A Computationally Intensive Ranking System for Paired Comparison Data

David Beaudoin and Tim B. Swartz *

Abstract

In this paper, we introduce a new ranking system where the data are preferences resulting from paired comparisons. When direct preferences are missing or unclear, then preferences are determined through indirect comparisons. Given that a ranking of n subjects implies $\binom{n}{2}$ paired preferences, the resultant computational problem is the determination of an optimal ranking where the agreement between the implied preferences via the ranking and the data preferences is maximized.

Keywords : Nonparametric methods, NCAA basketball, Ranking, Simulated annealing, Statistical computing.

*David Beaudoin is Associate Professor, Département Opérations et Systèmes de Décision, Faculté des Sciences de l'Administration, Pavillon Palasis-Prince, Bureau 2439, Université Laval, Québec (Québec), Canada G1V0A6. Tim Swartz is Professor, Department of Statistics and Actuarial Science, Simon Fraser University, 8888 University Drive, Burnaby BC, Canada V5A1S6. Both authors have been partially supported by the Natural Sciences and Engineering Research Council of Canada. This research was enabled in part by support provided by Calcul Québec (http://www.calculquebec.ca/en/) and Compute Canada (www.computecanada.ca). The authors thank two anonymous reviewers whose suggestions have improved the manuscript.

1 INTRODUCTION

The problem of ranking can be simply stated and has an extensive literature in the statistical sciences. Given data on n subjects, the objective is to determine a permutation (ranking) $R = (i_1, \ldots, i_n)$ where the interpretation is that subject i_j is preferable to subject i_k whenever $i_j < i_k$. The term "preferable" depends on the application and the methods used to determine the ranking depend on aspects of the data structure.

In sport, ranking is an important problem. For example, in National Collegiate Athletic Association (NCAA) basketball, there are over 300 teams competing in Division I where a typical team plays only a subset ($\sim$25) of the other teams during a season. At the end of the season, the NCAA Selection Committee is set with the task of creating a tournament structure known as "March Madness" involving 68 of these teams. In determining the invitees, team rankings (in terms of team quality) form part of the decision making process.

Similarly, in NCAA football, various team rankings are regularly reported during the regular season (e.g. Associated Press, FCS Coaches' Poll, Sagarin, etc.). Although such rankings are no longer used for determining Bowl bids (i.e. identifying pairs of teams that compete in prestigious holiday matches), the rankings receive considerable media attention and are available to the selection committee. Part of the intrigue involving the determination of the rankings is that there are not many crossover matches involving teams from different conferences.

Ranking also occurs in non-sporting contexts. For example, universities rank students, employers rank job candidates, there are rankings corresponding to the quality of journals, and so on. Clearly, the type of data used to inform the rankings varies greatly on the application.

In this paper, we focus on the ranking problem associated with NCAA basketball. More specifically, we consider the ranking of n Division I teams ($n = 351$ in 2015/2016). The data used to inform our ranking are the result of paired comparisons. Sometimes a comparison is explicit (e.g. based on the result of one team playing another team). In other instances, the comparison between two teams is determined by considering the results of matches involving common opponents. Our approach searches for an optimal ranking $R = (i_1, \ldots, i_n)$ which has maximal agreement between the $\binom{n}{2}$ implied paired

preferences via the ranking and the observed data preferences. The approach is appealing in its simplicity and its lack of assumptions. It may be regarded as nonparametric in the sense that there is no underlying probability model. However, the approach provides computational challenges. For example, a simple search amongst rankings is not possible since there are $n! \approx 10^{742}$ potential rankings.

Ranking methods based on paired comparison data originate from the work of Thurstone (1927) and Bradley and Terry (1952). The approach suggested by Park and Newman (2005) is most closely related to our approach in the sense that it is also nonparametric and extends comparisons to indirect matchups between teams. Park and Newman (2005) rank teams according to team wins w minus team losses l in both direct and discounted indirect matches. The statistics w and l correspond to a matrix-based network centrality measure involving adjacency matrices.

From the seminal work by Thurstone (1927) and Bradley and Terry (1952), the statistical literature on methods for paired comparison data has flourished. For example, many extensions to the original models have been considered such as the provision for ties in paired comparisons (Davidson 1970), multiple comparisons (Plackett 1975), Bayesian implementations (Leonard 1977, Chen and Smith 1984, Caron and Doucet 2012) and dynamic ranking models (Glickman 1999, 2001) which have been used in chess. The treatment of the margin of victory in paired comparison settings has also lead to various models and methods. For example, Harville (1977, 2003) considers linear models where truncations are imposed on large margins of victory. A central idea is that teams should not have an incentive for running up the score. Mease (2003) considers a model based on normal likelihoods and penalty terms that attempts to correspond to human judgments. A general review of the literature related to paired comparison methods is given by Cattelan (2012). Rotou, Qian and von Davier (2015) review methods that are primarily concerned with dynamic rankings where data are frequently generated such as in the gaming industry.

In Section 2, we describe our approach which is intuitive and simple to describe. However, the method gives rise to challenging computational hurdles for which we propose a stochastic search algorithm. For example, we demonstrate how time savings can be achieved in the calculation of our metric which measures the agreement between the im-

plied ranking preferences and the data preferences. The algorithm implements a simulated annealing procedure which optimizes over the $n!$ candidate rankings. Section 3 assesses the proposed ranking procedure by forecasting matches based on established ranks. We first investigate our procedure in the context of real data from previous NCAA basketball seasons. We compare our rankings with rankings obtained by other popular procedures. Our second example is based on simulated NCAA basketball data where the underlying strengths of the teams are specified. This allows us to compare forecasts against the truth. The final forecasting example is based on data from the 2016/2017 English Premier League season. This is a substantially different dataset in that we have a much smaller number of teams ($n = 20$). In Section 4, we consider various nuances related to our approach. In particular, we compare our procedure to the Bradley-Terry approach where we observe the proposed method places more importance on individual matchups than Bradley-Terry. We conclude with a short discussion in Section 5.

2 APPROACH

Our approach is based on data arising from paired comparisons. In basketball, this represents a data reduction since each team scores a specific number of points in a game. However, sometimes the actual number of points scored can be misleading. For example, in "blowouts", teams often "empty their benches" near the end of a game, meaning that regular players are replaced by players who do not typically play in competitive matches. In such cases, margins of victory may not be representative of true quality. Interestingly, a requirement of the computer rankings used in the former BCS (Bowl Championship Series) for NCAA football was that the computer rankings should not take into account margin of victory (dishingoutdimes 2010).

