

# The many routes to the ubiquitous Bradley-Terry model

Ian Hamilton, Nicholas Tawn, David Firth

December 22, 2023

## Abstract

The rating of items based on pairwise comparisons has been a topic of statistical investigation for many decades. Numerous approaches have been proposed. One of the best known is the Bradley-Terry model. This paper seeks to assemble and explain a variety of motivations for its use. Some are based on principles or on maximising an objective function; others are derived from well-known statistical models, or stylised game scenarios. They include both examples well-known in the literature as well as what are believed to be novel presentations.

## 1 Introduction

The first conference that the lead author attended as a PhD student was an American sports statistics conference. He presented a poster related to the Bradley-Terry model. As a retrodictive model on rugby union in a sea of American sports predictions, it felt a little out of place. But a kind attendee took pity on him and decided to engage him with a question. She asked, “Why would I choose Bradley-Terry rather than the Thurstone model?” (by which he took her to mean what is more commonly referred to as the Thurstone-Mosteller model). He flummed a vague response involving analytic niceness and simplicity — he suspects Occam’s razor even got a mention. She looked suitably unconvinced. It is to be hoped that this paper represents a more ordered response to the conference interlocutor and an aggregation of, as David (1988, p.13) puts it in his canonical survey of pairwise comparison methods, “the many routes to the ubiquitous Bradley-Terry model.”

Thus, the main original contribution of the work is in aggregating the motivations for the Bradley-Terry model, or as Bradley (1976) refers to them, the ‘bases for model formulation’. In collating these motivations, we hope that the work provides a useful resource to those encountering the model for the first time and some new perspectives for those more familiar with it. It may also complement other works, such as David (1988), Cattelan (2012), Vojnović (2015), or Wu et al. (2022) that provide alternative helpful summary perspectives on the model. The work takes in a diverse scope of motivating ideas including likelihood and entropy maximisation, psychological choice and sensation models, distance minimisation, a prominent Markov chain Monte Carlo method, other well-known rating models such as PageRank and the RPI of American college sports, sudden-death play-offs, pub pool norms and the British playground game of conkers. The aggregation of these motivations serves to demonstrate the broad appeal of the Bradley-Terry model in many settings.

The paper also offers a number of novelties including: a more extensive explicit discussion of the Bradley-Terry model in the context of an exponential family of distributions than has appeared previously, which provides a uniting theme to a number of the more notable motivations; a formalisation of perhaps the most intuitive motivation for the model, by proposing an explicit measure for the simplicity of a model in the pairwise comparison scenario and showing that, under plausible constraints, Bradley-Terry is the model that maximises this measure; and a demonstration of how the ideas behind the rating method of Wei (1952) and Kendall (1955) and of the Ratings Percentage Index (RPI) can be related to the Bradley-Terry model through Perron-Frobenius Theorem.

The scenario under consideration in this paper is one where there is a desire to create a ranking of items based on the observation of a set of binary-outcome pairwise comparisons. One popular approach to ranking is to determine a uni-dimensional rating, and then order items by their ratings. Statistical models such as Bradley-Terry or Thurstone-Mosteller achieve this by defining the probability of a preference for alternative  $i$  over alternative  $j$  in a pairwise comparison independently from other preferences conditional on the strengths of the items. In the Bradley-Terry model the probability is defined as

$$p_{ij} = \frac{\pi_i}{\pi_i + \pi_j},$$

where  $\pi_i$  is a positive-valued parameter that may be interpreted as a rating of alternative  $i$ , with a higher rating indicating greater ‘strength’ or ‘worth’.

This results in a model that generates independent Binomial realisations between pairs of items. Therefore, with a logit transformation of the above, one can equivalently state the model as a member of the class of generalised linear models

(McCullagh and Nelder, 1989) with

$$F(p_{ij}) = \lambda_i - \lambda_j,$$

where  $\lambda_i = \log(\pi_i)$  is a real-valued parameter indicating the strength of  $i$ , and  $F$  is taken as the logit function. The Thurstone-Mosteller model (Thurstone, 1927a; Mosteller, 1951), about which the interlocutor asked, is derived from taking  $F$  to be the probit function instead. In practice, as Stern (1992) notes, the models are often empirically very similar.

The Bradley-Terry model has formed the basis for many models and analyses in many contexts over time including, for example, those for journal citations (Stigler, 1994), college sports (Wobus, 2007), animal behavior (Stuart-Fox et al., 2006), risk analysis (Merrick et al., 2002), wine tasting (Oberfeld et al., 2009), university ranking (Dittrich et al., 1998), font selection (O’Donovan et al., 2014), educational assessment (Pollitt, 2012b), locational deprivation (Seymour et al., 2022) and of course in chess, which was the subject of the original work by Zermelo (1928), as well as being the subject of the popular closely-related ranking method proposed by Elo (1978), which is widely known and is still in use in the sport today.

Originally documented by Zermelo (1928), the Bradley-Terry model took the name by which it came to be commonly known when Bradley and Terry (1952) independently rediscovered it. Following the work of Thurstone (1927a,b,c) and Zermelo (1928), paired comparison methods saw little development for the best part of a quarter of a century until they became an active area of investigation in the 1950s and 60s. Much of this work took place in the context of the psychological literature, with Luce’s Choice Axiom (Luce, 1959) a particularly notable contribution, leading to the model sometimes being referred to as the Bradley-Terry-Luce (BTL) model. A number of these works showed how the Bradley-Terry model could be derived based on plausible axioms or desirable model features (Good, 1955; Luce, 1959; Bühlmann and Huber, 1963; Luce and Suppes, 1965). Towards the end of this period, Thompson and Singh (1967) demonstrated that a consideration of extreme value distributions within a discriminial process leads to the Bradley-Terry model, and Daniels (1969), in a highly original paper, noted the links between the Bradley-Terry model and what might now be recognised as an undamped PageRank (Page et al., 1999).

For further details of the development of the model up to this point David (1988) provides a thorough account of the paired comparison literature more generally, Bradley (1976) and Davidson and Farquhar (1976) give interesting perspectives on the literature related to the Bradley-Terry model at the end of this period, and Glickman (2013) is a highly readable account of the history, particularly as it pertains

to the contribution of Zermelo.

The next significant contributions to motivating the Bradley-Terry model came from Henery (1986) and Joe (1988) in identifying the model as the result of maximising an objective function subject to a suitable constraint. The later work (Joe, 1988) seems to have been unaware of Henery (1986), but provides a more complete presentation. As well as considering the Bradley-Terry model as a maximum entropy model and noting its relationship to an appropriate sufficient statistic, Joe (1988) also explicitly notes the link to a maximum likelihood derivation. A number of motivations in this paper are based on game-style scenarios. Perhaps the most interesting paper related to this also comes from this period (Stern, 1990). In the context of the purpose of this work, McCullagh (1993) provides a particularly pertinent contribution at the end of this period, demonstrating how the Bradley-Terry model can be motivated from a geometric perspective, as well as how, under certain conditions, it is essentially equivalent to two other well-known models for permutations and from directional statistics respectively.

More recently Slutzki and Volij (2006), Negahban et al. (2012), Maystre and Grossglauser (2015) and Selby (2020) provide more detailed accounts of the link between the Bradley-Terry model and the limiting distribution of a Markov Chain, and thereby to an undamped PageRank. The Social Choice literature provides an interesting perspective on this relationship, building on the approach of Rubinstein (1980) to provide axiomatic justifications for ranking methods. Slutzki and Volij (2006) is perhaps the most notable example in the present context.

The paper proceeds by dividing the motivations up into six types: axiomatic; objective function maximisation; discriminial processes; standard models; game scenarios; and quasi-symmetry and consistent estimators. These categorisations are somewhat arbitrary, and linkages exist across them which will be highlighted, but for the present purpose they provide a useful means to order the work. It begins with Section 2, the discussion of axiomatic approaches, which takes as a starting point features that one might reasonably desire of a pairwise comparison model. A number are very closely linked and might even be thought of as restatements of the same idea, but the intuitions behind them differ sufficiently, as evidenced by their separate appearances in the literature, such that they are presented separately here.

In Section 3, the selection of a rating model is cast in the familiar framework of a constrained optimisation, where an objective function is maximised or minimised subject to some plausible constraint. Section 4 takes the context of Thurstone’s discriminial processes, and discusses the distributions that lead to a Bradley-Terry model under this set-up, and how they might be motivated. In Section 5, it is noted how the Bradley-Terry model is apparent in other well-known statistical models,

as a conditional form of Rasch, Mallows  $\phi$ , von Mises-Fisher, hazard and network models. In Section 6, some examples are introduced that derive from realistic game scenarios picking up on the highly intuitive nature of the model. In Section 7, the quasi-symmetry model is discussed, and is used to show how the often intuitive approaches that underlie a number of other popular rating methods can be related to Bradley-Terry and produce consistent estimators for the Bradley-Terry strength parameters. This also leads to noting the link to Barker’s algorithm, a popular Markov chain Monte Carlo method.

In each subsection, the reference given in the title is that of the earliest work linking the approach explicitly to the Bradley-Terry model, and the subsections are ordered chronologically by these. The sections are ordered with statistical interest and chronology in mind.

In Section 8, the natural questions of how these motivations are linked and the usefulness of motivating the model from diverse perspectives is addressed. The linkages are established with an examination of the Bradley-Terry model in the context of an exponential family of distributions. In demonstrating the usefulness of the approach, two illustrative examples are provided where it may be natural to use the model based on one motivation, but its application can be aided by considering it through another motivation.

Throughout the paper,  $p_{ij}$  will be the probability of  $i$  beating  $j$  or for a preference for  $i$  over  $j$  given a comparison between  $i$  and  $j$ , where  $i, j \in T$  and  $T$  is of size  $n$ . The  $n \times n$  data matrix  $C = [c_{ij}]$  will be the ‘competition’ matrix of preferences or wins, such that  $c_{ij}$  is the number of times  $i$  was preferred over  $j$ .  $M = C + C^T$  is defined as the symmetric matrix where  $m_{ij}$  is the number of comparisons, or ‘matches’ in British sports parlance, between  $i$  and  $j$ . For the avoidance of doubt, no item is compared with itself, so that  $c_{ii} = m_{ii} = 0$  for all  $i$ . The observed wins for a team  $i$  is denoted by  $w_i = \sum_j c_{ij}$ .  $C$  is taken to be irreducible, that is, as described by Ford Jr (1957, p.29): “[I]n every possible partition of the objects into two non-empty subsets, some object in the second set has been preferred at least once to some object in the first set.” This ensures that strength estimates are finite. It is not assumed that there are the same number of comparisons between any two items, nor indeed that the number of comparisons between any two items is non-zero. Shortened summation notation is used such that  $\sum_{i,j}$  is taken to be  $\sum_{i=1}^n \sum_{j=1}^n$  and  $\sum_{i < j}$  is taken to be  $\sum_{j=1}^n \sum_{i=1}^{j-1}$ . Where appropriate, the language of sports — contests, scores, teams, wins — is used to aid in providing clear interpretability, though the motivations may be analogised outside this context.

## 2 Axiomatic motivations

It is sometimes possible to fix properties that we would desire of a model and use them to derive a unique model. In this section, we consider such properties that lead to the Bradley-Terry model.

### 2.1 Transitivity of odds (Good, 1955)

Consider four teams  $i, j, k, l$ . Suppose that the probability that  $j$  beats  $k$  is greater than the probability that  $j$  beats  $l$ ,

$$p_{jk} > p_{jl},$$

then it is intuitive to think that the probability that  $i$  beats  $k$  will be greater than the probability that  $i$  beats  $l$ ,

$$p_{ik} > p_{il}.$$

Perhaps the simplest way to enforce this is by insisting on the transitivity of odds as Good (1955) proposes, that is

$$\frac{p_{ij}}{p_{ji}} \times \frac{p_{jk}}{p_{kj}} = \frac{p_{ik}}{p_{ki}}.$$

Alternatively one might think of the same condition in the manner that Luce and Suppes (1965) refers to it as the *product rule*, where for any triple  $(i, j, k)$  the probability of the intransitive cycle  $i$  beats  $j$ ,  $j$  beats  $k$ ,  $k$  beats  $i$  is the same as that of the intransitive cycle  $i$  beats  $k$ ,  $k$  beats  $j$ ,  $j$  beats  $i$ , expressed

$$p_{ij}p_{jk}p_{ki} = p_{ik}p_{kj}p_{ji} \quad \text{for all triplets } (i, j, k).$$

Strang et al. (2022) characterise this as an ‘arbitrage free’ condition and it is also known as Kolmogorov’s criterion (Kolmogorov, 1936; Kelly, 1979).

Jech (1983) provides an alternative justification for the principle by considering estimating the odds of an item  $i$  beating an item  $k$  in the scenario where the comparison can only be made indirectly by comparing  $i$  to  $j$  and  $j$  to  $k$ . If  $i$  beats  $j$  and  $j$  beats  $k$  then  $i$  is taken to have beaten  $k$ . If  $i$  loses to  $j$  and  $j$  loses to  $k$  then  $k$  is taken to have beaten  $i$ . For other result combinations ( $i$  beats  $j$  and  $k$  beats  $j$ , or  $j$  beats  $i$  and  $j$  beats  $k$ ) judgement is reserved. In any given comparison, the probability that  $i$  beats  $k$  is thus  $p_{ik} = p_{ij}p_{jk}$  and the probability that  $k$  beats  $i$  is thus  $p_{ki} = p_{ji}p_{kj}$ . Taking the ratio of these probabilities, the odds conform to the transitivity condition. Jech (1983, p.246) claims that this leads to the “one and only

one correct way of comparing the records of teams in an incomplete tournament”, which seems a little bold, but the argument nevertheless demonstrates the intuitive appeal of the property.

