PROCEEDINGS A

rspa.royalsocietypublishing.org

Research

Article submitted to journal

Subject Areas:

xxxxx, xxxxx, xxxx

Keywords:

ranking, paired comparisons, Bradley-Terry model, dominance hierarchy

Author for correspondence:

Mark Newman

e-mail: mejn@umich.edu

Ranking with multiple types of pairwise comparisons

M. E. J. Newman

Center for the Study of Complex Systems, University of Michigan, Ann Arbor, MI 48109. U.S.A.

The task of ranking individuals or teams, based on a set of comparisons between pairs, arises in various contexts, including sporting competitions and the analysis of dominance hierarchies among animals and humans. Given data on which competitors beat which others, the challenge is to rank the competitors from best to worst. Here we study the problem of computing rankings when there are multiple, potentially conflicting types of comparison, such as multiple types of dominance behaviours among animals. We assume that we do not know *a priori* what information each behaviour conveys about the ranking, or even whether they convey any information at all. Nonetheless we show that it is possible to compute a ranking in this situation and present a fast method for doing so, based on a combination of an expectation-maximization algorithm and a modified Bradley-Terry model. We give a selection of example applications to both animal and human competition.

THE ROYAL SOCIETY
PUBLISHING

1. Introduction

Suppose a set of individuals or teams compete with one another in pairwise fashion, as in a sporting league or chess tournament. The outcome of each game is recorded and your task is to rank the players or teams from best to worst based on those outcomes. Similar ranking problems also arise in paired comparison tests in consumer research and behavioural psychology, and in dominance hierarchies in animals and humans, where individuals compete in pairwise fashion to establish dominance and the challenge is to estimate the hierarchy of individuals from their observed interactions.

Ranking problems of this kind have been widely studied for over a century [1–5]. They can be non-trivial because the outcomes of competitions are not always consistent—better players sometimes lose and worse players win—but excellent methods exist for computing rankings nonetheless. Here we study the more complex "multimodal" problem of computing rankings when there is more than one type or mode of competition between individuals. Many animals, for instance, compete in a range of different ways. Mountain goats attack opponents with their horns but also employ a type of rushing attack [6]. Spotted hyenas also use a rushing attack, or they may bite their opponents [7]. In traditional ranking analyses one treats these modes as equivalent, but different modes may contribute more or less to establishing dominance and by exploiting this variation one can perform a more fine-grained analysis that reveals things not readily accessible via standard methods.

Ranking analyses are frequently couched in the language of wins and losses, but there are cases where this framing does not apply. In observations of interactions among elementary school children, for instance, there is often a well established hierarchy but many interactions are not overtly competitive [8]. One child directing the actions of another, for instance, is commonly observed and non-competitive but is nonetheless believed to be indicative of dominance. In cases like this merely performing the action is the equivalent of a win. Similar situations also arise in animal behaviour, where dominance displays without physical competition are often preferred by both parties to avoid injury. American bison, for example, do fight one another, but they also employ a range of non-contact threats including a "nod threat" and a "broadside threat" [9]. Dominance is signalled merely by performing the action.

At the same time, not all actions necessarily indicate dominance: some may indicate subordination. In the case of schoolchildren, for instance, actions like directing behaviour indicate dominance, but actions like following, copying, or watching typically indicate subordination [8]. Both dominant and subordinate actions may be useful for inferring hierarchy, but we need to know which is which in order to analyse them appropriately.

In the work presented here, we explicitly allow for the possibility that different actions can play different roles in a given setting, but we do not assume that we know the role or valence of any action in advance. It turns out, however, that one can infer these parameters automatically from observational data, and once we know the nature of the interactions we can use this information to make an estimate of the complete ranking of individuals or teams, weighting each interaction type appropriately. Interactions that are strongly correlated with dominance carry heavy weight; those that are strongly correlated with subordination do so as well, but with an opposite sign. And interactions that are only weakly correlated are given low weight. In this way, we can construct a ranking that correctly allows for the different ways in which individuals interact.