With respect to paired comparison data, it is straightforward to determine the preference when one team plays another team in a single game. We let h denote the number of points corresponding to the home court advantage in NCAA basketball. If the home team defeats the road team by more than h points, then the home team is the preferred team in the paired comparison. Otherwise, the road team is the preferred team.

We set the home team advantage at $h = 3.5$ points. This is consistent with Bessire

(2016) who provided an average home team NCAA basketball advantage of $h = 3.7$ points. When a game is played at a neutral site, then $h = 0$. Whereas there is frequent discussion of differential home team advantages, we are inclined to believe that such differences are primarily the manifestation of multiple comparison issues (Swartz and Arce 2014) and unbalanced schedules. The value $h = 3.7$ roughly agrees with Gandar, Zuber and Lamb (2001) who estimated a home court advantage of 4.0 points in the National Basketball Association (NBA). Since an NBA game is 48 minutes in duration and a college game is only 40 minutes in duration, the mapping from $h = 4.0$ in the NBA to college is $4.0(40/48) = 3.3$. In an independent calculation, we studied pairs of NCAA basketball teams during the 2006/2007 through 2015/2016 seasons. In the 26,206 matchups where pairs of teams played more than once with both home and away matches, we estimated the home court advantage as $h = 3.4$ points. For example, suppose team A played at home and defeated team B by h_A points (h_A is negative for a loss). And similarly, suppose team B then played at home and defeated team A by h_B points (h_B is negative for a loss). In this matchup, home advantage is estimated by $(h_A + h_B)/2$, and h is obtained by averaging these terms over all matchups. For the determination of preferences in paired comparisons, we note that preferences are insensitive to $h \in (3.0, 4.0)$.

More generally, suppose that two teams have played each other more than once. Let $p_{\mathrm{A}i}$ and $p_{\mathrm{B}i}$ be the points scored by Team A and Team B respectively in the i-th game. Then from Team A's perspective, define the differential

$$d_i = \begin{cases} p_{\mathrm{A}i} - p_{\mathrm{B}i} - h & \text{if Team A is the home team} \\ p_{\mathrm{A}i} - p_{\mathrm{B}i} + h & \text{if Team B is the home team} \end{cases} \tag{1}$$

In this case, Team A is the preferred team in the particular paired comparison if the average of its d_i values is positive.

When two teams have played each other directly, then we use (1) to determine the preference, and we refer to this as a level L_1 preference. With $n = 351$ NCAA basketball teams, there are $\binom{n}{2} = 61,425$ potential paired comparisons. Based on the 5,948 matches that took place in 2015/2016, 3,918 level L_1 preferences were observed. The level L_1 preferences represent only 6.38% of the potential $\binom{n}{2}$ paired comparisons.

We now consider cases where Team A and Team B have not directly played against each other. Our approach for determining preferences in these situations borrows on ideas

from the RPI (Ratings Performance Index) where strength of schedule is considered; see Barrow et al. (2013) for a definition of RPI. Specifically, suppose that Team A and Team B have a common opponent Team C. Then (1) can be used to obtain an average differential $\bar{d}_{AC}$ from the point of view of Team A versus Team C. Similarly, (1) can be used to obtain an average differential $\bar{d}_{BC}$ from the point of view of Team B versus Team C. If $\bar{d}_{AC} > \bar{d}_{BC}$, then Team A is the preferred team in the paired comparison of Team A versus Team B. Now, suppose that Team A and Team B have multiple common opponents C_i. In this case, if $\sum_i \bar{d}_{AC_i} > \sum_i \bar{d}_{BC_i}$, then Team A is preferred to Team B. When two teams do not play one another directly but have common opponents, then we refer to the resulting preference as a level L_2 preference. In the 2015/2016 dataset, L_2 preferences represent 54.95% of the potential $\binom{n}{2}$ paired comparisons.

We extend the preference definition so that the data can be used to further determine preferences. Suppose now that Team A and Team B do not play each other directly and that they have no common opponents. However, imagine that Team A has an opponent and that Team B has an opponent who have a common opponent. For example, suppose Team A plays Team C, Team B plays Team D, Team C plays Team E and that Team D plays Team E. Without going into the notational details and using a similar approach as previously described, a differential $\bar{d}_{AE}$ can be determined via the AC and CE matchups. Similarly, a differential $\bar{d}_{BE}$ can be determined via the BD and DE matchups. Then $\bar{d}_{AE}$ can be compared with $\bar{d}_{BE}$ to determine the data preference between Team A and Team B. We refer to preferences of this type as level L_3 preferences. In the 2015/2016 dataset, L_3 preferences represent 38.67% of the potential $\binom{n}{2}$ paired comparisons. We therefore see that $6.38\% + 54.95\% + 38.67 = 100\%$ of the potential $\binom{n}{2}$ paired comparisons are either of levels L_1, L_2 or L_3. Referring to the popular 1993 movie "Six Degrees of Separation" starring Will Smith, we observe three (rather than six) degrees of separation in the 2015/2016 NCAA basketball season.

We now make a small adjustment in the definition of preferences. Occasionally, there are "ties" in the preferences. For example, in the 2015/2016 NCAA basketball season, there were 18 cases out of the 3,918 level L_1 preferences where a tie occurred. This was the result of two teams playing each other twice, one game on each team's home court. In both games, the home team won by the same margin leading to $\bar{d} = 0$. There are various

ways of breaking the tie to determine the preference. For example, you might set the preference according to the most recent match. Our approach which we use throughout the remainder of the paper is to set the preference for both teams equal to 0.5.

Having defined level L_1, L_2 and L_3 preferences using the NCAA basketball data, we note that the data preferences are not necessarily transitive. For example, it is possible that Team A is preferred to Team B, Team B is preferred to Team C, and yet Team C is preferred to Team A. If transitivity were present, then the ranking of teams would be trivial. In the absence of transitivity, what is a good ranking? Recall that a ranking $R = (i_1, \ldots, i_n)$ has an implicit set of preferences whereby Team i_j is preferred to Team i_k whenever $i_j < i_k$. We let L_i^C denote the number of times that the implied preferences based in the ranking R agree with the level L_i data preferences. In this sense, L_i^C is the number of "correct" preferences in R compared to the level L_i preferences determined by the data. We then define

$$C(R) = L_1^C + L_2^C + L_3^C \tag{2}$$

as the number of correct preferences. An optimal ranking R^* is one which maximizes $C(R)$ in (2) over the space of the $n!$ permutations. Although we considered assigning varying weights to the terms in (2), we were unable to determine weights having a theoretical justification.

