34], while maintaining the integrity of the rank order.

Since the $(1, \alpha)$ method is a modification of the Markov method, it will follow Axioms 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 Axiom II and not provide incentive to lose.

We are not able to rigorously prove that the $(1, \alpha)$ method will satisfy Axiom 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 Axiom II will be satisfied.

We revisit the previous example, 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 5.10. Again,

Table 5.10
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

we add an upset in which team E instead had defeated team A. The ranking can be seen in Table 5.11.

Table 5.11
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

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 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 5.12,

Table 5.12
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

losing the 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 before, 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 5.13.

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 before (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, Axiom II is satisfied. However, once $\alpha \geq 10$, the incentive to lose existed for GB because their rating increased. On analyzing data from the NFL seasons from 2002 to 2011, we found that for α values less than 5, there is never an incentive to lose a match. For

Table 5.13
$(1, \alpha)$ 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

Table 5.14
$(1, \alpha)$ 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

values 10 or greater, there were instances where losing a match was beneficial for a team.

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 Axiom 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 [34], 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 5.15 provide evidence of this convergence. (In 2007, the New England Patriots were undefeated, and thus as a result of the dangling node adjustment [19], the $(0, 1)$ method did not have an incentive to lose.)

Table 5.15
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

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 Axiom II. A future study is required that finds the relationship and/or threshold of α based on other league parameters to guarantee satisfaction of Axiom II.

It should be noted that our aim is only to show that the $(1, \alpha)$ method satisfies all three axioms, and thus is a fair and complete ranking method. Modification is required based on the sport and tournament structure to improve the predictive power of the $(1, \alpha)$ method. Examples of these modifications include but are not limited to: finding an optimal α that minimizes sensitivity and maximizes confidence level in victory, incorporating home and away advantages, incorporating score differential, etc.

5.6 Conclusion

In summary, we have outlined a set of ranking axioms 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.

We reviewed five popular sports ranking methods and found that none of the five adhered to all three of the axioms, although all of them satisfied exactly two of the axioms. 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 order of matches played, which is oftentimes (and always, in major professional sports) not in the teams' control.

Last, we conjectured that a newly proposed modification to the Markov method, known as the $(1, \alpha)$ method, will satisfy all three axioms under certain parametric conditions. We showed both a generic and case study example where the $(1, \alpha)$ method satisfied all three axioms and removed the previous case of having incentive to lose. However, for large values of α, the methods rating vector converges to the traditional Markov method rating vector, and Axiom II will be violated. Future work is needed to identify a relationship between α and league parameters that guarantees that Axiom II will be satisfied.

6. CONCLUSION

In summary, this dissertation document focused on research pertaining to the (1) application, (2) theory, and (3) framework for Markov-based ranking methods.

Application

In Chapter 3, we introduced the Crowd-Ranking method, which is an extension of the conventional Markov method applied to a recruiting scenario. This method uses collective wisdom and a dual-level decision process to rank a set of groups based on the quality of the individuals that select those groups.

We applied this method to NCAA football recruiting in the Big Ten conference, and developed rankings for Big Ten teams' recruiting classes from 2002−2013. Next, under the assumption that stronger recruiting rankings tend to indicate stronger team performance, we measured the predictive power of Crowd-Ranking against Rivals, a leading service in online recruiting information.

Crowd-Ranking performed significantly better than Rivals at predicting the winner of (1) Big Ten football games and (2) Big Ten conference champions over the course of 2006−2013.

Theory

Next, we examined the theory behind the Markov method, and one of its major drawbacks: sensitivity in its tail, especially to upset events. In turn, we introduced a new voting scheme, the $(1, \alpha)$ method, and compared it to the conventional $(0, 1)$ method in its ability to perform under undesirable conditions.

First, we showed rigorously that the $(1, \alpha)$ method could not obtain a periodic Markov chain, although the $(0, 1)$ method could obtain a periodic Markov chain. This is an issue because a periodic Markov chain will cause a loss of information in the rating vector and ranking.

In addition to the periodicity condition, the $(0, 1)$ method was shown to be extremely sensitive to upsets, particularly in the tail of its rating. We showed, in theory and in practice, that the $(1, \alpha)$ method subsided much of this sensitivity because it allowed the user to *control* the sensitivity of the method with the parameter α. By obtaining a ratio of rating increments, we saw that the $(0, 1)$ method displayed high sensitivity, but the $(1, \alpha)$ method was controlled and less erratic.

Framework

Finally, our last step in the research journey was to develop a framework to evaluate the (1) fairness and (2) comprehensiveness of ranking methods. We introduced three fundamental ranking axioms, and the reasoning behind their inclusion. Next, we introduced five popular sports ranking methods, and examined whether or not they adhere to the said axioms. We found that all of the methods violated exactly *one* axiom, and thus none of them adhered to all three. Last, we introduced our previously discussed $(1, \alpha)$ method, and showed that under certain parametric conditions for α, it did not violate the axioms when applied to the National Football League (NFL) from $2002-2011$. Further research is required to develop relationships and conditions for α values and tournament setup to determine whether or not the $(1, \alpha)$ method can guarantee satisfaction of all three axioms.

LIST OF REFERENCES

LIST OF REFERENCES

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

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

[3] T. L. Saaty, "The analytic hierarchy process: planning, priority setting, resources allocation," *New York: McGraw*, 1980.