A number of previous authors have studied related problems. Our multimodal ranking is in some ways reminiscent of the "multivariate" comparisons studied by Davidson and Bradley and others [10–12]. These authors consider problems of consumer choice in which items or products are rated by consumers on several criteria, such as a car rated on price, size, fuel economy, and so forth. A crucial difference between these studies and ours is that our goal is to infer only a single, one-dimensional ranking of items or individuals, whereas the multivariate approach produces multiple rankings, one for each criterion. While the multivariate approach can be appropriate in settings such as consumer choice, the approach of the present paper is more appropriate for applications to dominance hierarchies, sport, and other competitions where we either desire or believe axiomatically that the individuals or teams exist on a single rank scale. Perhaps a closer parallel to our approach is to be found in the literature on rank aggregation methods [13]. These are general methods for taking multiple, possibly conflicting rankings of the same set of items or individuals and combining them to generate a single consensus ranking. (Note that the phrase "rank aggregation" is also sometimes used to describe the unimodal ranking problem [13,14], so one must be careful when reading this literature.) Rank aggregation is used for instance to combine rankings of web pages for output by web search engines [15]. Although they are based on different principles to our approach, such as numerical minimization of the distance between rankings, rank aggregation methods could be applied to the types of problems we consider by generating separate rankings for each criterion or mode and then aggregating them. A related approach has been taken by Pritikin [16], who also constructs different ranking scales for different criteria but then performs a factor analysis to construct a consensus measure of merit. Like the multivariate approach, these methods contrast philosophically with our assumption that there exists only a single ranking, but the outputs could give similar results in some cases.

More broadly, the multimodal ranking problem belongs to the larger class of studies of "multiplex networks" [17,18]. The pattern of comparisons or interactions can be thought of as a network in which the nodes are the individuals or teams and the directed edges between them represent interactions, with the direction pointing from the winner or instigator of an interaction to the loser or recipient. The multiple modes of interaction can be thought of as multiple types of edges between the same set of participants, and such networks—known as multiplex networks in the literature—have been the subject of a substantial amount of work in recent years [18–20], including work on ranking problems [21,22], although the particular problem addressed here does not seem to have received attention.

2. The model

Suppose we have a population of N individuals or teams labelled by $u = 1 \dots N$ and a sequence of contests or interactions between them. Let there be M interactions in total denoted by $r = 1 \dots M$, let u_r be the individual who wins the rth interaction (in cases where there is a winner) or who instigates the interaction (in cases where there is no winner). Similarly let v_r be the loser or recipient of the rth interaction. For simplicity we assume that there are no ties or draws, although the approach we describe could be generalized to the case of ties using standard methods [23,24].

For each interaction we assume that there is a dominant individual and a subordinate individual. It is central to our treatment, however, that the dominant individual need not be the winner or instigator of the interaction: some games may be won by the weaker player and some actions may be instigated by a subordinate individual. Moreover, we explicitly allow for the possibility that the dominant individual in a pair may change from one interaction to another. In some pairs one individual may always be dominant, but in others the roles may vary. We define a stance variable σ_r which indicates which individual is dominant during interaction r: the stance is 1 if the winner u_r is the dominant individual and 0 if the loser v_r is dominant. (To simplify the discussion we will henceforth refer to u_r as the "winner", but this should be taken to include the instigator in cases where merely performing an action is equivalent to a win, and similarly for losses.)

The distinction here between the *dominant individual* and the *winner* is crucial. In the conventional theory of unimodal paired comparisons no distinction is made between these: the winner is dominant by definition. In the multimodal case, however, this approach is insufficient because it fails to separate the properties of the individuals from the properties of the interactions. It would not allow, for instance, for a situation in which one type of interaction is always instigated by the dominant individual and another by the subordinate individual. The model proposed here allows for the possibility that different interaction types may have different probabilities of being instigated or won, even by the same individual.

We assume the stance σ_r is independent for each interaction and model it using a standard Bradley-Terry model of dominance [1–3]. We define a score or ranking s_u for each individual u and we assume that the probability p_{uv} of individual u dominating over individual v on any particular interaction is a function $f(s_u - s_v)$ of the difference of their scores. In the Bradley-Terry model this function is chosen to be the logistic function $f(s) = 1/(1 + e^{-s})$, which means that

$$p_{uv} = \frac{e^{s_u}}{e^{s_u} + e^{s_v}}. \tag{2.1}$$

These probabilities are invariant under a uniform shift of all the scores s_u by an additive constant, so one commonly introduces some normalization or standardization to fix the origin of the score scale. Here we do this by fixing the average score to be zero, so that $s = 0$ indicates an average individual who is equally likely to be dominant or subordinate.