2.1 Computation

The first computational problem involves the calculation of the correct number of preferences $C(R)$ for a given ranking R. A naive approach in calculating $C(R)$ involves going through all of the L_1, L_2 and L_3 preferences and counting the number that agree with the implied preferences given by R. On an ordinarly laptop computer, such a calculation requires over one hour of computation for a single ranking R in the NCAA basketball dataset. Since our optimization problem involves searching over the space of permutations R, a more efficient way of calculating $C(R)$ is required.

To calculate $C(R)$ for a given ranking R, we pre-process the data by creating three matrices corresponding to preferences at levels L_1, L_2 and L_3. In the $n \times n$ matrix $D^{(k)} = (\bar{d}_{ij}^{(k)})$, k=1,2,3, we have the average differential $\bar{d}_{ij}^{(k)}$ from the point of view

of Team i versus Team j based on a level L_k paired comparison. Once these three matrices are constructed, it is easy to calculate $C(R) = C((i_1, \ldots, i_n))$ in (2) via $L_k^C = \sum_{j=1}^{n-1} \sum_{l=j+1}^{n} (I(\bar{d}_{i_j i_l}^{(k)} > 0) + 0.5 * I(\bar{d}_{i_j i_l}^{(k)} = 0))$ where I is the indicator function and the second term takes ties into account. With the pre-processing, the calculation of $C(R)$ for a new R now takes roughly one second of computation.

Recall that there are $n! \approx 10^{742}$ rankings R in the NCAA dataset, and therefore calculation of $C(R)$ for all rankings is impossible. To maximize $C(R)$ with respect to R over the space of the $n!$ rankings, we implemented a version of the simulated annealing algorithm (Kirkpatrick, Gelatt and Vecchi 1983). Simulated annealing is a stochastic search algorithm that explores the vast combinatorial space, spending more time in regions corresponding to promising rankings. In this problem, we begin with an initial ranking R_0. In the i-th step of the algorithm, a candidate ranking R_{new} is generated in a neighbourhood of the ranking R_{i-1} from step $i-1$. If $C(R_{\text{new}}) > C(R_{i-1})$, then the ranking $R_i = R_{\text{new}}$ is accepted as the current state. In the case where $C(R_{\text{new}}) \leq C(R_{i-1})$, then $R_i = R_{\text{new}}$ if a randomly generated uniform(0,1) variate $u < \exp\{(C(R_{\text{new}}) - C(R_{i-1}))/t_i\}$ where $t_i > 0$ is a parameter often referred to as the temperature. Otherwise the current ranking $R_i = R_{i-1}$ is set at the previous ranking. The algorithm iterates according to a sequence of non-increasing temperatures $t_i \to 0$. The states (rankings) $R_0, R_1, \ldots$ form a Markov chain. The algorithm terminates after a fixed number of iterations or when state changes occur infrequently. Under a 'suitable' neighbourhood structure, asymptotic results suggest that the final state will be nearly optimal.

Success of the simulated annealing algorithm depends greatly on fine tuning of the algorithm. In particular, the user must specify the cooling schedule (i.e. the temperatures t_i) and also the neighbourhood structure for generating successive states from a given state. Aarts and Korst (1989) discuss fine tuning of the algorithm.

Our implementation of simulated annealing begins with the recognition that our problem shares similarities with the well-studied travelling salesman problem. For example, like our problem, the state space in the travelling salesman problem consists of permutations, permutations of cities that are visited by the salesman. Also, in the same way that an interchange in the order of two adjacent cities in a permutation should not greatly affect the total travelling distance for the salesman, an interchange in the order of two

adjacent teams in a permutation (ranking) should not greatly affect the expected number of correct preferences $C(R)$. Accordingly, our implementation of simulated annealing uses an exponential cooling schedule in early stages defined by a sequence of temperature plateaux; this approach has been successively used in the travelling salesman problem (Aarts and Korst 1989).

After extensive experimentation, we have tuned our algorithm and we propose an optimization schedule that is suited to the NCAA basketball seasons under consideration. Specifically, we consider $m = 1, \ldots, 10$ blocks (procedures) where the first eight blocks correspond to simulated annealing. In simulated annealing, the Markov chain consists of B_m iterations in the m-th block with temperature t_m. The temperatures decrease exponentially from one block to the next according to $t_m = 20(0.82)^{m-1}$ where $m = 1, \ldots, 5$. Therefore, it is more difficult to accept downward moves (i.e. when $C(R_{\text{new}}) < C(R_{i-1})$) in the final blocks. In the first $m = 5$ blocks of simulated annealing, we refer to generation of candidate rankings as the "Permutation" procedure. Specifically, within the m-th block, consider the previous state $R_{i-1} = (i_1, \ldots, i_n)$ and generate a discrete uniform variable l on $(1, \ldots, n - k_m + 1)$ where the parameter k_m is user-specified. We then randomly permute $(i_l, i_{l+1}, \ldots, i_{l+k_m-1})$ yielding $(j_l, j_{l+1}, \ldots, j_{l+k_m-1})$. The candidate state in the algorithm is then given by $R_{\text{new}} = (i_1, \ldots, i_{l-1}, j_l, \ldots, j_{l+k_m-1}, i_{l+k_m}, \ldots, i_n)$. In the application, k_m is the number of consecutive teams in the previous ranking that are permuted. Once permuted, a candidate ranking is obtained. In keeping with the heuristic that state changes should be "smaller" as simulated annealing proceeds, we propose a schedule where the tuning parameter k_m decreases as m increases.