Returning to how this criterion leads to the Bradley-Terry model, and following Good (1955), it may alternatively be expressed as

$$\log \frac{p_{ij}}{p_{ji}} + \log \frac{p_{jk}}{p_{kj}} = \log \frac{p_{ik}}{p_{ki}}.$$

Letting  $p_{ij}/p_{ji} = \exp(\tau(\theta_i, \theta_j))$ , where  $\theta_i$  can be thought of as a parameter summarising the strength of  $i$ , then

$$\tau(\theta_i, \theta_j) + \tau(\theta_j, \theta_k) = \tau(\theta_i, \theta_k).$$

Setting  $\theta_j = \theta_i$ , it may be noted that  $\tau(\theta_i, \theta_i) = 0$  for all  $i$ . By setting  $\theta_k = \theta_i$  it may be noted that  $\tau$  is an antisymmetric function. Further, by differentiating with respect to  $\theta_i$  it may be noted that the partial derivative of  $\tau(\theta_i, \theta_j)$  with respect to  $\theta_i$  is independent of  $\theta_j$ , so that  $\tau(\theta_i, \theta_j)$  is some function of  $\theta_i$  alone plus some function of  $\theta_j$  alone, and since  $\tau$  is antisymmetric it must be of the form

$$\tau(\theta_i, \theta_j) = t(\theta_i) - t(\theta_j).$$

Since  $\theta_i$  is the strength of  $i$  then  $\tau(\theta_i, \theta_j)$  must be a monotone increasing function of  $\theta_i$  and so  $t(\theta_i)$  is a monotone increasing function of  $\theta_i$  also. Therefore,  $\lambda_i = t(\theta_i)$  is also a strength parameter for  $i$  and

$$\frac{p_{ij}}{p_{ji}} = \exp(\lambda_i - \lambda_j) \quad \text{for all } i, j,$$

giving the Bradley-Terry model.

## 2.2 Luce’s Choice Axiom (Luce, 1959)

Let  $p_S(i)$  be the probability that item  $i$  is chosen from a set  $S \subseteq T$ , then a complete system of choice probabilities satisfies Luce’s Choice Axiom if and only if for every  $i$  and for  $S \subseteq T$

$$p_S(i) = \frac{p_T(i)}{\sum_{k \in S} p_T(k)} \quad .$$

The choice axiom is a version of the decision theory axiom of the independence of irrelevant alternatives, the idea that a choice from  $S$  is independent of the other choices available in  $T$ . Luce (1959) introduces it with the assertion that many choice

situations are characterised by a multistage process, whereby a subset of the total choice set is selected, from which further subsets are selected iteratively, until a single choice is made from one of these subsets. While it is noted that the final result is likely to depend on these intermediate categorisations for complex choices and a multistage process, for a simple decision and a two stage process, it is argued that the two-stage choice, reflected by the product  $p_S(i) \sum_{k \in S} p_T(k)$ , does not depend on  $S$ , and by setting  $S = T$  it is apparent that this must be  $p_T(i)$ . The Choice Axiom itself has been motivated by appealing to the decomposition of a full ranking model (Block and Marschak, 1960), to invariance under uniform expansion of the choice set (Yellot, 1977), and, under specific assumptions, in a consideration of the utility of gambling (Luce et al., 2008).

A complete system satisfies the Choice Axiom if and only if there exist a set of numbers  $\pi_1, \pi_2, \dots, \pi_n$  such that for every  $i$  and  $S \subseteq T$

$$p_S(i) = \frac{\pi_i}{\sum_{k \in S} \pi_k} \quad .$$

In order to see this, let

$$\pi_i = \kappa p_T(i), \quad \kappa > 0,$$

then

$$\begin{aligned} p_S(i) &= \frac{p_T(i)}{\sum_{k \in S} p_T(k)} \\ &= \frac{\kappa p_T(i)}{\sum_{k \in S} \kappa p_T(k)} \\ &= \frac{\pi_i}{\sum_{k \in S} \pi_k}. \end{aligned}$$

$\pi_i$  is unique up to a multiplicative constant since suppose there is another  $\pi'_i$  satisfying this condition, then

$$\pi_i = \kappa p_T(i) = \frac{\kappa \pi'_i}{\sum_{k \in T} \pi'_k},$$

and setting  $\kappa' = \kappa / \sum_{k \in T} \pi'_k$  then  $\pi = \kappa' \pi'_i$

Taking  $S$  to be the two member set  $\{i, j\}$  gives the Bradley-Terry model.

## 2.3 Reciprocity (Block and Marschak, 1960)

What might be thought of as an alternative expression of the Choice Axiom is noted in Block and Marschak (1960). The idea is that the odds of  $i$  beating  $j$  should be



equivalent to the ratio of strength parameters of  $i$  and  $j$ .

$$\frac{p_{ij}}{p_{ji}} = \frac{\pi_i}{\pi_j} \quad \text{for all } i, j \quad .$$

Of course this condition can be framed in other familiar equivalent terms, either as detailed balance, more typically expressed as

$$p_{ij}\pi_j = p_{ji}\pi_i \quad \text{for all } i, j,$$

or that the irreducible, positive recurrent, aperiodic Markov chain for which  $P = [p_{ij}]$  is the transition matrix is reversible, which itself is the case if and only if the transitivity condition of Section 2.1 holds (Kelly, 1979). The condition leads immediately to

$$p_{ij} = \frac{\pi_i}{\pi_i + \pi_j} \quad .$$

The relationship to Markov chains is discussed further elsewhere in this work. An explicit motivation in the context of a discrete Markov chain is introduced in Section 6.4, and the discussion of Section 7 is also relevant, in particular with the link to Barker's algorithm, a prominent Markov chain Monte Carlo method, discussed in Section 7.5.

## 2.4 Wins as a sufficient statistic (Bühlmann and Huber, 1963)

Define a statistical model for pairwise comparison where the probability that  $i$  beats  $j$  is independent of other pairwise comparisons conditional on strengths  $\pi_i$  and  $\pi_j$ . Suppose  $w_i = \sum_j c_{ij}$  are the wins gained by team  $i$  and that the wins vector  $\mathbf{w} = (w_1, w_2, \dots, w_n)^T$  is a sufficient statistic for the strength vector  $\boldsymbol{\pi} = (\pi_1, \pi_2, \dots, \pi_n)^T$ .

Consider the comparison matrix  $C = [c_{ij}]$  with  $c_{kl}, c_{lm}, c_{mk}$  non-zero, for the triplet  $(k, l, m)$  where without loss of generality  $k < l < m$ . Now consider an alternative  $C'$  with  $c'_{kl} = c_{kl} - 1$ ,  $c'_{lm} = c_{lm} - 1$ ,  $c'_{mk} = c_{mk} - 1$  and  $c'_{lk} = c_{lk} + 1$ ,  $c'_{ml} = c_{ml} + 1$ ,  $c'_{km} = c_{km} + 1$ , and all else the same. Then the wins vectors for the tournaments represented by  $C$  and  $C'$  are identical. If wins are a sufficient statistic for the strength parameters then the likelihood is dependent on  $C$  only through  $\mathbf{w}$ , and so the likelihoods must also be identical. The likelihood is

$$\prod_{i < j} \binom{m_{ij}}{c_{ij}} p_{ij}^{c_{ij}} (1 - p_{ij})^{m_{ij} - c_{ij}},$$

so that the log-likelihood, up to a constant term, is

$$\sum_{i < j} c_{ij} \log \left( \frac{p_{ij}}{1 - p_{ij}} \right) + m_{ij} \log(1 - p_{ij}).$$

Setting these equal for  $C$  and  $C'$ , we get that

$$(c_{kl} - c'_{kl}) \log \frac{p_{kl}}{p_{lk}} + (c_{lm} - c'_{lm}) \log \frac{p_{lm}}{p_{ml}} + (c_{mk} - c'_{mk}) \log \frac{p_{mk}}{p_{km}} = 0,$$

and so

$$\log \frac{p_{kl}}{p_{lk}} + \log \frac{p_{lm}}{p_{ml}} + \log \frac{p_{mk}}{p_{km}} = 0,$$

by the specifications of  $c'_{kl}, c'_{lm}, c'_{mk}$ , giving the Bradley-Terry model following the same argument as in Section 2.1.

### 3 Objective function maximisation

It is a common procedure in quantitative analysis to identify an appropriate objective function and seek to maximise (or minimise) that function under certain plausible constraints. Indeed the familiarity of such procedures makes these motivations perhaps some of the most persuasive in the use of the Bradley-Terry model.

#### 3.1 Maximum entropy with retrodictive criterion (Henery, 1986; Joe, 1988)

In order to determine a functional form for the  $p_{ij}$ , we wish to select an appropriate objective function  $S(p)$ , a function of the probabilities  $p_{ij}$ , and then maximise this objective function subject to some appropriate constraint.

The proposed constraint is that of the ‘retrodictive criterion’, that the observed number of wins for each team is equal to the expected number of wins given the matches played. That is

$$w_i = \sum_j c_{ij} = \sum_j m_{ij} p_{ij} \quad \text{for all teams } i.$$

A justification for this criterion was pithily expressed by Stob (1984, p.280) in summarising the argument of Jech (1983): “What sort of a claim is it that a team solely on the basis of the results should have expected to win more games than they

did?” This would seem to fail to appreciate the bias present from finite observations; nevertheless, it reflects the intuitive appeal of the condition.

Alternatively the framework provided by Firth (2022) offers a justification for the retrodictive criterion based on two intuitive formulations for rating in this setting. The first formulation proposes that given the pairwise win probabilities  $p_{ij}$ , an intuitive rating for a team  $i$  would be the average win probability against all other competitors

$$\bar{p}_i = \frac{1}{n-1} \sum_{j=1, j \neq i}^n p_{ij}.$$

The second takes the ratio of observed wins for  $i$  divided by the ‘effective matches’ played by  $i$ ,  $w_i/m'_i$ . Effective matches played,  $m'_i$ , is chosen to account for the strength of opposition. Any definition of  $m'_i$  should meet two criteria. First, if the opponents played by  $i$  have been strictly stronger (weaker) than average, then  $m'_i$  is strictly less (greater) than  $m_i$ , the matches played by  $i$ , thus making the value of observed wins per effective matches played greater (less) than the value of observed wins per matches played. Second, observed wins per effective matches played  $w_i/m'_i$  is equal to the observed wins per matches played  $w_i/m_i$  in the case of a round-robin tournament, so that the rating accords with round-robin ranking. The simplest proposal meeting these two criteria is to scale each match played by the ratio of the probability of winning that match to the average probability of winning a match,

$$m'_i = \sum_j m_{ij} p_{ij} / \bar{p}_i.$$

If we then set these two ratings,  $\bar{p}_i$  and  $w_i/m'_i$  equal for all teams  $i$ , then we get the retrodictive criterion.

Turning to the objective function, the approach of maximising entropy is common in statistical physics. Entropy is a measure of the uncertainty of a random variable. By maximising it, roughly speaking, the assumptions in the model are minimised. Jaynes (1957) influentially advocated for the choice of entropy in a broader range of statistical settings, building on the ideas from information theory of Shannon (1948). Good et al. (1963) provides further discussion noting “[t]he mere fact that the principle of maximum entropy generates classical statistical mechanics, as a null hypothesis, would be sufficient reason for examining its implications in mathematical statistics.” Luce (1959), on the other hand, casts doubt on its applicability to choice contexts.

In this setting, the entropy is defined as

$$S(p) = - \sum_{i,j} m_{ij} p_{ij} \log p_{ij} = - \sum_{i < j} m_{ij} (p_{ij} \log p_{ij} + (1 - p_{ij}) \log(1 - p_{ij})).$$

We maximise the entropy subject to the retrodictive criterion using the method of Lagrange multipliers

$$\mathcal{L}(p, \boldsymbol{\eta}) = S(p) - \sum_{i=1}^n \eta_i \left( \sum_{j=1}^n (m_{ij} p_{ij} - c_{ij}) \right),$$

and setting  $\frac{\partial \mathcal{L}}{\partial p_{ij}} = 0$  for all  $p_{ij}$  in the normal way gives that

$$\frac{\partial S(p)}{\partial p_{ij}} = \frac{\partial}{\partial p_{ij}} \sum_{r=1}^n \eta_r \left( \sum_{s=1}^n (m_{rs} p_{rs} - c_{rs}) \right) \quad \text{for all } i, j.$$

So that for all  $i, j$  such that  $m_{ij} \neq 0$ ,

$$-\log p_{ij} + \log(1 - p_{ij}) = \eta_i - \eta_j,$$

or equivalently

$$p_{ij} = \frac{\pi_i}{\pi_i + \pi_j},$$

where  $\pi_i = \exp(-\eta_i)$ , and it can readily be checked by differentiating  $S(p)$  that this is a maximum.

### 3.2 Maximum likelihood estimation with retrodictive criterion (Joe, 1988)

Maintaining the retrodictive criterion of Section 3.1, we might consider the likelihood as an alternative objective function to maximise. This is consistent with the use of likelihood-based information criteria, such as AIC and BIC, for model choice. Suppose the probability of  $i$  being preferred to  $j$  is given by

$$p_{ij} = f(\lambda_i, \lambda_j),$$

where  $\lambda_i$  and  $\lambda_j$  are real-valued parameters describing the strength of items  $i$  and  $j$ , and  $f : \mathbb{R} \times \mathbb{R} \rightarrow [0, 1]$ . Then the likelihood function is given by

$$L(\boldsymbol{\lambda}) = \prod_{i < j} \binom{m_{ij}}{c_{ij}} p_{ij}^{c_{ij}} (1 - p_{ij})^{m_{ij} - c_{ij}} = \prod_{i < j} \binom{m_{ij}}{c_{ij}} p_{ij}^{c_{ij}} p_{ji}^{c_{ji}},$$

and the log-likelihood function, ignoring the constant term, is

$$l(\boldsymbol{\lambda}) = \sum_{i < j} c_{ij} \log(p_{ij}) + c_{ji} \log(p_{ji}).$$

At an extreme point of the log-likelihood, for all  $k$ ,

$$0 = \frac{\partial}{\partial \lambda_k} l(\boldsymbol{\lambda}) = \sum_j c_{kj} \frac{\partial}{\partial \lambda_k} \log(p_{kj}) + c_{jk} \frac{\partial}{\partial \lambda_k} \log(p_{jk}).$$

Considering the constraint we note that

$$\begin{aligned} 0 &= \sum_j c_{kj} - m_{kj} p_{kj} = \sum_j c_{kj} - (c_{kj} + c_{jk}) p_{kj} \\ &= \sum_j c_{kj} (1 - p_{kj}) - c_{jk} p_{kj} \\ &= \sum_j c_{kj} (1 - p_{kj}) - c_{jk} (1 - p_{jk}), \end{aligned}$$

and so there is an extreme point where

$$\begin{aligned} \frac{\partial}{\partial \lambda_k} \log(p_{kj}) &= (1 - p_{kj}) \quad \text{and} \\ \frac{\partial}{\partial \lambda_k} \log(p_{jk}) &= -(1 - p_{jk}), \end{aligned}$$

which gives

$$\begin{aligned} \frac{\partial p_{kj}}{\partial \lambda_k} &= p_{kj} (1 - p_{kj}) \quad \text{and} \\ \frac{\partial p_{jk}}{\partial \lambda_k} &= -p_{jk} (1 - p_{jk}). \end{aligned}$$

Solving these separable differential equations for  $p_{ij}$  gives

$$\begin{aligned} p_{ij} &= \frac{1}{1 + e^{-(\lambda_i - \lambda_j)}} \\ &= \frac{\pi_i}{\pi_i + \pi_j} \end{aligned}$$

where  $\pi_i = e^{\lambda_i}$ , and, as before, this is a maximum since the log-likelihood is strictly concave. So that the Bradley-Terry model is the likelihood maximising model.