For convenience, one also often introduces the shorthand $\lambda_u = e^{s_u}$, so that

$$p_{uv} = \frac{\lambda_u}{\lambda_u + \lambda_v}, \tag{2.2}$$

and we will do that here. Following Zermelo [1] we refer to λ_u as a strength parameter or simply a strength. Note that for the average individual with $s = 0$ we have $\lambda = e^0 = 1$, so the probability p_0 of an individual with strength parameter λ dominating against the average individual is

$$p_0 = \frac{\lambda}{\lambda + 1}, \tag{2.3}$$

and hence $\lambda = p_0/(1 - p_0)$. Thus λ has a simple interpretation: it is the odds of dominating against the average individual.

The winner of an interaction need not, as we have said, be the dominant individual. Sometimes the winner is the subordinate individual and the frequency with which this happens may depend on the type of the interaction. Let there be T interaction types, labelled by $t = 1 \dots T$. We define another parameter q_t, which we call the *valence probability*, equal to the probability that the currently dominant individual wins an interaction of type t (and $1 - q_t$ is the probability that the subordinate individual wins).

Given these definitions, we can now write down an expression for the likelihood of occurrence of a particular sequence of wins and losses. Consider an interaction r with winner u, loser v, and type t. The probability of the stance taking value $\sigma_r = 1$ is equal to p_{uv} and the probability that $\sigma = 0$ is $1 - p_{uv}$. In general,

$$P(\sigma_r|\lambda_u, \lambda_v) = p_{uv}^{\sigma_r}(1 - p_{uv})^{1-\sigma_r} = \frac{\lambda_u^{\sigma_r}\lambda_v^{1-\sigma_r}}{\lambda_u + \lambda_v}. \tag{2.4}$$

Given the value of σ_r we can calculate the probability that individual u did indeed win: if $\sigma_r = 1$, so that u is dominant, then the probability is q_t; if $\sigma_r = 0$, so that u is subordinate, then the probability is $1 - q_t$. Denoting the observation data for interaction r by x_r, these probabilities can be compactly combined as

$$P(x_r|\sigma_r, q_t) = q_t^{\sigma_r}(1 - q_t)^{1-\sigma_r}. \tag{2.5}$$

Then the probability that the stance is σ_r and u wins is

$$\begin{aligned} P(x_r, \sigma_r|\lambda_u, \lambda_v, q_t) &= P(x_r|\sigma_r, q_t)P(\sigma_r|\lambda_u, \lambda_v) = (p_{uv}q_t)^{\sigma_r}\left[(1 - p_{uv})(1 - q_t)\right]^{1-\sigma_r} \\ &= \frac{(\lambda_u q_t)^{\sigma_r}[\lambda_v(1 - q_t)]^{1-\sigma_r}}{\lambda_u + \lambda_v}. \end{aligned} \tag{2.6}$$

Finally, under the assumption that interactions r are independent of one another (conditioned on the values of λ_u and q_t), the likelihood of the entire set of M observations is the product of this expression over all r, which gives

$$P(x,\sigma|\lambda,q)=\prod_{r=1}^{M}\frac{(\lambda_{u_r}q_{t_r})^{\sigma_r}[\lambda_{v_r}(1-q_{t_r})]^{1-\sigma_r}}{\lambda_{u_r}+\lambda_{v_r}}, \tag{2.7}$$

where the unsubscripted variables x, σ, λ, and q indicate the complete sets of data and parameters. Our goal is to use Eq. (2.7) to estimate the values of λ_u and q_t for all individuals u and all interaction types t.

3. EM algorithm

One might imagine that the next step in the calculation would be to sum over the variables σ_r in Eq. (2.7), which can be done easily to give an expression for $P(x|\lambda,q)$ thus:

$$P(x|\lambda,q)=\sum_{\sigma}P(x,\sigma|\lambda,q)=\prod_{r=1}^{M}\frac{\lambda_{u_r}q_{t_r}+\lambda_{v_r}(1-q_{t_r})}{\lambda_{u_r}+\lambda_{v_r}}. \tag{3.1}$$

Then we could maximize this expression by differentiating with respect to λ and q to find maximum-likelihood estimates of the parameters, or alternatively maximize its logarithm, which is equivalent and usually simpler. Though the derivatives can be done, however, they yield a complicated set of implicit equations with no easy solution. Instead, therefore, we adopt a different approach, making use of an expectation-maximization (EM) algorithm.