[4] S. Brin and L. Page, "Reprint of: The anatomy of a large-scale hypertextual web search engine," *Computer networks*, vol. 56, no. 18, pp. 3825–3833, 2012.

[5] T. Callaghan, P. J. Mucha, and M. A. Porter, "Random walker ranking for ncaa division ia football," *American Mathematical Monthly*, vol. 114, no. 9, pp. 761–777, 2007.

[6] P. Kvam and J. S. Sokol, "A logistic regression/markov chain model for ncaa basketball," *Naval Research Logistics (NrL)*, vol. 53, no. 8, pp. 788–803, 2006.

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

[8] L. C. Ingram, *Ranking NCAA sports teams with Linear algebra*. PhD thesis, COLLEGE OF CHARLESTON, 2007.

[9] J. Park and M. E. Newman, "A network-based ranking system for us college football," *Journal of Statistical Mechanics: Theory and Experiment*, vol. 2005, no. 10, p. P10014, 2005.

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

[11] T. P. Chartier, E. Kreutzer, A. N. Langville, and K. E. Pedings, "Sensitivity and stability of ranking vectors," *SIAM Journal on Scientific Computing*, vol. 33, no. 3, pp. 1077–1102, 2011.

[12] J. P. Keener, "The perron-frobenius theorem and the ranking of football teams," *SIAM review*, vol. 35, no. 1, pp. 80–93, 1993.

[13] T. L. Saaty, "Axiomatic foundation of the analytic hierarchy process," *Management science*, vol. 32, no. 7, pp. 841–855, 1986.

[14] T. L. Saaty, "Rank according to perron: A new insight," *Mathematics Magazine*, pp. 211–213, 1987.

[15] M. G. Kendall and B. B. Smith, "On the method of paired comparisons," *Biometrika*, pp. 324–345, 1940.

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

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

[18] S. Herm, H.-M. Callsen-Bracker, and H. Kreis, "When the crowd evaluates soccer players market values: Accuracy and evaluation attributes of an online community," *Sport Management Review*, vol. 17, no. 4, pp. 484–492, 2014.

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

[20] T. Chartier, A. Langville, and P. Simov, "March madness to movies," *Math Horizons*, vol. 17, no. 4, pp. 16–19, 2010.

[21] S. M. Alem, F. Jolai, and S. Nazari-Shirkouhi, "An integrated fuzzy dea-fuzzy ahp approach: a new model for ranking decision-making units," *International Journal of Operational Research*, vol. 17, no. 1, pp. 38–58, 2013.

[22] T. Chartier and J. Peachey, "Reverse engineering college rankings," *Math Horizons*, vol. 22, no. 2, pp. 5–7, 2014.

[23] T. Kyriacos, Z. Christina, and S. George, "Ranking domain names using various rating methods,"

[24] S. Burer, "Robust rankings for college football," *Journal of Quantitative Analysis in Sports*, vol. 8, no. 2, 2012.

[25] U. Ramanathan and R. Ramanathan, "An investigation into rank reversal properties of the multiplicative ahp," *International Journal of Operational Research*, vol. 11, no. 1, pp. 54–77, 2011.

[26] S. Zahir, "Normalisation and rank reversals in the additive analytic hierarchy process: a new analysis," *International Journal of Operational Research*, vol. 4, no. 4, pp. 446–467, 2009.

[27] R. D. Baker and I. G. McHale, "A dynamic paired comparisons model: Who is the greatest tennis player?," *European Journal of Operational Research*, vol. 236, no. 2, pp. 677–684, 2014.

[28] A. Y. Govan, A. N. Langville, and C. D. Meyer, "Offense-defense approach to ranking team sports," *Journal of Quantitative Analysis in Sports*, vol. 5, no. 1, 2009.

[29] J. M. Kleinberg, "Authoritative sources in a hyperlinked environment," *Journal of the ACM (JACM)*, vol. 46, no. 5, pp. 604–632, 1999.

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

[31] Y. Koren, R. Bell, and C. Volinsky, "Matrix factorization techniques for recommender systems," *Computer*, no. 8, pp. 30–37, 2009.

[32] M. G. Kendall, "A new measure of rank correlation," *Biometrika*, pp. 81–93, 1938.

[33] 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*, vol. 9, no. 2, pp. 187–202, 2013.

[34] 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*, vol. Accepted, In Press.

[35] P. Scarf, M. M. Yusof, and M. Bilbao, "A numerical study of designs for sporting contests," *European Journal of Operational Research*, vol. 198, no. 1, pp. 190–198, 2009.

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

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

VITA

VITA

From a young age, I witnessed the life of an academic through my parents. I loved to learn, but I also loved to teach. Many of my teaching philosophies today stem from experiences in my childhood. In fact, more so than the subject matters, it is the teachers and mentors that I remember and am thankful for helping me in this journey.

Upon receiving my Bachelor's Degree in Industrial and Operations Engineering from the University of Michigan in 2006, I went to work as a process engineer in private industry. Soon enough, I realized my life path belonged in academia, and I returned to Purdue University as a PhD student. I hope to experience a long career in academia working as a researcher, educator, and community leader.

In my free time, I enjoy writing, performing, and listening to music.