### 3.3 Geometric minimisation (McCullagh, 1993)

If one were to conceive of the rating of  $n$  items under a geometric interpretation, a natural general framing might be that the observed results are represented as vectors

in some  $n$ -space and then a rating vector can be taken as the vector that minimises some aggregate quantity with respect to these observed result vectors. McCullagh (1993) presents just such a framing with the outcome and rating vectors confined to a  $n$ -sphere, taken to be of unit radius for convenience. For example, in a five-team tournament consisting of competitors  $A, B, C, D, E$  then a win for  $D$  over  $B$  would be represented by the result vector  $\mathbf{x} = (0, -1/\sqrt{2}, 0, 1/\sqrt{2}, 0)$ .

With both the rating vector and observed result vectors lying on the unit sphere, a natural quantity to seek to minimise is the angle between the rating vector and an observed result vector, or equivalently maximising the cosine of the angle as expressed through the dot product of the vectors,  $\mathbf{x} \cdot \lambda$ . Note that this is equivalent to minimising the squared Euclidean distance between the points on the sphere since

$$\|\mathbf{x} - \lambda\|^2 = 2 - 2\mathbf{x} \cdot \lambda.$$

So to find our rating vector  $\lambda$ , we would sum the dot product over all observed results and select  $\lambda$  such that it maximises this quantity.

In the notation used in this paper, and keeping the unit radius, any result vector  $\mathbf{x}_{ij}$  representing a win for  $i$  over  $j$  will have value  $1/\sqrt{2}$  in the  $i$ th position,  $-1/\sqrt{2}$  in the  $j$ th position and zero elsewhere. The sum over all such results is therefore

$$\frac{1}{\sqrt{2}} \sum_{i,j} c_{ij} (\lambda_i - \lambda_j) = \frac{1}{\sqrt{2}} \sum_{i,j} \lambda_i (c_{ij} - c_{ji}),$$

which is the form of the likelihood maximisation that gives the Bradley-Terry rating (see Section 8.1 for further details). Thus, a geometric interpretation of rating where one minimises the aggregate angles between results and rating on a sphere returns the Bradley-Terry ratings. One nice feature of this motivation is the ready extendability to scenarios of differing numbers of competitors in each contest, while maintaining consistency with Bradley-Terry in the pairwise contest case. This is discussed further in Section 8.2.1.

McCullagh (1993) also demonstrates the link to Mallows'  $\phi$ -model and the von Mises-Fisher distribution. These are presented in Section 5.

### 3.4 Definitional simplicity 1

Often when selecting a model, transparency and interpretability are desirable features. This may be especially so in contexts where fairness of a ranking system are a consideration. These sort of contexts are common in pairwise comparison with the methods being used to perform activities like ranking sports teams (Firth, 2022) or

in educational assessment (Pollitt, 2012b). Therefore, there may be a legitimate desire for definitionally simpler, more intuitive models. It is thus appealing to consider how one might select a model with the goal of maximising definitional simplicity.

Suppose one wished to determine a ranking by defining a probability for the preference for  $i$  over  $j$  related only to positive real-valued strength parameters  $\pi_i$  and  $\pi_j$  respectively,

$$p_{ij} = f(\pi_i, \pi_j).$$

A reasonable set of criteria for this function would be:

1.  $f : \mathbb{R}^+ \times \mathbb{R}^+ \rightarrow [0, 1]$ ,
2.  $f(\pi_i, \pi_j) = \frac{1}{2}$  when  $\pi_i = \pi_j$ ,
3.  $\lim_{\pi_i \rightarrow 0, \pi_j \text{ fixed}} f(\pi_i, \pi_j) = 0$ ,
4.  $\lim_{\pi_j \rightarrow 0, \pi_i \text{ fixed}} f(\pi_i, \pi_j) = 1$ ,
5.  $\lim_{\pi_i \rightarrow \infty, \pi_j \text{ fixed}} f(\pi_i, \pi_j) = 1$ ,
6.  $\lim_{\pi_j \rightarrow \infty, \pi_i \text{ fixed}} f(\pi_i, \pi_j) = 0$ .

where  $\mathbb{R}^+$  is taken to be the set of positive real numbers not including zero.

Assume that the simplest set of functions are those that may be defined solely using the four basic operators  $(+, -, \times, \div)$ , and that any measure of the simplicity of a function is a strictly decreasing function of the number of these operators used. So, in this setup, maximising definitional simplicity of a function is equivalent to minimising the number of basic operators in its definition. Bracketing anywhere, used in the conventional sense, to identify a functional subclause, is allowed without increasing or reducing simplicity. Constants are also allowed in place of parameters without increasing or reducing simplicity. In the language of Computer Science, this is therefore defining simplicity by the minimum number of floating point operations (flops).

No  $f$  with exactly zero or one operator can meet criterion 5 other than  $f(\pi_i, \pi_j) = 1$  or equivalents (for example,  $f(\pi_i, \pi_j) = \pi_i/\pi_i$ ), which violates criteria 2, 3 and 6. Likewise, considering a function with two operators and again considering criterion 5, then it must be that the operator  $\div$  is employed as otherwise the limit of criterion 5 would be infinite in absolute value other than in cases which are equivalent to a constant (for example,  $f(\pi_i, \pi_j) = \pi_i + (1 - \pi_i)$ ) or where  $\pi_i$  is not included, but if  $\pi_i$  is not included then criteria 2 and 3 will be in contradiction. So if there is a solution with exactly two operators then it must be of the form  $f(\pi_i, \pi_j) = g(\pi_i, \pi_j) \div h(\pi_i, \pi_j)$

where either  $g$  or  $h$  is equal to either one of the parameters or to a constant in order that only two operators are used, and the other must be a single operator function involving  $+$  or  $-$  in order to meet criterion 5 without being equivalent to a constant (for example,  $f(\pi_i, \pi_j) = \pi_i \div (c \times \pi_i)$ ). From criterion 3 it must be that  $g(\pi_i, \pi_j) = \pi_i$  and then from criterion 5,  $h$  must take  $\pi_i$  as one of its terms. Criterion 6 implies that the other term in  $h$  is  $\pi_j$  and criterion 2 then implies that  $h(\pi_i, \pi_j) = \pi_i + \pi_j$ . This gives  $f(\pi_i, \pi_j) = \pi_i \div (\pi_i + \pi_j)$ , which meets all the required criteria. It may be noted that not all the criteria were required for its unique derivation, and that other subsets of the criteria may be used to derive the same result. That is to say that

$$p_{ij} = \frac{\pi_i}{\pi_i + \pi_j}$$

will be the unique simplicity maximiser under a number of different subsets of the plausible criteria.

### 3.5 Definitional simplicity 2

Given positive-valued strength parameters  $\pi_i$  and  $\pi_j$  for  $i$  and  $j$  respectively, one may want to consider a model where the probability of  $i$  being preferred to  $j$  is a function of the ratio  $x_{ij} = \pi_i/\pi_j$ ,

$$p_{ij} = f(x_{ij}).$$

A reasonable set of criteria for this function would then be:

1.  $f : \mathbb{R}^+ \rightarrow [0, 1]$ ,
2.  $f(1) = \frac{1}{2}$ ,
3.  $\lim_{x \rightarrow 0} f(x) = 0$ ,
4.  $\lim_{x \rightarrow \infty} f(x) = 1$ ,

Proceeding in a similar fashion to the previous section, the only function including exactly zero or one flop that meets criterion 4 is  $f(x) = 1$  (or equivalents, for example,  $f(x) = x \div x$ ), but this violates criteria 2 and 3. Considering a function with two operators and again considering criterion 4, then it must be that the operator  $\div$  is employed as otherwise the limit would be infinite in absolute value other than in cases which are equivalent to a constant (for example,  $f(x) = x + (1 - x)$ ). So if there is a solution with exactly two operators then it must be of the form  $f(x) = g(x) \div h(x)$



where either  $g(x) = x$  or  $h(x) = x$  or  $g(x) = \text{constant}$  or  $h(x) = \text{constant}$  in order that only two operators are used, and the other must be a single-operator function involving  $+$  or  $-$  in order to meet criterion 4. Criterion 3 implies that  $g(x) = x$ , and criterion 2 then tells us that  $h(x) = 1 + x$ . Thus

$$f(x) = \frac{x}{1+x},$$

giving

$$p_{ij} = \frac{\pi_i}{\pi_i + \pi_j}.$$

## 4 Discriminal processes

Consider a scenario where the strength of each of two entities in a given pairwise interaction is independently observed with error and then compared. The item with the greater observed strength is preferred. This is the model of Thurstone's 'discriminal processes' (Thurstone, 1927a). Denote the observed strength of  $i$  as  $b_i$  with 'true' strength  $\lambda_i$ , so that  $b_i = \lambda_i + \epsilon_i$ , where  $\epsilon_i$  is an error term. Item  $i$  is preferred to item  $j$  if and only if  $b_i > b_j$ . Taking the error to be Gaussian, as Thurstone himself did, leads to what is commonly known as the Thurstone-Mosteller model (Thurstone, 1927a; Mosteller, 1951), but the set up may also be used to motivate the Bradley-Terry model by considering alternative distributions for  $b_i$ .

### 4.1 Exponential Distribution (Holman and Marley as cited by Luce and Suppes (1965, p.338))

Suppose  $b_i$  and  $b_j$  follow independent exponential distributions whose expected values are given by  $\pi_i$  and  $\pi_j$  respectively with the cdf,

$$F_i(x) = 1 - e^{-\frac{x}{\pi_i}}, \quad x \in \mathbb{R}^+.$$

Then, with  $F'$  denoting the pdf, the probability that  $i$  is preferred to  $j$  in a pairwise comparison is

$$\begin{aligned} p_{ij} &= \int_0^\infty F_j(x) F'_i(x) dx \\ &= \int_0^\infty \left(1 - e^{-\frac{x}{\pi_j}}\right) \frac{1}{\pi_i} e^{-\frac{x}{\pi_i}} dx \end{aligned}$$

$$\begin{aligned}
&= 1 - \frac{1}{\pi_i \left( \frac{1}{\pi_i} + \frac{1}{\pi_j} \right)} \int_0^\infty \left( \frac{1}{\pi_i} + \frac{1}{\pi_j} \right) e^{-\left( \frac{1}{\pi_i} + \frac{1}{\pi_j} \right) x} dx \\
&= 1 - \frac{\pi_j}{\pi_i + \pi_j} \\
&= \frac{\pi_i}{\pi_i + \pi_j} .
\end{aligned}$$

## 4.2 Extreme value distributions (Bradley, 1965; Thompson and Singh, 1967)

Thompson and Singh (1967) provide a rationale for a broader class of distributions that lead to a Bradley-Terry model under a discriminial process. Based on ideas from Psychology, sensations are hypothesised to be a result of a large number of stimuli. These stimuli are modeled as having independent identical distributions  $G(x)$ . One might then consider the distribution of the resultant sensation.

Two intuitive possibilities would be to model the distribution of the sensation  $F(x)$  either as the average of those stimuli or the maximum of those stimuli. Taking the average gives a normal distribution for  $F(x)$  in the limit, leading to a Thurstone-Mosteller comparison model. Taking the maximum of the stimuli, in the limit, gives, by extreme value theorem (Fisher and Tippett, 1928; Gnedenko, 1943; Gumbel, 1958), one of three distributions for  $F(x)$  — Gumbel, Weibull, or Frechet — depending on the underlying stimuli distribution  $G(x)$ , leading to a Bradley-Terry comparison model. The Gumbel is the most notable of these, being the sensation distribution for stimuli distributions such as the normal, lognormal, logistic, and exponential.

While Thompson and Singh (1967) provided a clear motivation for considering such models and do not assume that the underlying stimuli distributions need have the same location parameters for  $i$  and  $j$ , Lehmann (1953) had previously considered a family of distributions in the context of the power of rank tests of the form  $F_{X_i}(x; \pi_i) = G^{\pi_i}(x)$ , where  $G(x)$  is itself a distribution function. Bradley (1965) discussed this family of distributions with respect to the Bradley-Terry model. As Bradley (1976) notes, if  $G(x)$  is a distribution function, and  $X_i$  is the random variable relating to a sensation  $i$ , with distribution function

$$\mathbb{P}(X_i \leq x) = G^{\pi_i}(x),$$

where  $\pi_i > 0$ , then comparing sensations  $i$  and  $j$ ,

$$p_{ij} = \mathbb{P}(X_i > X_j) = \int_{x_i > x_j} dG^{\pi_i}(x_i) dG^{\pi_j}(x_j) = \frac{\pi_i}{\pi_i + \pi_j}, \quad i \neq j.$$

#### 4.2.1 Gumbel distribution (Thompson and Singh, 1967)

Suppose  $b_i$  follows a Gumbel distribution with mean  $\lambda_i$ . Then

$$\Pr(b_i \leq x) = F_i(x) = \exp(-\pi_i e^{-\alpha x}) \text{ for } x \in \mathbb{R} \text{ and parameter } \alpha > 0,$$

where  $\pi_i = e^{\alpha \lambda_i - \gamma}$ , with  $\gamma$  the Euler-Mascheroni constant. Then the probability that  $i$  is preferred to  $j$  in a pairwise comparison is

$$\begin{aligned} p_{ij} &= \int_{-\infty}^{\infty} F_j(x) F'_i(x) dx \\ &= \int_{-\infty}^{\infty} \exp(-\pi_j e^{-\alpha x}) \alpha \pi_i \exp(-\alpha x - \pi_i e^{-\alpha x}) dx \\ &= \frac{\pi_i}{\pi_i + \pi_j} \int_{-\infty}^{\infty} \alpha (\pi_i + \pi_j) \exp(-\alpha x - (\pi_i + \pi_j) e^{-\alpha x}) dx \\ &= \frac{\pi_i}{\pi_i + \pi_j}. \end{aligned}$$

#### 4.2.2 Weibull distribution (Thompson and Singh, 1967)

Suppose  $b_i$  follows a Weibull distribution

$$\mathbb{P}(b_i \leq x) = F_i(x) = 1 - \exp(-(x/\lambda_i)^\alpha) \text{ for } x \in \mathbb{R}^+ \text{ and parameter } \alpha > 0.$$

Then the probability that  $i$  is preferred to  $j$  in a pairwise comparison is

$$\begin{aligned} p_{ij} &= \int_0^{\infty} F_j(x) F'_i(x) dx \\ &= \int_0^{\infty} [1 - \exp(-(x/\lambda_j)^\alpha)] \frac{\alpha}{\lambda_i} (x/\lambda_i)^{\alpha-1} \exp(-(x/\lambda_i)^\alpha) dx \\ &= 1 - \int_0^{\infty} \frac{\alpha}{\lambda_i} (x/\lambda_i)^{\alpha-1} \exp(-(x/\lambda_j)^\alpha) - (x/\lambda_i)^\alpha dx \\ &= 1 - \frac{\lambda_j^\alpha}{\lambda_i^\alpha + \lambda_j^\alpha} \int_0^{\infty} \frac{\alpha}{\lambda_i \lambda_j} (x/\lambda_i \lambda_j)^{\alpha-1} (\lambda_i^\alpha + \lambda_j^\alpha) \exp(-(x/\lambda_i \lambda_j)^\alpha (\lambda_i^\alpha + \lambda_j^\alpha)) dx \\ &= \frac{\pi_i}{\pi_i + \pi_j}, \end{aligned}$$

where  $\pi_i = \lambda_i^\alpha$ .