Instead of maximizing $\log P(x|\lambda,q)$ directly, we apply Jensen's inequality, which says for any set of non-negative quantities z_i and weights π_i satisfying $\sum_i \pi_i=1$ that $\log\sum_i z_i \geq \sum_i \pi_i \log(z_i/\pi_i)$. Applied to the present case, this gives

$$\log P(x|\lambda,q)=\log\sum_{\sigma}P(x,\sigma|\lambda,q)\geq\sum_{\sigma}\pi(\sigma)\log\frac{P(x,\sigma|\lambda,q)}{\pi(\sigma)}, \tag{3.2}$$

where $\pi(\sigma)$ is any probability distribution over the set σ of stance variables satisfying $\sum_\sigma \pi(\sigma)=1$. The exact equality is achieved—and hence the right-hand side of (3.2) maximized—when

$$\pi(\sigma)=\frac{P(x,\sigma|\lambda,q)}{\sum_\sigma P(x,\sigma|\lambda,q)}=\frac{\prod_r(\lambda_{u_r}q_{t_r})^{\sigma_r}[\lambda_{v_r}(1-q_{t_r})]^{1-\sigma_r}}{\prod_r[\lambda_{u_r}q_{t_r}+\lambda_{v_r}(1-q_{t_r})]}=\prod_r \pi_r^{\sigma_r}(1-\pi_r)^{1-\sigma_r}, \tag{3.3}$$

where

$$\pi_r=\frac{\lambda_{u_r}q_{t_r}}{\lambda_{u_r}q_{t_r}+\lambda_{v_r}(1-q_{t_r})} \tag{3.4}$$

and we have used Eqs. (2.7) and (3.1). The quantity π_r can be interpreted as the posterior likelihood that u_r is the dominant participant in interaction r.

Since the choice (3.3) maximizes the right-hand side of (3.2) and simultaneously makes the two sides equal, a further maximization with respect to λ and q will then achieve our goal of finding the maximum-likelihood values of these parameters. Equivalently, a double maximization of the right-hand side with respect to both $\pi(\sigma)$ and the parameters λ, q will achieve the same goal. In an EM algorithm we perform this double maximization by alternately and repeatedly maximizing with respect to $\pi(\sigma)$ and with respect to λ, q until convergence is reached. The first we do using Eqs. (3.3) and (3.4) and the second by differentiation as follows.

Substituting from Eqs. (2.7) and (3.3) into (3.2) and neglecting terms that do not depend on λ, q, we find that

$$\begin{aligned}
&\sum_\sigma \pi(\sigma) \log P(x,\sigma|\lambda,q) \\
&= \sum_\sigma \left[\prod_r \pi_r^{\sigma_r}(1-\pi_r)^{1-\sigma_r}\right] \sum_r \Bigl[\sigma_r \log(\lambda_{u_r} q_{t_r}) + (1-\sigma_r)\log\bigl[\lambda_{v_r}(1-q_{t_r})\bigr] - \log(\lambda_{u_r}+\lambda_{v_r})\Bigr] \\
&= \sum_r \Bigl[\pi_r \log(\lambda_{u_r} q_{t_r}) + (1-\pi_r)\log\bigl[\lambda_{v_r}(1-q_{t_r})\bigr] - \log(\lambda_{u_r}+\lambda_{v_r})\Bigr] \\
&= \sum_r \sum_{uv} \delta_{u_r u}\delta_{v_r v}\bigl[\pi_r \log\lambda_u + (1-\pi_r)\log\lambda_v - \log(\lambda_u+\lambda_v)\bigr] \\
&\qquad\qquad + \sum_r \sum_t \delta_{t_r t}\bigl[\pi_r \log q_t + (1-\pi_r)\log(1-q_t)\bigr],
\end{aligned} \tag{3.5}$$

where δ_{ab} is the Kronecker delta. Now we differentiate this expression to calculate the maximum-likelihood estimates of the parameters. Differentiating with respect to q_t gives the straightforward result

$$q_t = \frac{\sum_r \delta_{t_r t}\pi_r}{\sum_r \delta_{t_r t}}. \tag{3.6}$$

Calculating the λ parameters is a little more complicated. Differentiating (3.5) with respect to λ_i and setting the result to zero, we get

$$\sum_r \frac{\delta_{u_r i}\pi_r + \delta_{v_r i}(1-\pi_r)}{\lambda_i} - \sum_r \left[\frac{\delta_{u_r i}}{\lambda_i + \lambda_{v_r}} + \frac{\delta_{v_r i}}{\lambda_{u_r} + \lambda_i}\right] = 0, \tag{3.7}$$