When the first five blocks of the algorithm have completed, we carry out a procedure referred to as "Shuffle" in blocks $m = 6, 7, 8, 9$. The idea behind Shuffle is that whereas the Permutation procedure can lead to candidate rankings that differ considerably from the current ranking, Shuffle produces new rankings where only one "misplaced" team shuffles from its current position. Specifically, given the previous ranking R_{i-1}, Shuffle proceeds by generating a discrete uniform random variable l on $(1, \ldots, n)$. Then another discrete uniform random variable j is generated on $(\max(1, l - 50), \ldots, \min(l + 50, n))$. Shuffle updates from R_{i-1} to R_{new} if R_{new} is accepted and where R_{new} has the same ordering as R_{i-1} except that the team ranked l is moved to position j. In blocks $m = 6, 7, 8$, Shuffle is

more accurately described as Non-Greedy Shuffle (NGShuffle) where temperatures t_6, t_7, t_8 are specified as part of the simulated annealing procedure. In block $m = 9$, the Shuffle procedure is modified as Greedy Shuffle (GShuffle). A greedy procedure is one where only non-negative moves towards the maximum are allowed (i.e. $C(R_i) \geq C(R_{i-1})$). The motivation is that when the algorithm nears termination, we only want to be moving in directions which provide improvements.

Finally, in block $m = 10$ of the algorithm, we carry out another greedy procedure which we refer to as "Housekeeping". Housekeeping investigates the effect of even smaller changes to the ranking R following the GShuffle procedure (i.e. block $m = 9$). Specifically, we take $R = (i_1, i_2, \ldots, i_n)$ and we sweep through the solution by inspecting quintuples $(i_j, i_{j+1}, i_{j+2}, i_{j+3}, i_{j+4})$ beginning with $j = 1$ and ending with $j = n - 4$. For each quintuple, we calculate $C(R)$ for the 120 permutations of the quintuple to see if any of the potential rankings lead to an improved solution. Whenever an improved permutation is detected, the ranking is updated accordingly.

Table 1 summarizes the schedule for the optimization algorithm. In the NCAA basketball example, one run of the optimization procedure takes approximately 36 hours of computation. This is not onerous for a task that might be expected to be carried out once per week.

m	Procedure	B_m	k_m	t_m	m	Procedure	B_m	k_m	t_m
1	Permutation	2000	65	20.00	6	NGShuffle	25000		3.00
2	Permutation	3000	60	16.40	7	NGShuffle	25000		2.00
3	Permutation	4000	55	13.45	8	NGShuffle	25000		1.00
4	Permutation	5000	45	11.03	9	GShuffle	75000		
5	Permutation	6000	40	9.04	10	Housekeeping			

Table 1: Schedule for the optimization algorithm. For the m-th block in the Permutation procedure we provide the block size B_m and the number of consecutive teams k_m in the permutation. For the non-greedy procedures, we also provide the temperature t_m.

Figure 1 provides a plot of the optimization algorithm corresponding to the preferences obtained in the 2015/2016 NCAA basketball season. We see that the algorithm moves quickly towards an optimal ranking and then slowly improves. The algorithm was initiated from a promising ranking R_0 (the 2015/2016 season ending RPI rankings). However, the

algorithm works equally well using less promising initial states.

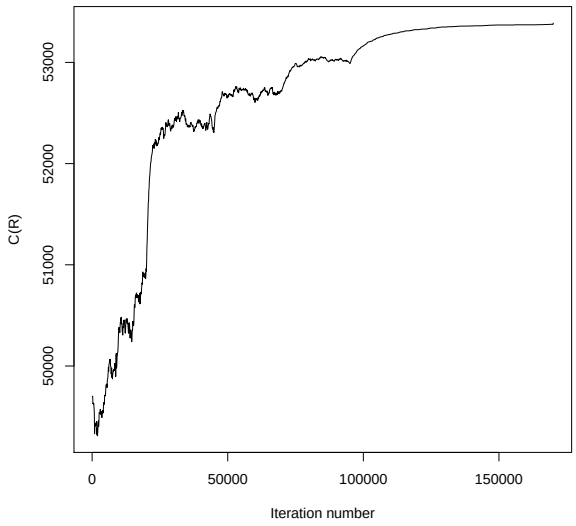

Figure 1: A plot of the number of correct preferences $C(R_i)$ versus the iteration number i in the optimization algorithm.

The simulated annealing algorithm provides guarantees of convergence to a global maximum. However, in practical computing times, it may be the case that our proposed algorithm gets stuck in a local mode and only gets "close" to a maximum. Because of this, we propose multiple runs of the algorithm. In our case, we choose to run the algorithm $M = 20$ times which does not take any extra time because we are able to submit our job to a cluster colony of processors. We then choose the ranking R^* which corresponds to the maximum value of $C(R)$ from the M runs.

The multiple runs also provide us with some confidence that our resultant ranking R^* yields $C^* = C(R^*)$ which is close to the global maximum. From the $M = 20$ runs, we have observed that the resultant maxima $C_1, \ldots, C_M$ are roughly symmetric. To gain some insight, we therefore make the assumption that the maxima are approximately normally distributed with mean $\bar{C}$ and standard deviation given by the sample standard deviation s_C. In extreme value theory, the probability density function of C^*, the M-th

order statistic of $C_1, \ldots, C_M$ is therefore approximately

$$f(C^*) = M \ \frac{1}{s_C}\phi\left(\frac{C^* - \bar{C}}{s_C}\right) \ \Phi\left(\frac{C^* - \bar{C}}{s_C}\right)^{M-1} \tag{3}$$

where ϕ and Φ are the density function and the distribution function of the standard normal distribution, respectively. In Figure 2, we plot the density function of C^* given by (3) based on the observed maxima $C_1, \ldots, C_M$ from the $M = 20$ runs. Based on the observed value $C^* = 53388.5$, the plot suggests that we can be confident that we are close to the global maximum. In particular, it looks unlikely that $C^* = 53388.5$ could be off from the global maximum by much more than 6.0.

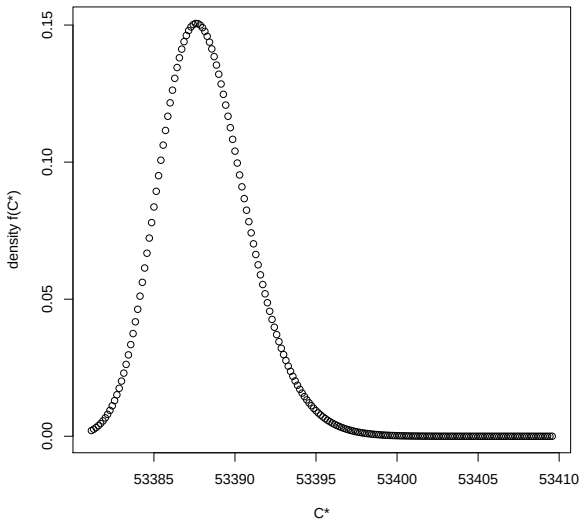

Figure 2: A plot of the density function (3) of the largest order statistic C^* based on the optimal $C(R)$ values $C_1, \ldots, C_M$ from the $M = 20$ runs of the optimization algorithm using the 2015/2016 NCAA basketball dataset.