### 4.2.3 Fréchet distribution (Thompson and Singh, 1967)

Suppose  $b_i$  follows a Frechet distribution

$$\mathbb{P}(b_i \leq x) = F_i(x) = \exp(-\pi_i x^{-\alpha}) \text{ for } x \in \mathbb{R}^+ \text{ and parameter } \alpha > 0.$$

Then the probability that  $i$  is preferred to  $j$  in a pairwise comparison is

$$\begin{aligned} p_{ij} &= \int_0^\infty F_j(x) F'_i(x) dx \\ &= \int_0^\infty \exp(-\pi_j x^{-\alpha}) \frac{\pi_i \alpha}{x^{\alpha+1}} \exp(-\pi_i x^{-\alpha}) dx \\ &= \frac{\pi_i}{\pi_i + \pi_j} \int_0^\infty \alpha \frac{\pi_i + \pi_j}{x^{\alpha+1}} \exp(-( \pi_i + \pi_j ) x^{-\alpha}) dx \\ &= \frac{\pi_i}{\pi_i + \pi_j}. \end{aligned}$$

## 5 Standard models

A number of models familiar to statisticians may be related to the Bradley-Terry model by considering conditional forms. In Section 3.3, we noted how McCullagh (1993) demonstrated links to Mallows'  $\phi$ -model and the von Mises-Fisher distribution. We expand on those links here and also discuss the relation to three more models familiar to statisticians.

### 5.1 Rasch model (Andrich, 1978)

Let  $X_{vi}$  be a binary random variable, representing the outcome of a test  $v$  taken by candidate  $i$ , where  $X_{vi} = 1$  represents passing the test, and  $X_{vi} = 0$  denotes failure. Under the Rasch simple logistic model (Rasch, 1960, 1961) the probability of the outcome  $X_{vi} = 1$  is taken to be

$$\mathbb{P}(X_{vi} = 1) = \frac{e^{\lambda_i - \delta_v}}{1 + e^{\lambda_i - \delta_v}},$$

where  $\lambda_i$  represents the ability of candidate  $i$  and  $\delta_v$  the difficulty of test  $v$ .

There are two conceptualisations by which we might derive the Bradley-Terry model from this. First, as Andrich (1978) notes, if we take

$$p_{ij} = \mathbb{P}(i \text{ passes a test } v \mid \text{exactly one of } i \text{ and } j \text{ pass the test } v),$$

then since

$$\mathbb{P}(X_{vi} = 1, X_{vj} = 0) = \frac{e^{\lambda_i - \delta_v}}{(1 + e^{\lambda_i - \delta_v})(1 + e^{\lambda_j - \delta_v})},$$

and

$$\mathbb{P}(X_{vi} + X_{vj} = 1) = \frac{e^{\lambda_i - \delta_v} + e^{\lambda_j - \delta_v}}{(1 + e^{\lambda_i - \delta_v})(1 + e^{\lambda_j - \delta_v})}$$

then conditional on being able to discern that one of the test-takers has performed better based on the binary test outcome and taking their test outcomes to be independent conditional on their abilities and the test difficulty then the probability that  $i$  has beaten  $j$  is

$$p_{ij} = \frac{\mathbb{P}(X_{vi} = 1, X_{vj} = 0)}{\mathbb{P}(X_{vi} + X_{vj} = 1)} = \frac{e^{\lambda_i}}{e^{\lambda_i} + e^{\lambda_j}} = \frac{\pi_i}{\pi_i + \pi_j},$$

where  $\pi_i = e^{\lambda_i}$ .

Second, we might more directly consider that in comparing  $i$  with  $j$  we are setting a test for  $i$  of difficulty equal to the strength of the comparator  $\lambda_j$  (or equivalently setting a test for  $j$  of difficulty equal to the strength of the comparator  $\lambda_i$ ), so that

$$p_{ij} = \frac{e^{\lambda_i - \lambda_j}}{1 + e^{\lambda_i - \lambda_j}} = \frac{\pi_i}{\pi_i + \pi_j}.$$

## 5.2 Mallows' $\phi$ -model (McCullagh, 1993)

Mallows (1957) discusses models on the space of permutations. In the context of this paper, a permutation might equivalently be thought of as a ranking. The simplest of these model families is Mallows'  $\phi$ -model,

$$p(\mathbf{x}) = K_\phi \exp \{-\phi d(\mathbf{x}, \lambda)\},$$

where  $d$  is a distance measure between an observed permutation  $\mathbf{x}$  and the 'modal permutation'  $\lambda$ ,  $\phi$  is a concentration parameter and  $K_\phi$  is a constant of proportionality. Thus, in maximising the likelihood of the model given observed permutations, the modal permutation is the permutation that has the minimum aggregate distance to the observed permutations. In considering distances on permutations or ranks, the Spearman rank correlation coefficient is a natural candidate and is the one considered here.

McCullagh (1993) notes that the use of ordinal numbers to represent ranks, while a strong norm, is somewhat arbitrary. He proposes that the ranks are transformed

such that

$$k' = \frac{k - (n + 1)/2}{\sqrt{n(n^2 - 1)/12}},$$

where  $k$  is an integer from 1 to  $n$  representing a rank. With this transformation, the rank permutations are projected onto the unit sphere. For this paper, we consider a ‘modal rating’ rather than a modal permutation or ranking and thus take the negative of this transformation to ensure that higher-ranked items have higher value. For example, a rank vector  $(2, 3, 1, 4)$ , expressing that item 1 came second, item 2 third etc. would be transformed to the vector  $\frac{1}{2\sqrt{5}}(1, -1, 3, -3)$ . The observed pairwise results may be projected onto the unit sphere with the pairwise ranking of a win for  $i$  over  $j$  represented by a vector with the value  $1/\sqrt{2}$  in the  $i$ th position,  $-1/\sqrt{2}$  in the  $j$ th position and zero elsewhere.

If the pairwise results are represented in this way and the distance measure is taken to be the Spearman rank correlation coefficient, which is equivalent to the squared Euclidean distance, then following the argument from Section 3.3, the Mallows’  $\phi$ -model becomes equivalent in form to the Bradley-Terry model, with the modal permutation vector equal to the vector of Bradley-Terry ratings. McCullagh (1993) notes that, strictly speaking, the models are not equivalent. Under the proposed ranking transformation, the Mallows’  $\phi$ -model is defined on the sample space of permutations represented on the unit sphere and has a ranking as the parameter  $\lambda$ , whereas the Bradley-Terry model is defined on the sample space of pairwise unit vectors and takes  $\lambda$  to be any point on the unit sphere.

### 5.3 von Mises-Fisher distribution (McCullagh, 1993)

The von Mises-Fisher distribution (von Mises, 1918; Fisher, 1953) is a well-known model in directional statistics. It defines a probability density for a random  $n$ -dimensional unit vector  $\mathbf{x}$  as

$$p(\mathbf{x}; \lambda) = C_\kappa \exp \{ \kappa \mathbf{x} \cdot \lambda \}.$$

As discussed in Sections 3.3 and 5.2, if we take the pairwise result outcomes and the Bradley-Terry rating vector to be defined on the unit sphere then this takes the same form as the Bradley-Terry model. As with the Mallows’ model, McCullagh (1993) notes that strictly speaking they are not equivalent due to being defined on different sample spaces. In this case, the von Mises-Fisher distribution is defined on the continuous sample space of the unit sphere, whereas the Bradley-Terry model takes the pairwise rankings projected onto the unit sphere as its sample space.

## 5.4 Cox proprtrional hazards model (Su and Zhou, 2006)

Consider a proportional hazards model (Cox, 1972) on random variables  $T_i$  with hazard function given by

$$h_i(t) = h(t)\pi_i.$$

Thus the hazard rate for object  $i$  is given by a multiplicative factor  $\pi_i$ . Then

$$\begin{aligned}\mathbb{P}(T_i < T_j) &= \int_0^\infty F_{T_i}(t) f_{T_j}(t) dt \\ &= \int_0^\infty \left(1 - \exp\left\{-\int_0^t h(x)\pi_i dx\right\}\right) h(t)\pi_j \exp\left\{-\int_0^t h(x)\pi_j dx\right\} dt \\ &= 1 - \int_0^\infty h(t)\pi_j \exp\left\{-(\pi_i + \pi_j) \int_0^t h(x) dx\right\} dt \\ &= 1 - \frac{\pi_j}{\pi_i + \pi_j} \\ &= \frac{\pi_i}{\pi_i + \pi_j}.\end{aligned}$$

Further, as Su and Zhou (2006) note, if a stratified proportional hazards model is used such that each stratum represents a different match with

$$h_{ij}(t) = h_{s_{ij}}(t)\pi_i,$$

where  $s_{ij}$  is the stratum for a match between  $i$  and  $j$  then the contribution to the partial likelihood from the random variables  $T_i$  and  $T_j$  with the event  $\{T_i < T_j\}$  is  $\pi_i/(\pi_i + \pi_j)$ .

## 5.5 Network models

Consider a binary directed network  $Y$ , with an edge  $i \rightarrow j$  taking the value  $y_{ij}$ . A common class of models in network analysis takes a conditional independence approach, assuming that the value of any directed edge is independent of all other edge values given an appropriate set of parameters. In a generalised form for the current purposes it can be expressed as

$$\begin{aligned}\mu_{ij} &= \mathbb{P}(y_{ij} = 1) \\ \text{logit}(\mu_{ij}; \delta_i, \gamma_j, f_{ij}) &= \delta_i + \gamma_j + f_{ij},\end{aligned}$$

where  $\delta_i$  and  $\gamma_j$ , sometimes referred to as *sociality* and *attractivity* parameters (Krivitsky et al., 2009), reflect the heterogeneity of out-degree and in-degree respectively, and  $f_{ij} = f(i, j)$  is a symmetric function capturing the propensity for an

edge in either direction to exist. For example, Hoff et al. (2002) takes  $f(i, j)$  to be the Euclidean distance between points associated with  $i$  and  $j$  in a latent space but note that  $f(i, j)$  could be any distance measure satisfying the triangle inequality  $f(i, j) \leq f(i, k) + f(k, j)$ . Often models also incorporate a term of the form  $\beta^T x_{ij}$  within  $f(i, j)$ , where  $x_{ij}$  is a vector of pair-specific characteristics, in order to capture known homophilies.

Applying the conditional independence assumption and looking at the probability of an edge being present only in the direction  $i \rightarrow j$  and not the  $j \rightarrow i$  direction,

$$\mathbb{P}(y_{ij} = 1, y_{ji} = 0; \delta_i, \delta_j, \gamma_i, \gamma_j, f_{ij}) = \frac{e^{\delta_i + \gamma_j + f_{ij}}}{(1 + e^{\delta_i + \gamma_j + f_{ij}})(1 + e^{\delta_j + \gamma_i + f_{ji}})},$$

so

$$\begin{aligned} \mathbb{P}(y_{ij} = 1 \mid y_{ij} + y_{ji} = 1; \delta_i, \delta_j, \gamma_i, \gamma_j, f_{ij}) &= \frac{e^{\delta_i + \gamma_j + f_{ij}}}{e^{\delta_i + \gamma_j + f_{ij}} + e^{\delta_j + \gamma_i + f_{ji}}} \\ &= \frac{e^{\delta_i - \gamma_i}}{e^{\delta_i - \gamma_i} + e^{\delta_j - \gamma_j}} \\ &= \frac{e^{\lambda_i}}{e^{\lambda_i} + e^{\lambda_j}} \\ &= \frac{\pi_i}{\pi_i + \pi_j}, \end{aligned}$$

where  $\pi_i = e^{\lambda_i}$  and  $\lambda_i = \delta_i - \gamma_i$ . If  $Y$  is considered as a tournament matrix with a directed edge  $i \rightarrow j$  indicating  $i$  beats  $j$ , then *sociality* is a team's propensity for winning and *attractivity* the propensity for losing so that assessing the strength of a team as the difference between these is readily intuitive.

## 6 Game scenarios

The Bradley-Terry model has frequently been associated with an analysis of sport. So it is perhaps not surprising that there are a number of game scenarios in which the model may be very naturally motivated. Some of these are presented here.

### 6.1 Poisson scoring (Audley, 1960; Stern, 1990)

Consider two teams  $i$  and  $j$  who score according to independent Poisson processes  $X_i(t)$  and  $X_j(t)$  with rate parameters  $\pi_i$  and  $\pi_j$  respectively. The winner is the first



team to score. Then by Poisson thinning, for any value of  $t$ ,

$$p_{ij} = \mathbb{P}(X_i(t) = 1 \mid X_i(t) + X_j(t) = 1) = \frac{\pi_i}{\pi_i + \pi_j} \quad .$$

Audley (1960) presents an argument for this framing based in the psychological literature, considering the probability of one response occurring before another, where the probability of a response occurring in any given small time interval is determined by a response-specific parameter. While the argument is presented in terms of discrete time, it notes that the continuous alternative would be to consider Poisson distributions. Stern (1990) notes that the context may be widened to that of two gamma random variables with the same shape parameter and different scale parameters, showing that taking a shape parameter of one returns the Bradley-Terry model, whereas allowing it to tend to infinity sees the model tend to the Thurstone-Mosteller model. The idea might also be considered in the context of the discriminial process on exponential distributions of Section 4.1, since the interarrival time of a homogeneous Poisson process with rate parameter  $\lambda$  has an exponential distribution with a mean  $1/\lambda$ .