which can be rearranged into the form

$$\lambda_i = \frac{\sum_r \bigl[\pi_r \delta_{u_r i} + (1-\pi_r)\delta_{v_r i}\bigr]}{\sum_u A_{iu}/(\lambda_i + \lambda_u)}, \tag{3.8}$$

where

$$A_{ij} = \sum_r \bigl[\delta_{u_r i}\delta_{v_r j} + \delta_{u_r j}\delta_{v_r i}\bigr] \tag{3.9}$$

is the total number of times that i and j interact. Equation (3.8) can be solved for λ_i by simple iteration starting from any convenient set of initial values $\lambda_i > 0$. This iterative procedure can be thought of as a variation on the well-known algorithm of Zermelo [1] for the traditional Bradley-Terry model with just a single mode of interaction. (For small problems with only a few individuals or teams it may be faster to solve (3.7) directly by Newton's method or Fisher scoring, but the speed advantage is relatively small and for larger instances the simple iterative approach is faster because it does not require a matrix inversion. On balance, we recommend the iterative approach in most cases.)

The full algorithm for computing a ranking now consists of the following steps:

(i) Choose initial values for the parameters λ_u and q_t for all individuals u and interaction types t. These values can for instance be chosen at random.
(ii) Compute the quantities π_r from Eq. (3.4).
(iii) Compute new estimates of the valence probabilities q_t from Eq. (3.6).
(iv) Compute new estimates of the strengths λ_u by iterating Eq. (3.8) to convergence.
(v) Normalize the λ_u so that the average individual has rank score zero, which is equivalent to dividing by the geometric mean of the λ_u thus:

$$\lambda_u \leftarrow \frac{\lambda_u}{\left(\prod_{v=1}^N \lambda_v\right)^{1/N}}. \tag{3.10}$$

(vi) Repeat from step 2 until overall convergence is achieved.

Note that, by contrast with EM algorithms in some other applications, this algorithm is guaranteed to always converge to the global maximum of the likelihood, Eq. (3.1). It is true generally that an EM algorithm must converge to some maximum of the likelihood, but it is possible, depending on initial conditions, to find a local rather than global maximum. In the present case, however, the likelihood has only a single maximum, a property it inherits from the original Bradley-Terry model [1], and hence convergence to the global maximum is assured.

The end result of the algorithm is a complete set of strengths λ_u that specify the ranking of the individuals—higher λ_u implies higher ranking. If we wish, we can convert these parameters back to the original scores $s_u = \log \lambda_u$, which are more symmetric and arguably easier to interpret. A minor technical issue is that the likelihood of Eq. (3.1) is invariant under the change $\lambda_u \to 1/\lambda_u$, $q_t \to 1 - q_t$ for all u and t, meaning that, depending on the initial conditions, the final ranking may end up being upside down, so that the most dominant individuals have the lowest scores and the most subordinate ones have the highest. If this happens one need merely invert all the values of the λ_u to set them the right way up.

4. Prior on the strength parameters

The procedure described in the previous section amounts to a complete algorithm for calculating the maximum-likelihood values of all the parameters in our model. In some cases, however, this method gives poor results, for well understood reasons. Maximum-likelihood estimates of Bradley-Terry style models can perform poorly because they give undue weight to very large and very small values of the scores s_u. This is because the maximum-likelihood approach effectively assumes an (improper) uniform prior on s_u, but s_u has infinite support $s_u \in [-\infty, +\infty]$, so all but a vanishing fraction of the prior weight is on arbitrarily large values. This results in a number of well-known problems, particularly that the value of s_u for any individual who is dominant in every interaction is automatically infinite, which makes it impossible to tell any two such individuals apart.

One solution to these problems is to impose a more appropriate prior on s_u and a natural choice is a logistic prior [25,26]. Consider again the quantity p_0 defined in Eq. (2.3), which is the probability that an individual with strength λ dominates against the average individual:

$$p_0 = \frac{\lambda}{\lambda + 1} = \frac{e^s}{e^s + 1}. \tag{4.1}$$

In the absence of any information to the contrary, we assume this probability to be uniformly distributed between zero and one—the minimally informative or maximum-entropy prior $P(p_0) = 1$. Then the corresponding prior on the score s is

$$P(s) = P(p_0)\frac{dp_0}{ds} = \frac{dp_0}{ds} = \frac{1}{(e^s + 1)(e^{-s} + 1)}, \tag{4.2}$$

which is the logistic distribution. This is the prior we use in our work.