3 FORECASTING

3.1 NCAA Basketball Data

We now compare our proposed ranking procedure with four widely reported ranking systems used in NCAA basketball (Bihl, Massey, Pomeroy and RPI).

We consider rankings that have been published over five seasons (2011/2012 through 2015/2016) where we note that the Pomeroy rankings were unavailable in the 2012/2013 and 2013/2014 seasons. For each season, we consider 7 time points $t_1, \ldots, t_7$ where rankings are reported. The time periods roughly correspond to mid December, early January, mid January, early February, mid February, early March and mid March. For each ranking system and for each time period (t_i, t_{i+1}), our evaluation considers matches played in the time period and the ranking based at time t_i. Except for the last time period coinciding with March Madness, there are approximately 500 matches played in each time period per year. In a given match, the outcome is categorized as correct if the home team wins by more than $h = 3.4$ points and the home team has the higher ranking. The match outcome is also categorized as correct if the road team wins or loses by less than $h = 3.4$ points and the road team has the higher ranking. On a neutral court, the match outcome is considered correct if the higher ranked team wins.

Over all the predictions made during the five year period, we calculated the percentage of correct predictions by each of the ranking systems. In order of the highest percentages, we observed Pomeroy (69.6%), Massey (69.4%), our proposed method (69.2%), Bihl (68.7%) and RPI (67.8%). Although the percentages are reasonably close, the five methods exhibited fairly consistent orderings on a yearly basis. It is interesting that the RPI approach exhibited the lowest percentage, yet RPI is used by the NCAA Selection Committee in their March Madness deliberations.

Table 2 provides the top five ranked teams using the five ranking methods at the end of the 2015/2016 NCAA basketball season. We observe a lot of agreement in the sets of rankings. However, our proposed approach is interesting in that it provides a notable difference from the other rankings. In particular, Kansas is excluded from the top five whereas Xavier is included. This perspective is interesting as Kansas had a good season (33 wins versus 5 losses). However, their five losses came against strong teams (Michigan

St, West Virginia, Oklahoma State, Iowa State and Villanova), all top 20 AP (Associated Press) teams except for Oklahoma State. This highlights the importance of the head-to-head matchups which is discussed in Section 4.1. We note that our ranking had Kansas in the seventh position. On the other hand, Xavier (not a traditional powerhouse school) had a strong 28-6 record and may have been overlooked by some of the other ranking methods.

Method	First	Second	Third	Fourth	Fifth
Pomeroy	Villanova	N Carolina	Virginia	Kansas	Michigan St
Massey	Villanova	Kansas	N Carolina	Virginia	Oklahoma
Proposed	Villanova	Michigan St	N Carolina	Xavier	Virginia
Bihl	Kansas	Villanova	N Carolina	Oklahoma	Virginia
RPI	Kansas	Villanova	Virginia	Oregon	N Carolina

Table 2: Final rankings of the top five teams at the end of the 2015/2016 season.

3.2 Simulated NCAA Basketball Data

Although the previous example using actual NCAA basketball data was instructive, it did not allow us to make comparisons with the "truth" since the correct rankings based on team strengths in actual seasons are always unknown. In this example, we consider simulated data sets where we can initially set team strengths so that the true rankings are known to us.

Therefore, in the context of NCAA basketball, we consider $n = 351$ teams where the team rankings are set according to alphabetical order. For example, Team 1 is the best team and its schedule is determined by the 2015/2016 schedule for Abilene Christian. Team 351 is the weakest team and its schedule is determined by the 2015/2016 schedule for Youngstown State. For a match between Team i and Team j on a neutral court, the observed point differential in favour of Team i is modeled according to the Normal($\mu_i - \mu_j, \sigma^2$) distribution where the normal distribution is a common assumption for NCAA basketball (Stern and Mock 1998), and we set $\sigma = 9.3$ which is consistent with Swartz et al. (2011). If the normal variate is greater (less) than zero, then Team i is the winning (losing) team. For home and road matches, winners and losers are determined by using the same procedure as if the match was played on a neutral court.

We consider two team strength scenarios. In the first case, we set team strengths according to $\mu_i = 35.1 - (0.1)i$ such that Team 1 has strength 35.0 and Team 351 has strength 0.0. This implies, for example, that the strongest team is expected to defeat the weakest team by 35 points on a neutral court. In the second case, we set team strengths according to $\mu_i = 52.65 - (0.15)i$ which implies that the strongest team is expected to defeat the weakest team by 52.5 points on a neutral court.

Our comparison via simulation proceeds by generating $M = 10$ seasons of matches according to the above description where $h = 3.5$ is set as the home team advantage. In the j-th season, we take the resultant ranking $R_j = (j_1, \ldots, j_n)$ and compare it to the true ranking $R_{\text{true}} = (1, \ldots, n)$. We do this using two comparison metrics, $C_j^{(1)} = \frac{1}{n}\sum_{i=1}^{n} \mid j_i - i \mid$ and $C_j^{(2)} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(j_i - i)^2}$. We repeat the procedure over the $M = 10$ seasons to obtain the overall comparison metrics $C^{(1)} = \frac{1}{M}\sum_{j=1}^{M} C_j^{(1)}$ and $C^{(2)} = \frac{1}{M}\sum_{j=1}^{M} C_j^{(2)}$.

Our simulation involves a comparison of our proposed ranking method with the Bradley-Terry approach which is considered the benchmark procedure for paired comparison data. Bradley-Terry estimation procedure fails when there is more than one winless team. For this reason, we assign 0.5 wins in those rare cases where there are winless teams. We note that Bayesian implementations of Bradley-Terry as mentioned in the Introduction do not suffer from this drawback. We are unable to make comparisons with some of the systems that are frequently reported in NCAA basketball (e.g. Sagarin, Pomeroy, Massey or RPI) since the systems are proprietary and the code is unavailable.