More directly it is simply an expression of the standard equivalence between a multinomial distribution, in this case Bernoulli, and independent Poisson distributions conditional on their total, sometimes referred to as the ‘‘Poisson trick’’ (Fienberg and Larntz, 1976; Lee et al., 2017).

## 6.2 Sudden death (Stirzaker, 1999; Vojnović, 2015)

Consider two teams  $i$  and  $j$  involved in a ‘sudden death’ shoot-out. They play a game where in each round they succeed with independent probabilities  $p_i$  and  $p_j$  respectively. The winner is the team who first has more successes than the other team. Let  $(i \succ j)_n$  be the event that  $i$  wins the ‘sudden death’ contest in round  $n$ . Then

$$\begin{aligned} p_{ij} &= \sum_{n=1}^{\infty} \mathbb{P}[(i \succ j)_n] \\ &= \sum_{n=1}^{\infty} \sum_{k=0}^{n-1} p_i(1-p_j) \binom{n-1}{k} (p_i p_j)^k ((1-p_i)(1-p_j))^{n-k-1} \\ &= p_i(1-p_j) \sum_{m=0}^{\infty} \sum_{k=0}^m \binom{m}{k} (p_i p_j)^k ((1-p_i)(1-p_j))^{m-k} \end{aligned}$$

$$\begin{aligned}
&= p_i(1-p_j) \sum_{m=0}^{\infty} (p_i p_j + (1-p_i)(1-p_j))^m \\
&= p_i(1-p_j) \sum_{m=0}^{\infty} (2p_i p_j - p_i - p_j + 1)^m \\
&= \frac{p_i(1-p_j)}{p_i + p_j - 2p_i p_j} \\
&= \frac{p_i(1-p_j)}{p_i(1-p_j) + p_j(1-p_i)} \\
&= \frac{\frac{p_i}{1-p_i}}{\frac{p_i}{1-p_i} + \frac{p_j}{1-p_j}} \\
&= \frac{\pi_i}{\pi_i + \pi_j},
\end{aligned}$$

where  $\pi_i = \frac{p_i}{1-p_i}$ .

Further, suppose there is an alternative contest but now the winner is the team that is the first to have  $r$  more successes than the opposition. Such a contest may be understood as an aggregation of the sudden death contests described above, such that the winner is the first team to win  $r$  more sudden death contests than the opposition. Based on the result above, given that there is a winner to a sudden death contest, the probability that the winner is  $i$  is  $q_i = p_i/(1-p_i)$ . Let  $A_i$  be the event that  $i$  wins and  $A^{r+k}$  be the event that a result, either  $i$  or  $j$  winning, occurs after the winning team has won exactly  $r+k$  sudden death contests then

$$\begin{aligned}
p_{ij} = \mathbb{P}(A_i) &= \sum_{k=0}^{\infty} \mathbb{P}(A_i | A^{r+k}) \mathbb{P}(A^{r+k}) \\
&= \sum_{k=0}^{\infty} \frac{q_i^{r+k} q_j^k}{q_i^{r+k} q_j^k + q_i^k q_j^{r+k}} P(A^{r+k}) \\
&= \frac{q_i^r}{q_i^r + q_j^r} \sum_{k=0}^{\infty} P(A^{r+k}) \\
&= \frac{q_i^r}{q_i^r + q_j^r} \\
&= \frac{\pi_i}{\pi_i + \pi_j},
\end{aligned}$$

where  $\pi_i = q_i^r$ .

### 6.3 Accumulated win ratio (Vojnović, 2015)

Take a sequence of matches between two players,  $i$  and  $j$ , where the probability that team  $i$  wins is proportional to the accumulated number of wins in previous matches. Suppose that the probability that  $i$  wins the first match is  $\pi_i/(\pi_i + \pi_j)$ . Then consider the probability that  $i$  will win the  $n$ th match. The claim is that this is  $\pi_i/(\pi_i + \pi_j)$ . We proceed to show this by induction. Define notation  $(i \succ j)_n$  as meaning  $i$  beats  $j$  in match  $n$  then our base case is

$$\mathbb{P}[(i \succ j)_1] = \frac{\pi_i}{\pi_i + \pi_j}.$$

Now assume the inductive hypothesis for some  $k > 1$

$$\mathbb{P}[(i \succ j)_k] = \frac{\pi_i}{\pi_i + \pi_j}.$$

Then proceeding by induction

$$\begin{aligned} \mathbb{P}[(i \succ j)_{k+1}] &= \mathbb{P}[(i \succ j)_{k+1} \mid (i \succ j)_k] \mathbb{P}[(i \succ j)_k] \\ &\quad + \mathbb{P}[(i \succ j)_{k+1} \mid (j \succ i)_k] \mathbb{P}[(j \succ i)_k] \\ &= \frac{\pi_i + 1}{\pi_i + 1 + \pi_j} \frac{\pi_i}{\pi_i + \pi_j} + \frac{\pi_i}{\pi_i + 1 + \pi_j} \frac{\pi_j}{\pi_i + \pi_j} \\ &= \frac{\pi_i(\pi_i + 1 + \pi_j)}{(\pi_i + 1 + \pi_j)(\pi_i + \pi_j)} \\ &= \frac{\pi_i}{\pi_i + \pi_j}. \end{aligned}$$

### 6.4 Continuous time state transition

Consider a match where the winner is the team winning at the end of a defined period of play. We choose to model the continuous state of ‘winning’ by a continuous time Markov chain on a binary state space  $I = \{i \text{ winning}, j \text{ winning}\}$ . Let the rate at which there is a switch from the state ‘ $i$  winning’ to the state ‘ $j$  winning’ be denoted by  $\pi_j$ , and the rate at which the switch from the state ‘ $j$  winning’ to the state ‘ $i$  winning’ be denoted by  $\pi_i$ . Then the intensity matrix is

$$Q = \begin{pmatrix} -\pi_j & \pi_j \\ \pi_i & -\pi_i \end{pmatrix}$$

and the equilibrium distribution vector of this process  $\mathbf{p}$  is such that

$$\mathbf{p}Q = \mathbf{0},$$

and in this case is given by the probability vector  $\mathbf{p} = (\frac{\pi_i}{\pi_i + \pi_j}, \frac{\pi_j}{\pi_i + \pi_j})$ .

Assuming that we are likely to see a large number of state changes during the course of the match or the probability of the initial state being ‘ $i$  winning’ may be approximated by  $\pi_i/(\pi_i + \pi_j)$  then the probability that  $i$  beats  $j$  may be approximated by

$$p_{ij} = \frac{\pi_i}{\pi_i + \pi_j}.$$

The authors are not aware of published work that uses this continuous-time model, which might reasonably be called the “Bradley-Terry process” model.

## 7 Quasi-symmetry and consistent estimators

The quasi-symmetry model was proposed by Caussinus (1965). A matrix  $C$  is quasi-symmetric if it can be decomposed such that

$$c_{ij} = \alpha_i \beta_j \gamma_{ij},$$

where  $\gamma_{ij} = \gamma_{ji}$ . The form of this can be simplified by taking  $a_i = \alpha_i/\beta_i$  and  $s_{ij} = \beta_i \beta_j \gamma_{ij}$ , so that

$$c_{ij} = a_i s_{ij},$$

or in matrix form

$$C = AS,$$

where  $A$  is a diagonal matrix and  $S$  is symmetric. Informally, one might think of the symmetric matrix representing the intensity of interactions, and the diagonal matrix as the relative ratings. Asymptotically, where the number of matches between each pair of teams tends to infinity and the number of teams is held constant, by the Law of Large Numbers, under a Bradley-Terry data generating process, we would expect the results matrix to be quasi-symmetric, since

$$\mathbb{E}[c_{ij}] = p_{ij} m_{ij} = \frac{\pi_i}{\pi_i + \pi_j} m_{ij} = a_{ii} s_{ij},$$

where  $s_{ij} = m_{ij}/(\pi_i + \pi_j) = s_{ji}$  and  $\pi_i = a_{ii}$ . So, rating methods that accord with Bradley-Terry in the case of a quasi-symmetric results matrix are consistent estimators for the Bradley-Terry model given a Bradley-Terry data generating process, and thus motivations for those rating methods are of interest in the context of this paper. This is especially so as it provides a link to a number of other, sometimes familiar, rating methods.

## 7.1 PageRank (Daniels, 1969)

Daniels (1969) appears to have been the first to document the link between the Bradley-Terry model and what might now be recognised as an undamped PageRank (Page et al., 1999). PageRank has come to be widely known as it formed the basis for the original Google search algorithm. An intuitive explanation for the way it functions is the so-called ‘random surfer’ model. It envisages a (web-)surfer, who is randomly assigned to a node in a directed network. The random surfer then moves randomly to one of the other nodes. With a given probability they may move to any node (teleportation) or alternatively they move to a node to which there is a weighted directed edge from the node where they are currently. The probability of moving to any particular destination node if they do not teleport is set equal to the weight of the edge between the origin node and the destination node divided by the total weight of edges from the origin node. This process continues indefinitely with the proportion of time spent at each node representing the PageRank for that node. What we refer to here as ‘undamped PageRank’ is the algorithm with the teleportation probability set to zero.

In the notation of this paper, we may take the comparison matrix to define the relevant weighted directed network, with  $c_{ij}$  the weight of the directed edge from  $j$  to  $i$ . Define  $D$  as the diagonal matrix of column sums with  $d_{jj} = \sum_k c_{kj}$ . The undamped PageRank rating vector  $\alpha_{PR}$  is the stationary distribution of the Markov chain with column-normalised comparison matrix  $CD^{-1}$  as a left stochastic transition matrix. That is

$$\alpha_{PR} = CD^{-1}\alpha_{PR}.$$

While this rating is perhaps best known from its link to PageRank, it had been previously identified as the ‘total influence’ metric in Pinski and Narin (1976) in the context of bibliometrics. It has been independently axiomatised in Altman and Tennenholtz (2005) and in Slutzki and Volij (2006). More prosaically, such a measure might be motivated in the context of sports competition by the idea of a ‘glory-seeker’ fan, or, as Langville and Meyer (2012, p. 68) terms it, the ‘fair weather’ fan. Consider a fan who begins by selecting a team to support at random. At each step they transfer their allegiance to one of the teams that has beaten the team they previously supported. This decision is made at random in proportion to the number of their defeats that were against each team. Each team is then rated by the proportion of time that the glory-seeker has spent supporting them.

While there is a pleasing intuition to this approach, there are situations where using PageRank is questionable. We present two toy examples that demonstrate just such circumstances. First, consider a five team round-robin tournament between

teams A, B, C, D and E. A beats B, C and D; B beats C, D and E; C beats D and E; D beats E; and E beats A, as represented in Table 1.

	A	B	C	D	E	Wins
A	0	1	1	1	0	3
B	0	0	1	1	1	3
C	0	0	0	1	1	2
D	0	0	0	0	1	1
E	1	0	0	0	0	1

Table 1: Five-team round-robin tournament

Undamped PageRank would rate A and E joint first, because every time the glory-seeker selects team A, they will subsequently select team E, whereas standard round-robin ranking by the number of wins would rate A as joint first and E as joint last. Rubinstein (1980) established axiomatic grounds for why number of wins should be taken as the rating in a round-robin tournament and, beyond that, it is a strong norm in competitive sport, so in this situation PageRank might be deemed inappropriate.

Second, consider three teams F, G, and H. Their strengths are such that we would expect F to beat G in 2/3 of matches, F to beat H in 4/5 of matches, and G to beat H in 2/3 of matches. Now consider two tournaments between these three teams. In the first of these tournaments each team plays each other team 15 times and the proportion of results follow expectations. These results are represented in Table 2(a). In the second tournament the teams win their match-ups in the same proportions, but H plays six times more matches against both F and G; while F and G play each other the same number of times as in the first tournament, with results represented in Table 2(b).

	F	G	H		F	G	H
F	0	10	12	F	0	10	72
G	5	0	10	G	5	0	60
H	3	5	0	H	18	30	0

(a)
(b)

Table 2: Three-team tournaments

It seems clear that based on propensity to win, in either tournament (a) or (b), team F should be ranked higher than team G and team G should be ranked higher

than team H. PageRank meets this requirement for tournament (a), but ranks H highest based on the results of tournament (b).

In both examples, it seems that undamped PageRank focuses too much on the wins of a team, ignoring the losses. In the first example, it was E's win against A that drove its high ranking rather than being balanced by its losses to B, C and D. In the second example, the number of H's wins saw it ranked highest, ignoring its higher number of losses. Therefore one suggestion to address this would be to construct a rating,  $\boldsymbol{\pi} = D^{-1}\boldsymbol{\alpha}_{PR}$ , by scaling the undamped PageRank rating of each competitor by dividing by their number of losses.

$$\boldsymbol{\pi} = D^{-1}\boldsymbol{\alpha}_{PR} = D^{-1}CD^{-1}\boldsymbol{\alpha}_{PR} = D^{-1}C\boldsymbol{\pi},$$

so that  $\boldsymbol{\pi}$  is an eigenvector for  $\hat{C} = D^{-1}C$ .

A vector  $\boldsymbol{\pi}$  is an eigenvector for  $\hat{C} = D^{-1}C$  with an eigenvalue of 1 if and only if

$$\sum_j c_{ij}\pi_j = d_{ii}\pi_i \quad \text{for all } i,$$

but if  $C = AS$  is quasi-symmetric such that  $A$  is a diagonal matrix and  $S$  is symmetric then choosing  $\pi_i = a_{ii}$  yields

$$\sum_j c_{ij}\pi_j = \sum_j a_{ii}s_{ij}a_{jj} = a_{ii} \sum_j s_{ji}a_{jj} = \pi_i \sum_j c_{ji} = d_{ii}\pi_i \quad \text{for all } i,$$

so that the scaled undamped PageRank  $\boldsymbol{\pi} = D^{-1}\boldsymbol{\alpha}_{PR}$  is the diagonal component of a quasi-symmetric matrix. Equivalently it is the Bradley-Terry rating vector in the special case of a quasi-symmetric comparison matrix  $C$  and thus a consistent estimator for the Bradley-Terry rating vector given a Bradley-Terry data-generating process.

In the context of bibliometrics, this rating method was proposed as the ‘influence weight’ measure by Pinski and Narin (1976) and as ‘Scrooge factor’ by Selby (2020), the name we will adopt for the rating for the remainder of this section. In the bibliometric context,  $c_{ij}$  within the comparison matrix represents a citation in journal  $j$  of an article in journal  $i$ . It was motivated by noting that journals are likely to be of different sizes and that one may be interested in determining influence independent of size. The proposal was therefore to normalise the citations received by  $i$  by the citations given by  $i$ . More recently, the ‘Rank Centrality’ algorithm of Negahban et al. (2012) proposes the same estimator applied to ratio matrices, and it is also equivalent to the ‘Luce Spectral Ranking’ of Maystre and Grossglauser (2015) in the  $k = 2$  case. A more detailed discussion of these links was provided by Selby (2020).

As a brief illustration, we return to our examples. In the first example, with results from Table 1, the results do not make up a quasi-symmetric matrix, so that the Bradley-Terry rating and Scrooge factor do not align. As can be seen in Table 3, Bradley-Terry produces the same ranking as the convention of taking the number of wins, since the vector of the number of wins is a sufficient statistic for the Bradley-Terry rating as we showed in Section 2.4. Undamped PageRank and Scrooge factor both rank the teams in the descending order A, B, C, D, but undamped PageRank ranks E as being first equal, whereas Scrooge factor places it third. If we take number of wins to be the correct ranking, then Scrooge Factor gives a more accurate ranking in placing E closer to last equal.

	A	B	C	D	E
Wins	3(1=)	3(1=)	2(3)	1(4=)	1(4=)
Bradley-Terry	7.57(1=)	7.57(1=)	2.75(3)	1.00(4=)	1(4=)
PageRank	1.00(1=)	0.67(3)	0.44(4)	0.33(5)	1(1=)
Scrooge factor	3.00(1)	2.00(2)	0.67(4)	0.33(5)	1(3)

Table 3: Five-team round-robin tournament rating(ranking), with rating of E standardised to 1. PageRank here is undamped.

In the second example, there is no convention such as the number of wins to anchor our methodology on. But given the ratio of wins and losses for each pair, it seems clear that the teams should be ranked in descending order F, G, H. Since both results matrices are quasi-symmetric then Bradley-Terry and Scrooge factor are the same and provide a ranking in the appropriate ordering. As can be seen in Table 4, this is matched by undamped PageRank in the the first of the tournaments where every team plays every other the same number of times, but undamped PageRank disagrees when H has a higher number of match-ups against the other two teams.

## 7.2 Fair bets (Daniels, 1969)

Daniels (1969) introduces an idea referred to as ‘fair scores’. It was elaborated on and cast in the perhaps more intuitive language of ‘bets’ by Moon and Pullman (1970). Both provide interesting discussions of more general approaches. More recently, Slutzki and Volij (2006) provides an excellent summary of the approach, providing two axiomatisations for it, a presentation of a more informal motivation due to Laslier (1997), the link to undamped PageRank, and an argument for why the axiomatisations may lead us to believe that the ‘fair bets’ method is more appropriate for sports tournaments, while the undamped PageRank is more suitable for citation networks.



	F	G	H
Bradley-Terry	4.00(1)	2.00(2)	1(3)
PageRank	1.45(1)	1.36(2)	1(3)
ScroogeFactor	4.00(1)	2.00(2)	1(3)
(a)			
	F	G	H
Bradley-Terry	4.00(1)	2.00(2)	1(3)
PageRank	0.98(2)	0.86(3)	1(1)
ScroogeFactor	4.00(1)	2.00(2)	1(3)
(b)			

Table 4: Three-team tournament rating(ranking) with rating of H standardised to 1. PageRank is undamped.

The first of the axiomatisations shows that the ‘fair bets’ model is the unique ranking derived under the three simultaneous requirements of uniformity, inverse proportionality to losses, and neutrality. Uniformity here requires that if a tournament outcome is balanced in the sense that every competitor has the same number of wins and losses then the competitors must be ranked equally. Inverse proportionality to losses requires that if one begins with a balanced tournament outcome, and then a single competitor’s losses are multiplied by a constant then its rating will be divided by the same constant relative to the other competitors. Neutrality requires that if one begins with a balanced tournament outcome and some new matches are added between two teams where they share the wins equally then competitors will remain equally ranked.

The second of the axiomatisations requires two axioms, consistency between a ranking and its reduced forms, and reciprocity. Reciprocity here requires that, in a two-player tournament, the ratio of the two competitors’ ratings is equal to the ratio of their wins in matches between them, assuming that there are a non-zero number of matches between them. The reduced form condition considers a reduced tournament without a team  $k$ , with the comparison matrix modified to, in effect, reallocate results involving  $k$  so that the comparison matrix is redefined as

$$c_{ij} = \begin{cases} 0 & i = j \\ c_{ij} + \frac{c_{ik}c_{kj}}{\sum_t c_{tk}} & \text{otherwise.} \end{cases}$$

The axiom requires that the relative ratings of two teams in any reduced tournament are equal to their ratio in the full tournament. Consistency requirements of this type are a common feature of axiomatic approaches to ranking (Thomson et al., 1996).

Alternatively, inkeeping with the original presentation of Daniels (1969), suppose one retrospectively wishes to assign a betting scheme to a tournament, where the loser pays to the winner an amount on the result of each match. This is subject to two conditions. First, the amount that is paid to the winner by the loser is a value dependent solely on the strength of the loser. So that if  $i$  beats  $j$  then  $i$  will receive an amount  $\alpha_j^{\text{FB}}$  from  $j$ . Second, the betting scheme is fair. Here ‘fair’ is taken to mean that the wagered amounts will have led to the result that betting on any team throughout the tournament will have a net gain of zero. Then one has the condition that, for all  $i$ ,

$$\sum_j c_{ij} \alpha_j^{\text{FB}} = \sum_j c_{ji} \alpha_i^{\text{FB}},$$

where  $\alpha^{\text{FB}}$  may be taken as a rating vector for the participants, with the intuition being that one would be prepared to wager more on a strong team.

If  $C = AS$  is quasi-symmetric then we have for all  $i$

$$\sum_j a_{ii} s_{ij} \alpha_j^{\text{FB}} = \sum_j a_{jj} s_{ji} \alpha_i^{\text{FB}},$$

so that

$$\sum_j s_{ij} (a_{ii} \alpha_j^{\text{FB}} - a_{jj} \alpha_i^{\text{FB}}) = 0.$$

Thus,  $\alpha_i^{\text{FB}} = a_{ii} = \pi_i$ , and the Fair Bets rating is a consistent estimator for the Bradley-Terry rating vector given a Bradley-Terry generating process.

### 7.3 Wei-Kendall

The rating method introduced in Wei (1952) and Kendall (1955) relies on an iterative application of the comparison matrix. The motivation for such a procedure might be seen by taking the five-team tournament example from Section 7.1. One might argue that ranking D and E equally is unfair as E’s single victory occurred against a top-ranked team A, whereas D gained its only victory against bottom-ranked E. An approach to address this suggested by Wei (1952) is to weight each victory by the rating of the defeated team. The notion of inheriting the wins of a defeated opponent to inform a rating is intuitive enough that it forms the basis for the predominant rating system of the British playground game of conkers (Barrow, 2014). Under the Wei-Kendall method we would begin with a rating vector defined by the sum of wins

$$\mathbf{1} \alpha_{wK} = C \mathbf{e} = \{3, 3, 2, 1, 1\}^T,$$

where  $\mathbf{e}$  is a  $n \times 1$  vector of 1s. Then we assign to each team the sum of the first iteration ratings of each team they have beaten

$$\mathbf{2}\alpha_{WK} = C\mathbf{1}\alpha_{WK} = C^2\mathbf{e} = \{6, 4, 2, 1, 3\}^T.$$

This second iteration measure is sometimes used in chess for tie-breaking, where it is known as the Sonneborn-Berger score (Hooper and Whyld, 1996). But then one might reason that the victories should instead have been weighted by this updated rating. Proceeding in this way for the next five iterations we have Wei-Kendall rating vectors

$$\begin{aligned}\mathbf{3}\alpha_{WK} &= \{7, 6, 4, 3, 6\}^T, \\ \mathbf{4}\alpha_{WK} &= \{13, 13, 9, 6, 7\}^T, \\ \mathbf{5}\alpha_{WK} &= \{28, 22, 13, 7, 13\}^T, \\ \mathbf{6}\alpha_{WK} &= \{42, 33, 20, 13, 28\}^T, \\ \mathbf{7}\alpha_{WK} &= \{66, 61, 41, 28, 42\}^T.\end{aligned}$$

Note that  $E$  continues to be ranked higher than  $D$  and  $C$ .

Generalising, one may define a series of rating vectors

$$\mathbf{k}\alpha_{WK} = C^k\mathbf{e}.$$

It is then natural to consider the limit, but this is clearly not convergent. However, as Moon (1968) notes, since the matrix  $C$  is irreducible then by the Perron-Frobenius theorem (Frobenius, 1912) the rating vector defined by

$$\alpha_{WK} = \lim_{k \rightarrow \infty} \left( \frac{C}{\rho} \right)^k \mathbf{e},$$

where  $\rho$  is the dominant eigenvalue of  $C$ , is convergent, and this normalised limit may be thought of as a rating vector. In the case considered above this gives

$$\alpha_{WK} = \{1.63, 1.38, 0.87, 0.55, 0.95\}^T.$$

The same motivational construct can be applied to give a consistent estimator of the Bradley-Terry rating vector in the case of a Bradley-Terry data-generating process. In both cases, the idea is that we start with an intuitive rating method. It is then noted that the initial wins should not be considered equal and instead those wins should be weighted using the best rating available. This may be done

iteratively defining a rating in the limit. In the case of the Wei-Kendall method, the sum of wins is used as the initial rating. Here, the initial rating is based on the win-loss ratio of each team,  $\hat{C}\mathbf{e} = D^{-1}C\mathbf{e}$ . In the reweighting step in the Wei-Kendall method, the wins are simply weighted by the rating of the losing team. Here, leaning on the intuition of needing to account for losses as well as wins, we scale the vector of rating-weighted wins by the losses for each team. Proceeding in this manner, we define a rating vector

$$\boldsymbol{\pi} = \lim_{k \rightarrow \infty} \hat{C}^k \mathbf{e}.$$

Since the scaled matrix  $\hat{C}$  has unit dominant eigenvalue, then by Perron-Frobenius Theorem the limit is convergent and  $\boldsymbol{\pi}$  is equal to the leading eigenvector of  $\hat{C}$ . If additionally  $\hat{C}$  is quasi-symmetric, which it will be if  $C$  is quasi-symmetric, then this leading eigenvector will be the vector of Bradley-Terry ratings. Thus by applying the same reasoning used to motivate the Wei-Kendall method, but starting with an alternative plausible rating method, win-loss ratio, and accounting for losses as well as wins in the reweighting step, we derive a consistent estimator for the Bradley-Terry rating vector given a Bradley-Terry data-generating process.

## 7.4 Ratings Percentage Index

A rating measure that until recently was prevalent in college sports in North America is the Ratings Percentage Index (RPI). It is commonly defined as

$$\begin{aligned} \text{RPI} = & 25\% \times \text{Win Percentage} \\ & + 50\% \times \text{Opposition's Win Percentage} \\ & + 25\% \times \text{Opposition's Opposition's Win Percentage.} \end{aligned}$$

In the notation of this article, recalling that  $M$  is the matrix of the number of matches, let the matrix  $\hat{M} = [\hat{m}_{ij}]$  with  $\hat{m}_{ij} = m_{ij} / \sum_j m_{ij}$ , so that  $\hat{m}_{ij}$  is the proportion of  $i$ 's matches that are against team  $j$ . Define the win percentage vector  $\mathbf{x} = (x_1, x_2, \dots, x_n)^T$  where  $x_i = w_i / m_i = \sum_j c_{ij} / \sum_j m_{ij}$ , then the RPI rating vector  $\mathbf{RPI} = (\text{RPI}_1, \text{RPI}_2, \dots, \text{RPI}_n)^T$  may be defined as

$$\mathbf{RPI} = 0.25\mathbf{x} + 0.5\hat{M}\mathbf{x} + 0.25\hat{M}^2\mathbf{x}$$

An argument very much like the one in the previous section may be followed to motivate this, that we must consider the strength of opposition in aggregating wins and that we can do this iteratively. In the RPI it is assumed that the previous iterations carry information that should be retained in the overall rating and that three such applications is sufficient.

The choice of win percentage as the initial rating vector and of the proportion of matches as the relevant weighting factor when taking account of the strength of opposition is not unintuitive, but not exclusively so. For example, one might instead take each team’s win-loss ratio as the initial rating. To account for the strength of opposition one might weight wins by the opposition’s rating rather than weighting matches, and then normalise those weighted wins by the number of losses. The 0.25/0.5/0.25 weighting is arbitrary and indeed has been criticised as overweighting the strength of a team’s opposition and for producing perverse incentives (Baker, 2014). In the absence of any clear reason to do otherwise, an equal weighting might instead be applied. This would give an initial rating vector

$$\alpha_1 = \hat{C}e,$$

and considering down to an opposition’s opposition’s strength as in RPI

$$\alpha_3 = \frac{1}{3}\hat{C}^2\alpha_1 + \frac{1}{3}\hat{C}\alpha_1 + \frac{1}{3}\alpha_1 = \frac{1}{3}(\hat{C}^3 + \hat{C}^2 + \hat{C})e.$$

Clearly there is no particular reason to stop after recursively considering two levels of opposition antecedents and so one might more generally consider

$$\pi = \lim_{r \rightarrow \infty} \frac{1}{r} \sum_{k=1}^r \hat{C}^k e.$$

This is the row sum vector of the Cesaro average for  $\hat{C}$  and so

$$\pi = \lim_{k \rightarrow \infty} \hat{C}^k e,$$

giving the same result as in the previous section. And so we have that an RPI-style rating based on win-loss ratios is a consistent estimator for the Bradley-Terry rating vector, given a Bradley-Terry data-generating process.

## 7.5 “Winner stays on” - Barker’s algorithm

It is a convention in some settings, for example pub pool tables, to play on the basis of “winner stays on”, where the winner of any match continues to play the next competitor. While rarely part of an official ranking system, it is intuitive that players who spend more games as “reigning champion” might be considered stronger.