Table 3 reports the results of the simulation procedure. The metrics are interesting as they may be interpreted as the average deviation between a team's ranking and its true ranking. We observe that both ranking procedures are improved in the second simulation case. This makes sense as there is more variability between teams in Case 2 than in Case 1, and it is therefore more likely for a ranking method to differentiate between teams. Further, we observe that in both simulation cases and using both metrics that the proposed ranking procedure gives better rankings than Bradley-Terry. The reported standard errors suggest that the improvements are statistically significant in the second simulation case. We believe that Case 2 is more realistic than Case 1 in describing a wider range in quality between NCAA teams.

It would be interesting to repeat the simulation where team strengths were not lin-

ear but followed a Gaussian specification. For example, one could generate μ_i from a Normal$(0, 36)$ distribution and then sort the μ_i such that μ_1 is the largest and μ_{351} is the smallest. This may be a more realistic description of a population of team strengths.

	Proposed Ranking Procedure	Bradley-Terry Procedure
Case 1: $\mu_i = 35.1 - (0.1)i$	$C^{(1)} = 12.7$ (0.39)	$C^{(1)} = 13.2$ (0.85)
	$C^{(2)} = 15.4$ (0.44)	$C^{(2)} = 16.9$ (0.99)
Case 2: $\mu_i = 52.65 - (0.15)i$	$C^{(1)} = 09.1$ (0.25)	$C^{(1)} = 11.7$ (0.44)
	$C^{(2)} = 11.0$ (0.34)	$C^{(2)} = 15.0$ (0.47)

Table 3: Comparison metrics with standard errors in parentheses for two ranking systems (Proposed versus Bradley-Terry) studied under two simulation cases.

3.3 English Premier League Data

Whereas NCAA basketball consisted of $n = 351$ teams in 2015/2016, the English Premier League (EPL) is a much smaller league with $n = 20$ teams. Therefore, the EPL provides a different type of challenge for our ranking procedure.

In the EPL, each team plays both a home and a road game against every other team for a total of 38 matches in a season. We begin by setting two dates during the 2016/2017 EPL season where ranks based on our procedure are determined at each date. These dates roughly correspond to weeks 19 and 27 of the season. We chose not to extend the dates to the latter part of the season as unusual playing behaviours sometimes occur. For example, in the latter portion of the 2016/2017 season, Manchester United was more focused on their Europa Cup matches than in their EPL matches. It was believed that they had a greater chance for Champions League qualification from the Europa Cup route. Also, we wanted to use some of the matches beyond week 27 to assess the predictive power of the rankings. Based on these dates, each team had played every other team at least once, and therefore team comparisons were based entirely on level L_1 preferences. The data matrix $D^{(1)}$ was constructed for each of the two dates where the home field advantage was set

at $h = 0.5$ (Roeder and Curley 2014). Recall that the calculation of paired comparison preferences is insensitive to the choice of h in the wide interval $h \in (0, 1)$.

We note that the full strength of the optimization algorithm described in Table 1 was not required since we have fewer teams. We instead initiated the procedure beginning in block $m = 6$. We also modified the Shuffle procedure where we now generate independent uniform variates l and j on $(1, \ldots, 20)$. Under Shuffle, the candidate ranking R_{new} has the same ordering as R_{i-1} except that team l is inserted into position j. In this case, the optimization procedure was carried out in roughly 45 seconds of computing for each of the two time periods.

An advantage of working with a smaller league is the increased confidence that optimal rankings are obtained. Multiple runs of the algorithm based on different initial rankings typically gave the same value of $C(R)$. However, we did discover that the rankings were not unique. We found three optimal rankings at the first date and two optimal rankings at the second date.

Table 4 provides both the EPL table (standings) and the optimal rankings at the two dates. We observe some meaningful differences between the tables and the ranks. On the Jan 1/17 date, the largest discrepancies between the table and the optimal ranks involve Middlesbrough, Arsenal and Watford. The optimal rankings suggest that Middlesbrough is stronger (9 placings), Arsenal is weaker (6 placings) and Watford is weaker (6 placings) than the table indicates. Middlesbrough's strength was aided by "wins" (i.e. taking into account home team advantage) over Manchester City, Arsenal and West Brom. We also observe that the three optimal rankings R_1^*, R_2^* and R_3^* on Jan 1/17 are similar; the only differences involve the top three sides Chelsea, Liverpool and Manchester United. On the Mar 6/17 date, the largest discrepancies between the table and the optimal rankings involve Manchester City (7 places lower according to R_1^*), Leicester City (6 places higher according to R_1^*) and Sunderland (6 places higher according to both R_1^* and R_2^*).

Having observed some of the large discrepancies between the standings and the optimal rankings in Table 4, it is difficult to assess which lists are more sensible as measures of team strength. Perhaps large discrepancies indicate to gamblers that there is something interesting about such teams, that there may be a partial explanation for their standings at a given point in time. We now use the rankings at the two dates in Table 4 to

	Jan 1/17						Mar 6/17					Final
Pos	Table	Gms	Pts	R_1^*	R_2^*	R_3^*	Table	Gms	Pts	R_1^*	R_2^*	Table
1	CHE	19	49	CHE	MUN	LIV	CHE	27	66	LIV	LIV	CHE
2	LIV	19	43	MUN	LIV	CHE	TOT	27	56	TOT	TOT	TOT
3	ARS	19	40	LIV	CHE	MUN	MCI	26	55	ARS	CHE	MCI
4	TOT	19	39	TOT	TOT	TOT	LIV	27	52	CHE	EVE	LIV
5	MCI	19	39	SOU	SOU	SOU	ARS	26	50	MUN	MCI	ARS
6	MUN	19	36	EVE	EVE	EVE	MUN	26	49	EVE	ARS	MUN
7	EVE	19	27	MID	MID	MID	EVE	27	44	WBA	MUN	EVE
8	WBA	19	26	MCI	MCI	MCI	WBA	27	40	SOU	WBA	SOU
9	SOU	19	24	ARS	ARS	ARS	STK	27	35	LEI	SOU	BOU
10	BOU	19	24	BOU	BOU	BOU	SOU	26	33	MCI	LEI	WBA
11	BUR	19	23	WBA	WBA	WBA	WHU	27	33	STK	STK	WHU
12	WHU	19	22	LEI	LEI	LEI	BUR	27	31	WHU	WHU	LEI
13	WAT	19	22	STK	STK	STK	WAT	27	31	BUR	BUR	STK
14	STK	19	21	WHU	WHU	WHU	BOU	27	27	SUN	SUN	CRY
15	LEI	19	20	SWA	SWA	SWA	LEI	27	27	CRY	CRY	SWA
16	MID	19	18	BUR	BUR	BUR	SWA	27	27	WAT	WAT	BUR
17	CRY	19	16	CRY	CRY	CRY	CRY	27	25	MID	MID	WAT
18	SUN	19	14	SUN	SUN	SUN	MID	27	22	BOU	BOU	HUL
19	HUL	19	13	WAT	WAT	WAT	HUL	27	21	HUL	HUL	MID
20	SWA	19	12	HUL	HUL	HUL	SUN	27	19	SWA	SWA	SUN