Suppose that one would like to design a “winner stays on” tournament with the property that the long-term proportion of time spent as the “reigning champion”

is directly proportional to their strength. For a countable collection of players, let player  $i$  have a specified strength of  $\pi_i$ . Denoting the indicator that player  $i$  is the reigning champion after the  $k^{th}$  game by  $T_i^k$ , then the design requirement can be specified as

$$\lim_{K \rightarrow \infty} \frac{1}{K} \sum_{k=1}^K T_i^k = \frac{\pi_i}{\sum_{i=1}^n \pi_i}.$$

To make progress with this one must specify the selection probability for the next opponent. Suppose that the current reigning champion is player  $i$ , then the probability their next opponent is chosen to be player  $j$  is denoted  $\phi_{ij}$ . Assuming all games are conditionally independent given the players involved then this construction is a Markov chain on the player identities with transition probability of switching “reigning champion” from player  $i$  to player  $j$  given by  $\phi_{ij}p_{ij}$ .

This setup is akin to the Markov chain Monte Carlo problem of generating samples from a probability distribution only known up to a scaling constant. Satisfying the above requirement for the tournament is equivalent to ensuring that the constructed Markov chain has an invariant distribution that is given by the (normalised) strengths.

There are many ways that the  $p_{ij}$  can be specified to achieve this goal but a natural way is to invoke reversibility by designing the chain so that it satisfies the detailed balance equations. Again, there are many choices of  $p_{ij}$  here but if one wishes the game outcomes to be determined directly by a ratio involving the strengths of the teams, then the natural choice would be to use the acceptance ratio from Barker’s algorithm (Barker,1965)

$$p_{ij} = \frac{\pi_i \phi_{ij}}{\pi_i \phi_{ij} + \pi_j \phi_{ji}}.$$

This can be interpreted as a game being decided by a Bradley-Terry type probability where the player’s strength is biased for that particular game by a multiplicative factor accounting for the imbalance of symmetry for proposing that particular opponent as their next opponent. Hence, the biased strength of player  $i$  is given by  $\pi_i \phi_{ij}$  which is the original strength multiplied by the proposed opponent probability of choosing  $j$  which is independent of the strength of player  $j$ .

Suppose further that the opponent proposal distribution is symmetric, i.e.  $\phi_{ij} = \phi_{ji}$  for all pairs  $i$  and  $j$ . This would be the case if the next opponent was selected uniformly at random in a finite collection of players or if there was some local standardised symmetric proposal centred about the current player’s identity. Then, the

above probability that team  $i$  beats team  $j$  is given by

$$p_{ij} = \frac{\pi_i}{\pi_i + \pi_j}.$$

## 8 Discussion

Faced with “the many routes to the ubiquitous Bradley-Terry model”, two natural questions to ask are: how are these motivations linked? and how is a recognition of these diverse motivations useful in statistical modelling?

In seeking to address the linkages, we discuss the Bradley-Terry model in the context of an exponential family of distributions (Darmois, 1935; Pitman, 1936; Koopman, 1936). This provides a direct link between perhaps the most substantial motivations, those of Sections 2 and Sections 3.1 and 3.2, by showing that the motivations are the specific expression in the Bradley-Terry context of general features of exponential family models. It also provides a direct link to the motivations of Sections 3.3, 5.2 and 5.3, as these are explicitly exponential family models from alternative contexts translated to be applicable to the context under consideration with the Bradley-Terry model.

The usefulness of being able to compare motivations is illustrated in two examples, where the initial motivation for using the model comes from one motivation, but by applying the insight from another motivation we are able to substantiate and advance the method.

### 8.1 The Bradley-Terry model as an exponential family of distributions

Following Geyer (2020), a statistical model is an exponential family of distributions if it has a log-likelihood of the form

$$l(\theta) = \langle y, \theta \rangle - k(\theta),$$

where  $y$  is a vector-valued canonical statistic;  $\theta$  is a vector-valued canonical parameter;  $\langle \cdot, \cdot \rangle$  represents an inner product; and  $k$  is a real-valued function, the cumulant function, which is defined such that  $\nabla k(\theta) = \mathbb{E}_\theta(Y)$ . In seeking a maximum likelihood estimate, the derivative is taken and set equal to zero

$$0 = \nabla l(\theta) = y - \nabla k(\theta) = y - \mathbb{E}_\theta(Y),$$

by the definition of the cumulant function within an exponential family.

In the model discussed here, the likelihood is

$$\prod_{i < j} \binom{m_{ij}}{c_{ij}} p_{ij}^{c_{ij}} (1 - p_{ij})^{m_{ij} - c_{ij}},$$

so that the log-likelihood, up to a constant term, may be taken to be

$$\frac{1}{2} \sum_{i,j} c_{ij} \log \left( \frac{p_{ij}}{1 - p_{ij}} \right) + m_{ij} \log(1 - p_{ij}),$$

and may be rewritten in the form

$$l(\theta) = \frac{1}{2} \sum_{i,j} c_{ij} \theta_{ij} - m_{ij} \log(1 + e^{\theta_{ij}}),$$

where  $\theta$  is the canonical parameter, a vector of length  $n(n - 1)$  corresponding to the directed pairwise comparisons, and with  $\theta_{ij} = \log(p_{ij}/(1 - p_{ij}))$ ; the canonical statistic vector  $y$  takes scaled outcomes  $c_{ij}/2$  as its elements; and the cumulant function is  $k(\theta) = \sum_{i,j} m_{ij} \log(1 + e^{\theta_{ij}})/2$ .

What Geyer et al. (2007) refer to as an *affine canonical submodel* may be parametrised through the linear transformation

$$\theta = a + X\beta,$$

where  $a$  is an offset vector,  $X$  is a design matrix, and  $\beta$  is the canonical parameter for the submodel, giving a log-likelihood of

$$l(\beta) = \langle X^T y, \beta \rangle - k_{SUB}(\beta),$$

where  $k_{SUB}(\beta) = k(a + X\beta)$ , so that this defines a new exponential family with canonical statistic vector  $X^T y$ , canonical parameter vector  $\beta$ , and cumulant function  $k_{SUB}$ .

In the context of the Bradley-Terry model, one may take  $a = 0, \beta = \lambda$ , where  $\lambda$  is the vector of log-strengths  $\lambda_i = \log \pi_i$ , and  $X$  to be the design matrix with the columns representing the  $n$  participants, and the rows representing the  $n(n - 1)$  directed pairwise comparisons. The entry in the row corresponding to a preference for  $i$  over  $j$  has 1 in column  $i$ ,  $-1$  in column  $j$  and zero elsewhere. This gives a log-likelihood

$$l(\lambda) = \frac{1}{2} \sum_{i,j} (c_{ij} - c_{ji}) \lambda_i - \frac{1}{2} \sum_{i,j} m_{ij} \log(1 + e^{\lambda_i - \lambda_j})$$



$$\begin{aligned}
&= \frac{1}{2} \sum_{i,j} (2c_{ij} - m_{ij}) \lambda_i - \frac{1}{2} \sum_{i,j} m_{ij} \log(1 + e^{\lambda_i - \lambda_j}) \\
&= \sum_{i,j} c_{ij} \lambda_i - \frac{1}{2} \sum_{i,j} m_{ij} (\lambda_i + \log(1 + e^{\lambda_i - \lambda_j})).
\end{aligned}$$

Define a vector of wins  $\mathbf{w}$  by  $w_i = \sum_j c_{ij}$ , then

$$l(\lambda) = \sum_i w_i \lambda_i - \frac{1}{2} \sum_{i,j} m_{ij} (\lambda_i + \log(1 + e^{\lambda_i - \lambda_j})),$$

defining an exponential family where the number of wins is the vector-valued canonical statistic and log-strength is the vector-valued canonical parameter. It is a feature of an exponential family of distributions that ‘observed equals expected’, or more precisely that the observed value of the canonical statistic vector equals its expected value under the MLE distribution, that is to say

$$y = \mathbb{E}_{\hat{\theta}}(Y) = \nabla k(\hat{\theta}),$$

which under this affine canonical submodel translates to

$$\begin{aligned}
w_k &= \frac{1}{2} \sum_j m_{kj} \left( 1 + \frac{e^{\lambda_k - \lambda_j}}{1 + e^{\lambda_k - \lambda_j}} \right) - \frac{1}{2} \sum_i m_{ik} \frac{e^{\lambda_i - \lambda_k}}{1 + e^{\lambda_i - \lambda_k}} \\
&= \sum_j m_{kj} \frac{e^{\lambda_k}}{e^{\lambda_k} + e^{\lambda_j}} \quad \text{for all } k,
\end{aligned}$$

noting that  $p_{kj} = e^{\lambda_k} / (e^{\lambda_k} + e^{\lambda_j})$  gives what was referred to as the retrodictive criterion in Sections 3.1 and 3.2.

The motivations based on wins as a sufficient statistic, maximum entropy and maximum likelihood of Sections 2.4, 3.1, and 3.2 may thus be seen as an example of a general fact about exponential families. If one starts with a canonical statistic, then the corresponding affine submodel, if it exists, will be uniquely determined and it will be the maximum entropy and maximum likelihood model subject to the ‘observed equals expected’ constraint on the canonical statistic. As shown in Section 2.4, the requirement to take wins as a sufficient statistic leads directly to the same statistical condition as the other axiomatic motivations presented in Section 2. Thus, a consideration of the Bradley-Terry model as an exponential family of distributions gives a synthesis to the axiomatic and objective function motivations.

## 8.2 Motivation-switching

In this section we present two brief examples to illustrate the usefulness of being able to consider the Bradley-Terry model from a diverse set of motivations. They are characterised by the selection of the model being based on one motivation but then justification and advancement of the methods employed being based on the consideration of other motivations.

### 8.2.1 Sports ranking

The Bradley-Terry model is frequently employed in analysis of sports competitions. Indeed, the original work by Zermelo (1928) was an analysis of competitive chess. Many times the choice of the Bradley-Terry model for sports ranking may be based on its familiarity in the context, or perhaps on an informal version of the definitional simplicity motivations of Sections 3.4 and 3.5. However, a more principled motivation for its application could rest in its status as the unique statistical pairwise comparison model for which the number of wins is a sufficient statistic. Taking the number of wins as the defining ranking measure in balanced sports tournaments is a strong norm and was axiomatised in Rubinstein (1980). It is then natural to generalise this principle by maintaining wins as a sufficient statistic to unbalanced tournaments, where competitors may play differing number of matches against differing opponents of varying strength.

This perspective also provides a natural way to extend the principle to situations where it is points rather than wins that are taken as the determinating data in round-robin tournaments, allowing for result outcomes other than win/loss. Taking points as a sufficient statistic provides a principled motivation to the use of the ties model of Davidson (1970) for unbalanced tournaments in sports where the number of points on offer for a draw is half that for a win, or in employing David Firth’s alt-3 model (Firth, 2022) for soccer, where the norm is 3 points for a win and 1 for a draw.

The geometric motivation of Section 3.3 and the permutation-based model of Section 5.2 may also be applied to extend the situations covered by ranking in a way that is consistent with these well-established sports norms. For example, in athletics — or track and field in North American parlance — it is common for races to be of variable size and to have different entrants at each race. If  $T$ , of size  $n$ , is the total set of competitors, let  $A_k$ , of size  $n_k$ , be the set of competitors in race  $k$ , and  $r_{ik}$  the finishing position for competitor  $i \in T$ . Then we can define a result vector of length  $n$  for race  $k$ ,  $\mathbf{x}_k$ , with value

$$\frac{(n_k + 1)/2 - r_{ik}}{\sqrt{n_k(n_k^2 - 1)/12}}$$

if  $i \in A_k$  and zero otherwise. Consistent with Sections 3.3 and 5.2, a rating vector  $\lambda$  can then be determined by minimising the cumulative squared Euclidean distance

$$\sum_k d(\mathbf{x}_k, \lambda),$$

giving a rating consistent with the Bradley-Terry model in the pairwise comparison case.

### 8.2.2 Comparative Judgement

Comparative Judgement is a form of educational assessment. It creates ratings for a set of items by having judges rank subsets of the items. These comparisons are most commonly pairwise with the Bradley-Terry model being fitted to determine the ratings. Andrich (1978) is often cited in that literature and so it seems reasonable to speculate that the familiarity of the Rasch model in educational assessment may be a significant reason for the model choice. But given the nature of the outcome — the rating of academic work — there might be a legitimate desire to be able to demonstrate the fairness of any method used. While there are not the strong norms around number of wins as a rating measure in this context like in the sports example, the idea of maximising entropy and in that sense minimising the assumptions in the modelling may be attractive as a justification.

The motivations discussed here might also influence some of the practices employed in Comparative Judgement. Often the comparisons are scheduled in order to be able to produce ratings of equivalent reliability with fewer judgements than would be achieved with random scheduling. These adaptive scheduling schemes work by scheduling comparisons between items that are similar in strength so that the information from each pairwise comparison is maximised (Pollitt, 2012a). The Swiss scheduling scheme, where competitors with the same, or as similar as possible, number of wins are scheduled to play each other, is a well-known example. More sophisticated approaches use an online rating that accounts for the observed comparators in order to schedule the next comparisons. These ratings could be the Bradley-Terry ratings, but their computational expense may make them unsuitable. The motivations for the consistent estimators to the Bradley-Terry model discussed in Section 7 may provide grounds for using computationally faster spectral methods for the online rating used for scheduling, even if the final rating is based on fitting the Bradley-Terry model directly, based on the fairness justification.

## 9 Concluding Remarks

In concluding, we highlight four aspects that we hope the reader may take from this work. First is a general interest in the model. Special status is accorded to models and phenomena that become apparent from a diversity of seemingly unrelated perspectives. It is in this spirit, and with a certain affection for the Bradley-Terry model, that this work was initially undertaken. Undoubtedly some of the motivations presented here carry more weight than others. Being the unique solution to maximising entropy subject to the retrodictive criterion will be a relevant motivation in more scenarios than being a readily hypothesised model for a sudden death contest on a difference of  $r$  points. Nevertheless, the number and diversity of motivations is suggestive of the applicability and attractiveness of the model, and lays the basis for its use in a wide variety of contexts.

Second is an appreciation for the importance of model motivations. Often the motivation for using a particular model is a pragmatic one based on goodness of fit, predictive ability, computational ease or simply familiarity to the practitioner. However, there can be scenarios where a more principled motivation matters. This is likely to be the case where there are issues of fairness involved. Such scenarios are not uncommon where the output of a model is a rating, as with the examples of official sports ranking and educational assessment. The ‘wins as a sufficient statistic’ and ‘maximum entropy’ motivations may be particularly pertinent in those scenarios. Third is an appreciation for how understanding different motivations can aid in modelling practice, as illustrated with the examples of Sections 8.2.1 and 8.2.2. The setting of the Bradley-Terry model in the context of an exponential family of distributions, and the directly related motivations, may be particularly useful in advancing or expanding its application. Finally, we hope the work may be useful in devising material for engaging wider audiences. Some of the subject matter that the Bradley-Terry model relates to — ratings in general, especially when applied to fields like sports — are ones that can be of great interest to student and outside audiences, and so it is to be hoped that this work can assist in that engagement.

## References

- Altman, A. and Tennenholtz, M. (2005). Ranking systems: the Pagerank axioms. In *Proceedings of the 6th ACM conference on Electronic Commerce*, pages 1–8.
- Andrich, D. (1978). Relationships between the Thurstone and Rasch approaches to item scaling. *Applied Psychological Measurement*, 2(3):451–462.

Audley, R. (1960). A stochastic model for individual choice behavior. *Psychological Review*, 67(1):1.

Baker, D. (2014). Death to the RPI. <https://www.theonlycolors.com/2014/2/26/5444872/death-t> accessed March 31, 2021.

Barrow, M. (2014). Conkers. A favourite children's game. <http://projectbritain.com/conkers.html>, accessed May 18, 2022.

Block, H. D. and Marschak, J. (1960). Random orderings and stochastic theories of responses. In *Contributions to probability and statistics*. Stanford University Press Stanford, CA.

Bradley, R. A. (1965). Another interpretation of a model for paired comparisons. *Psychometrika*, 30(3):315–318.

Bradley, R. A. (1976). Science, statistics, and paired comparisons. *Biometrics*, 32(2):213–239.

Bradley, R. A. and Terry, M. E. (1952). Rank analysis of incomplete block designs: I. The method of paired comparisons. *Biometrika*, 39(3/4):324–345.

Bühlmann, H. and Huber, P. J. (1963). Pairwise comparison and ranking in tournaments. *The Annals of Mathematical Statistics*, 34(2):501–510.

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

Caussinus, H. (1965). Contribution à l'analyse statistique des tableaux de corrélation. In *Annales de la Faculté des sciences de Toulouse: Mathématiques*, volume 29, pages 77–183.

Cox, D. R. (1972). Regression models and life-tables. *Journal of the Royal Statistical Society: Series B*, 34(2):187–202.

Daniels, H. E. (1969). Round-robin tournament scores. *Biometrika*, 56(2):295–299.

Darmois, G. (1935). Sur les lois de probabilité à estimation exhaustive. *CR Acad. Sci. Paris*, 260(1265):85.

David, H. A. (1988). *The method of paired comparisons*. Charles Griffin, London, second edition.

- Davidson, R. R. (1970). On extending the Bradley-Terry model to accommodate ties in paired comparison experiments. *Journal of the American Statistical Association*, 65(329):317–328.
- Davidson, R. R. and Farquhar, P. H. (1976). A bibliography on the method of paired comparisons. *Biometrics*, pages 241–252.
- Dittrich, R., Hatzinger, R., and Katzenbeisser, W. (1998). Modelling the effect of subject-specific covariates in paired comparison studies with an application to university rankings. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 47(4):511–525.
- Elo, A. E. (1978). *The rating of chessplayers, past and present*. Arco Pub.
- Fienberg, S. E. and Larntz, K. (1976). Log linear representation for paired and multiple comparisons models. *Biometrika*, 63(2):245–254.
- Firth, D. (2022). *Maths. Football. That’s all*. <https://alt3.uk/>, accessed September 16, 2022.
- Fisher, R. A. (1953). Dispersion on a sphere. *Proceedings of the Royal Society of London. Series A. Mathematical and Physical Sciences*, 217(1130):295–305.
- Fisher, R. A. and Tippett, L. H. C. (1928). Limiting forms of the frequency distribution of the largest or smallest member of a sample. In *Mathematical Proceedings of the Cambridge Philosophical Society*, volume 24, pages 180–190. Cambridge University Press.
- Ford Jr, L. R. (1957). Solution of a ranking problem from binary comparisons. *The American Mathematical Monthly*, 64(8P2):28–33.
- Frobenius, G. (1912). Über matrizen aus nicht negativen elementen. *Sitzungsberichte der Königl. Akademie der Wissenschaften, Berlin*, 23:456 – 477.
- Geyer, C. J. (2020). Stat 8054 lecture notes: Exponential families. <https://www.stat.umn.edu/geyer/8054/notes/expfam.html>, accessed October 4, 2022.
- Geyer, C. J., Wagenius, S., and Shaw, R. G. (2007). Aster models for life history analysis. *Biometrika*, 94(2):415–426.
- Glickman, M. E. (2013). Introductory note to 1928 (= 1929). In *Ernst Zermelo-collected works/Gesammelte Werke II*, pages 616–671. Springer.

- Gnedenko, B. (1943). Sur la distribution limite du terme maximum d'une serie aleatoire. *Annals of Mathematics*, pages 423–453.
- Good, I. J. (1955). On the marking of chess-players. *The Mathematical Gazette*, 39(330):292–296.
- Good, I. J. et al. (1963). Maximum entropy for hypothesis formulation, especially for multidimensional contingency tables. *The Annals of Mathematical Statistics*, 34(3):911–934.
- Gumbel, E. J. (1958). *Statistics of extremes*. Columbia University Press.
- Henery, R. J. (1986). Interpretation of average ranks. *Biometrika*, 73(1):224–227.
- Hoff, P. D., Raftery, A. E., and Handcock, M. S. (2002). Latent space approaches to social network analysis. *Journal of the American Statistical Association*, 97(460):1090–1098.
- Hooper, D. and Whyld, K. (1996). *The Oxford companion to chess*. Oxford University Press, USA.
- Jaynes, E. T. (1957). Information theory and statistical mechanics. *Physical Review*, 106(4):620.
- Jech, T. (1983). The ranking of incomplete tournaments: A mathematician's guide to popular sports. *The American Mathematical Monthly*, 90(4):246–266.
- Joe, H. (1988). Majorization, entropy and paired comparisons. *The Annals of Statistics*, pages 915–925.
- Kelly, F. P. (1979). *Reversibility and stochastic networks*. Cambridge University Press.
- Kendall, M. G. (1955). Further contributions to the theory of paired comparisons. *Biometrics*, 11(1):43–62.
- Kolmogorov, A. (1936). Zur Theorie der Markoffschen Ketten. *Mathematische Annalen*, 112(1):155–160.
- Koopman, B. O. (1936). On distributions admitting a sufficient statistic. *Transactions of the American Mathematical society*, 39(3):399–409.

- Krivitsky, P. N., Handcock, M. S., Raftery, A. E., and Hoff, P. D. (2009). Representing degree distributions, clustering, and homophily in social networks with latent cluster random effects models. *Social Networks*, 31(3):204–213.
- Langville, A. N. and Meyer, C. D. (2012). *Who’s# 1?: the science of rating and ranking*. Princeton University Press.
- Laslier, J.-F. (1997). *Tournament solutions and majority voting*, volume 7. Springer Verlag.
- Lee, J. Y., Green, P. J., and Ryan, L. M. (2017). On the” poisson trick” and its extensions for fitting multinomial regression models. *arXiv preprint arXiv:1707.08538*.
- Lehmann, E. L. (1953). The power of rank tests. *The Annals of Mathematical Statistics*, pages 23–43.
- Luce, R. and Suppes, P. (1965). Preference, utility, and subjective utility. *Handbook of Mathematical Psychology, III, New York: Wiley*, pages 249–409.
- Luce, R. D. (1959). *Individual choice behavior*. Wiley, New York.
- Luce, R. D., Ng, C., Marley, A., and Aczél, J. (2008). Utility of gambling I: entropy modified linear weighted utility. *Economic Theory*, 36(1):1–33.
- Mallows, C. L. (1957). Non-null ranking models. i. *Biometrika*, 44(1/2):114–130.
- Maystre, L. and Grossglauser, M. (2015). Fast and accurate inference of Plackett-Luce models. In *Advances in Neural Information Processing Systems*, pages 172–180.
- McCullagh, P. (1993). Models on spheres and models for permutations. In *Probability models and statistical analyses for ranking data*, pages 278–283. Springer.
- McCullagh, P. and Nelder, J. A. (1989). *Generalized linear models*. Chapman and Hall.
- Merrick, J. R., Van Dorp, J. R., Mazzuchi, T., Harrauld, J. R., Spahn, J. E., and Grabowski, M. (2002). The Prince William Sound risk assessment. *Interfaces*, 32(6):25–40.
- Moon, J. W. (1968). *Topics on tournaments in graph theory*. Holt, Rinehart and Winston.



- Moon, J. W. and Pullman, N. (1970). On generalized tournament matrices. *SIAM Review*, 12(3):384–399.
- Mosteller, F. (1951). Remarks on the method of paired comparisons. I. The least squares solution assuming equal standard deviations and equal correlations. *Psychometrika*, 16:6–9.
- Negahban, S., Oh, S., and Shah, D. (2012). Iterative ranking from pair-wise comparisons. *Advances in Neural Information Processing Systems*, 25:2474–2482.
- Oberfeld, D., Hecht, H., Allendorf, U., and Wickelmaier, F. (2009). Ambient lighting modifies the flavor of wine. *Journal of Sensory Studies*, 24(6):797–832.
- O’Donovan, P., Libeks, J., Agarwala, A., and Hertzmann, A. (2014). Exploratory font selection using crowdsourced attributes. *ACM Transactions on Graphics (TOG)*, 33(4):92.
- Page, L., Brin, S., Motwani, R., and Winograd, T. (1999). The Pagerank citation ranking: Bringing order to the web. Technical report, Stanford InfoLab.
- Pinski, G. and Narin, F. (1976). Citation influence for journal aggregates of scientific publications: Theory, with application to the literature of physics. *Information Processing & Management*, 12(5):297–312.
- Pitman, E. J. G. (1936). Sufficient statistics and intrinsic accuracy. In *Mathematical Proceedings of the Cambridge Philosophical Society*, volume 32, pages 567–579. Cambridge University Press.
- Pollitt, A. (2012a). Comparative judgement for assessment. *International Journal of Technology and Design Education*, 22(2):157–170.
- Pollitt, A. (2012b). The method of adaptive comparative judgement. *Assessment in Education: principles, policy & practice*, 19(3):281–300.
- Rasch, G. (1960). Probabilistic models for some intelligence and attainment tests. *Danish institute for Educational Research*.
- Rasch, G. (1961). On general laws and the meaning of measurement in psychology. In *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability*, volume 4, pages 321–333.
- Rubinstein, A. (1980). Ranking the participants in a tournament. *SIAM Journal on Applied Mathematics*, 38(1):108–111.

- Selby, D. (2020). *Statistical modelling of citation networks, research influence and journal prestige*. PhD thesis, Department of Statistics, University of Warwick. Unpublished.
- Seymour, R. G., Sirl, D., Preston, S. P., Dryden, I. L., Ellis, M. J., Perrat, B., and Goulding, J. (2022). The bayesian spatial bradley–terry model: Urban deprivation modelling in tanzania. *Journal of the Royal Statistical Society Series C: Applied Statistics*, 71(2):288–308.
- Shannon, C. E. (1948). A mathematical theory of communication. *Bell system technical journal*, 27(3):379–423.
- Slutzki, G. and Volij, O. (2006). Scoring of web pages and tournaments—axiomatizations. *Social Choice and Welfare*, 26(1):75–92.
- Stern, H. (1990). A continuum of paired comparisons models. *Biometrika*, 77(2):265–273.
- Stern, H. (1992). Are all linear paired comparison models empirically equivalent? *Mathematical Social Sciences*, 23(1):103–117.
- Stigler, S. M. (1994). Citation patterns in the journals of statistics and probability. *Statistical Science*, pages 94–108.
- Stirzaker, D. (1999). *Probability and random variables: A beginner’s guide*. Cambridge University Press.
- Stob, M. (1984). A supplement to “A mathematician’s guide to popular sports”. *The American Mathematical Monthly*, 91(5):277–282.
- Strang, A., Abbott, K. C., and Thomas, P. J. (2022). The network hhd: Quantifying cyclic competition in trait-performance models of tournaments. *SIAM Review*, 64(2):360–391.
- Stuart-Fox, D. M., Firth, D., Moussalli, A., and Whiting, M. J. (2006). Multiple signals in chameleon contests: designing and analysing animal contests as a tournament. *Animal Behaviour*, 71(6):1263–1271.
- Su, Y. and Zhou, M. (2006). On a connection between the Bradley-Terry model and the Cox proportional hazards model. *Statistics & Probability Letters*, 76(7):698–702.

- Thompson, W. and Singh, J. (1967). The use of limit theorems in paired comparison model building. *Psychometrika*, 32(3):255–264.
- Thomson, W. et al. (1996). Consistent allocation rules. Technical report, University of Rochester-Center for Economic Research (RCER).
- Thurstone, L. L. (1927a). A law of comparative judgment. *Psychological Review*, 34(4):273.
- Thurstone, L. L. (1927b). The method of paired comparisons for social values. *The Journal of Abnormal and Social Psychology*, 21(4):384.
- Thurstone, L. L. (1927c). Psychophysical analysis. *The American Journal of Psychology*, 38(3):368–389.
- Vojnović, M. (2015). *Contest theory: Incentive mechanisms and ranking methods*. Cambridge University Press.
- von Mises, R. (1918). Über die “Ganzzahligkeit” der Atomgewichte und verwandte Fragen. *Physikalische Zeitschrift*, 19:490–500.
- Wei, T.-H. (1952). *Algebraic foundations of ranking theory*. PhD thesis, University of Cambridge.
- Wobus, J. (2007). Krach ratings. <http://sports.vaporio.com/krach.html>, accessed October 4, 2022.
- Wu, W., Niezink, N., and Junker, B. (2022). A diagnostic framework for the bradley–terry model. *Journal of the Royal Statistical Society Series A: Statistics in Society*, 185(Supplement\_2):S461–S484.
- Yellot, J. (1977). The relationship between Luce’s choice axiom, Thurstone’s theory of comparative judgment, and the double exponential distribution. *Journal of Mathematical Psychology*, 15(2):109–144.
- Zermelo, E. (1928). Die Berechnung der Turnier-ergebnisse als ein Maximumproblem der Wahrscheinlichkeitsrechnung. *Mathematische Zeitschrift*, 29(1):436–460.