Table 4: EPL table (standings) and optimal rankings R_i^* corresponding to our method. The chosen intervals in the 2016/2017 season include matches up to and including the specified dates. We also include the final season ending table.

forecast matches. For example, based on the time interval from Jan 2/17 to Mar 6/17, we investigated the 78 matches played. We calculated the percentage of correct forecasts as implied by the table and by our rankings as of Jan 1/17. We repeated the procedure for the time interval Mar 7/17 to May 1/17. The results are provided in Table 5. It is difficult to conclude much from Table 5. In the first predictive period, the table appears to do slightly better than the optimal rankings. In the second period, the opposite pattern emerges. It seems that both the table and the optimal rankings use past results in a sensible way to assess team strength. Although the EPL exercise was interesting, we believe the methods developed in this paper are particularly suited to the more challenging problem involving

large numbers of teams such as the NCAA where most pairs of teams do not compete.

		Forecast Accuracy			
Time Interval	Matches	Table	R_1^*	R_2^*	R_3^*
Jan 2/17 - Mar 6/17	78	70.5%	66.7%	65.4%	66.7%
Mar 7/17 - May 1/17	76	61.8%	63.2%	65.8%	

Table 5: Percentage accuracy of forecasts implied by the table and the optimal rankings during two time intervals.

4 NUANCES OF THE APPROACH

4.1 Importance of Individual Matchups

Although the proposed ranking scheme is conceptually simple and does not rely on parametric assumptions, it is sometimes instructive to look at pathological cases to gain a deeper understanding of the approach.

We therefore consider the case where the number of teams n is large, say $n > 100$. Suppose further that there are two very strong teams, Team A and Team B, that are preferred in 99% and 90% of the paired comparisons resulting from matches, respectively. And suppose that the remaining $n - 2$ teams are not nearly as strong as Team A and Team B. Then, following (2), it is apparent that

$$C((B, A, i_3, \ldots, i_n)) = C((A, B, i_3, \ldots, i_n)) + 1 \tag{4}$$

if Team B has "defeated" Team A (i.e. if Team B is preferred to Team A in terms of the actual matches).

The question is whether it is sensible to rank Team B above Team A according to (4) given that Team A has won a much larger proportion of matches than Team B (99% versus 90%). The answer to the question depends on how one views the importance of individual matchups. We believe that in NCAA basketball and football, head-to-head matchups are considered vitally important.

The above discussion illuminates the importance of individual matchups in the proposed ranking system. We now contrast this with the Bradley-Terry ranking system. In

Bradley-Terry, the i-th team is characterized by a parameter π_i such that the probability that Team i defeats Team j is $\pi_i/(\pi_i + \pi_j)$. Therefore, the π_i's (by virtue of their magnitude) determine a ranking of the teams.

In Bradley-Terry, consider the case where two teams i_1 and i_2 have the same strength (i.e. $\pi_{i_1} = \pi_{i_2}$). Assume further that each of the n teams has played all of the other $n-1$ teams exactly once. Then, according to the steady state of the Bradley-Terry iterative estimation procedure (Bradley and Terry 1952), we have

$$\sum_{j \neq i_1} x_{i_1 j} \left(\sum_{j \neq i_1} \frac{1}{\pi_{i_1} + \pi_j} \right)^{-1} = \sum_{j \neq i_2} x_{i_2 j} \left(\sum_{j \neq i_2} \frac{1}{\pi_{i_2} + \pi_j} \right)^{-1} \tag{5}$$

where $x_{ij} = 1$ if Team i defeated Team j and $x_{ij} = 0$ if Team i lost to Team j. From (5), we obtain $\sum_{j \neq i_1} x_{i_1 j} = \sum_{j \neq i_2} x_{i_2 j}$. The implication for Bradley-Terry is that the equal ranking of the two teams i_1 and i_2 is based on their total number of victories $\sum_{j \neq i_1} x_{i_1 j}$ and $\sum_{j \neq i_2} x_{i_2 j}$ which is only slightly dependent on the result of their particular matchup. Buhlmann and Huber (1963) develop properties of ranking procedures including the recognition that the number of wins by each team is the sufficient statistic for the Bradley-Terry model.

Given that our procedure is nonparametric, there is no underlying likelihood and no capability for the calculation of model-based probabilities. For example, there are no parameters to estimate and one cannot assess the closeness of rankings. Therefore, it is interesting to consider how the resultant rankings are affected by match outcome variability. It seems the only way to assess the impact of match variability is to change a particular match outcome (or a set of match outcomes). Then one re-runs the ranking algorithm and simply observes how the rankings change.

4.2 Uniqueness of the Optimal Ranking

Since $C(R)$ is a discrete function, it is important to consider whether a derived optimal ranking R^* is unique. It is clear that a solution may not be unique as demonstrated in the following simple example. Let $n = 3$ and suppose that the data preferences are: A is preferred to B, B is preferred to C, and C is preferred to A. In this case, there are three

optimal rankings since

$$C(A, B, C) = C(C, A, B) = C(B, C, A) = 2 \ .$$

The following proposition addresses an aspect of the uniqueness of optimal solutions.

Proposition 1: Suppose that B_1, B_2 and B_3 are ordered blocks of subjects and that $R_1 = (B_1, B_2, j, B_3)$ is an optimal ranking. Then $R_2 = (B_1, j, B_2, B_3)$ is an optimal ranking only if $C(j, B_2) = C(B_2, j)$.
Proof: $C(R_1) - C(R_2) = C(B_2, j) - C(j, B_2)$. Therefore, the only way that R_2 can be an optimal ranking is if $C(j, B_2) = C(B_2, j)$.

The condition $C(j, B_2) = C(B_2, j)$ in Proposition 1 implies that j is preferred to exactly half of the subjects in B_2. However, there is more that can be said concerning Proposition 1. Suppose that $B_2 = (i_1, i_2, \ldots, i_k)$ where k is necessarily even. Then it must be the case that $C(j, i_1) = 1$. For if it were not the case, then $C(B_1, i_1, j, i_2, \ldots, i_k, B_3) > C(R_2) = C(B_1, j, i_1, i_2, \ldots, i_k, B_3)$ and therefore R_2 would not be optimal. Using similar reasoning, it must also be the case that $C(j, i_k) = 0$. In other words, j must be preferred to the best subject i_1 in B_2 and j must not be preferred to the worst subject i_k in B_2. We suggest that this is an atypical situation and it becomes more unusual as k increases.

"Small" deviations from optimal rankings which lead to alternative optimal rankings are not a big concern. What we would not like to see is "very different" rankings which are optimal. Now although there can be other optimal rankings which do not take the form described in Proposition 1, our investigations suggest that again, optimal rankings tend to be "close". The intuition is that in an optimal ranking R_1, the strong subjects appear early in the ranking going from left to right. Any reshuffling of the subjects moves away from this heuristic and is therefore likely not optimal.

5 DISCUSSION

With paired comparison data, the ranking problem is a common problem which has applications to sport. The ranking method proposed in this paper is conceptually simple although the calculation of an optimal ranking poses computational challenges.

For NCAA basketball, the ranking problem is particulary challenging due to the sparsity of games relative to the number of teams $n = 351$. In an ideal world, our method would lead to a unique global maximum. However, the lack of uniqueness in some datasets is a consequence of the simplicity of the optimality criterion. The good news is that the rankings belonging to the set of optimal rankings tend to be "close" to one another. For getting a sense of the overall quality of teams, we have demonstrated that our ranking system compares favourably with other ranking systems. The only occasions where the uniqueness issue may cause distress is when there is disparity between the top teams. In such situations, one may consider the introduction of tie-breaking procedures to differentiate between optimal rankings.

We note that our ranking criterion $C(R)$ in equation (2) assigned equal weights to team preferences. That is, we assigned the same weight to pairs of teams that competed directly versus pairs of teams that competed indirectly. For future research, we may consider obtaining weights that provide optimal predictive power. The procedure would be computationally demanding and optimal weights may vary by sport. Another avenue for future research might be to assign data preference scores other than 0/1. For example, suppose that Team A had a data preference of q points over Team B. Then, the contribution to $C(R)$ for having Team A ranked above Team B in the ranking R would be worth q points. For example, data preference scores could be assigned via considerations based on margin of victory.

6 REFERENCES

Aarts, E. and Korst, J. (1989). *Simulated Annealing and Boltzmann Machines*, Wiley: New York.

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

Bessire, P. (2016). "College Homecourts (2/16/12)". In *Prediction Machine.com*, http://www.predictionmachine.com/college-basketball-homecourt-advantage.

Bradley, R.A. and Terry, M.E. (1952). "Rank analysis of incomplete block designs. I. The method of paired comparisons", *Biometrika*, 39, 324-345.

Buhlmann, H. and Huber, P.J. (1963). "Pairwise comparison and ranking in tournaments", *Annals of Mathematical Statistics*, 34, 501-510.

Caron, F. and Doucet, A. (2012). "Efficient Bayesian inference for generalized Bradley-Terry models", *Journal of Computational and Graphical Statistics*, 21, 174-196.

Cattelan, M. (2012). "Models for paired comparison data: A review with emphasis on dependent data", *Statistical Science*, 27(3), 412-433.

Chen, C. and Smith, T.M. (1984). "A Bayes-type estimator for the Bradley-Terry model for paired comparison", *Journal of Statistical Planning and Inference*, 10, 9-14.

Davidson, R.R. (1970). "On extending the Bradley-Terry model to accommodate ties in paired comparison experiments", *Journal of the American Statistical Association*, 65, 317-328.

dishingoutdimes (2010). "How the BCS Rankings are calculated?", In *SB Nation*, http://www.crimsonandcreammachine.com/2010/10/7/1737405/how-the-bcs-rankings-are-calculated.

Glickman, M.E. (1999). "Parameter estimation in large dynamic paired comparison experiments", *Applied Statistics*, 48, 377-394.

Glickman, M.E. (2001). "Dynamic paired comparison models with stochastic variances", *Journal of Applied Statistics*, 28, 673-689.

Gandar, J.M., Zuber, R.A. and Lamb, R.P. (2001). "The home field advantage revisited: A search for the bias in other sports betting markets", *Journal of Economics and Business*, 53(4), 439-453.

Harville, D. (1977). "The use of linear-model methodology to rate high school or college football teams", *Journal of the American Statistical Association*, 72, 278-289.

Harville, D. (2003). "The selection or seeding of college basketball or football teams for postseason competition", *Journal of the American Statistical Association*, 98, 17-27.

Kirkpatrick, S., Gelatt Jr. C.D. and Vecchi, M.P. (1983). "Optimization by simulated annealing", *Science*, 220, 671-680.

Leonard, T. (1977). "An alternative Bayesian approach to the Bradley-Terry model for paired comparisons", *Biometrics*, 33, 121-132.

Mease, D. (2003). "A penalized maximum likelihood approach for the ranking of college football teams independent of victory margins", *The American Statistician*, 57, 241-248.

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

Plackett, R. (1975). "The analysis of permutations", *Applied Statistics*, 24, 193-202.

Roeder, O. and Curley, J. (2014). "Home-field advantage doesn't mean what it used to in English football", In *FiveThirtyEight*, https://fivethirtyeight.com/features/home-field-advantage-english-premier-league/

Rotou, O., Qian, X. and von Davier, M. (2015). "Ranking systems used in gaming assessments and/or competitive games", *ETS Research Memorandum Series*, ETS RM-15-03.

Stern, H.S. and Mock, B. (1998). "College basketball upsets: will a 16-seed ever beat a 1-seed?" In the column, "A Statistician Reads the Sports Pages", *Chance*, 11, 26-31.

Swartz, T.B. and Arce, A. (2014). "New insights involving the home team advantage", *International Journal of Sports Science and Coaching*, 9(4), 681-692.

Swartz, T.B., Tennakoon, A., Nathoo, F., Tsao, M. and Sarohia, P.S. (2011). "Ups and downs: team performance in best-of-seven playoff series", *Journal of Quantitative Analysis in Sports*, 7(4), Article 2.

Thurstone, L.L. (1927). "A law of comparative judgment", *Psychological Review*, 34, 368-389.