Now instead of maximizing the likelihood of Eq. (2.7) we maximize the posterior probability, assuming that the prior on q is also uniform. It is important to recognize that the resulting estimate, like all maximum a posteriori (MAP) estimates, depends on the parametrization of the likelihood, and specifically in this case on whether we maximize with respect to s_u or λ_u. The maximum with respect to one will not in general lie in the same place as the maximum with respect to the other. In our work we consider the s_u to be the more fundamental set of variables and maximize the posterior distribution

$$P(s, q, \sigma|x) = P(x, \sigma|s, q)\frac{P(s)P(q)}{P(x)} \tag{4.3}$$

with respect to s. With $P(s)$ as in Eq. (4.2) and uniform $P(q)$, and changing variables back to λ_u, this gives us

$$P(s,q,\sigma|x) = \prod_{r=1}^{M} \frac{(\lambda_{u_r} q_{t_r})^{\sigma_r} [\lambda_{v_r}(1-q_{t_r})]^{1-\sigma_r}}{\lambda_{u_r}+\lambda_{v_r}} \prod_{u=1}^{N} \frac{\lambda_u}{(\lambda_u+1)^2}. \tag{4.4}$$

In addition to regularizing the score parameters, the addition of the prior also eliminates the invariance of the model under a uniform shift of the scores and hence eliminates the need to normalize them. (Note that, although we have for convenience changed variables to λ_u, Eq. (4.4) still represents the probability of s_u and not of λ_u. The probability of λ_u would have a different form because there would be an additional Jacobian factor.)

With the addition of the extra term in (4.4), the derivation of the algorithm proceeds as before. Equations (3.4) and (3.6) for the quantities π_r and q_t remain unchanged, while the equation for the strengths λ_u now becomes

$$\lambda_i = \frac{1+\sum_r [\pi_r \delta_{u_r i} + (1-\pi_r)\delta_{v_r i}]}{2/(\lambda_i+1) + \sum_u A_{iu}/(\lambda_i+\lambda_u)}. \tag{4.5}$$

Other than this change, and the omission of the now-unnecessary normalization step, the algorithm is the same as before.

5. Results

We demonstrate our approach with example applications to a set of computer-generated test data and to three real-world data sets, an animal dominance hierarchy, a human hierarchy, and an example from competitive team sport.

Synthetic tests

As a first demonstration of our approach we present the results of a set of tests using synthetic (computer-generated) data. In these tests we generated a large number of random data sets with $N = 100$ individuals each and interactions between pairs chosen uniformly at random (a random graph in the language of network theory). Because the sparsity of the interactions can affect our ability to perform accurate inference, we study cases with both relatively dense interactions ($M = 5000$) and sparser ones ($M = 1000$). Tests were also performed with two different choices of the number of interaction types, $T = 5$ and 10. The winners of the interactions were generated using the model of Section 2 with independent scores s_u drawn from the logistic distribution of Eq. (4.2) and valence probabilities q_t drawn uniformly in an interval $[q_{\text{min}}, q_{\text{max}}]$ for various values of q_{min} and q_{max} as described below.

For each choice of M, T, q_{min}, and q_{max} we generated 1000 data sets and analysed each one by the MAP estimation method of Section 4 and also by fitting to a traditional unimodal Bradley-Terry model in which all interactions are considered equivalent—the standard procedure for calculations of dominance hierarchies for example. Following each analysis we ranked the fictional participants according to their inferred scores s_u and computed a Spearman rank correlation between these inferred ranks and the ground-truth ranks implied by the original scores used to generate the data, thereby testing our ability to recover the true ranking of the individuals. The results are shown in Table 1.

We explore three different choices for the valence probabilities q_t. In the first column of results in Table 1 we choose values of q_t uniformly at random in the interval $[0.5, 1]$. This means that all interactions are dominant in the sense of being instigated or won by the dominant individual more often than not. In this situation our method does well at recovering the ground-truth ranking with Spearman $R^2 > 0.5$ in all cases, but the standard Bradley-Terry analysis does almost as well. This is expected since the standard method assumes that indeed all interactions are dominant. There is nonetheless some daylight between the two methods—our method does better by a small

Table 1. The results of tests of our method on computer-generated data, compared with analyses of the same data using a standard ranking algorithm that assumes all interactions to be equivalent. All generated examples have $N = 100$ individuals and interactions placed between pairs chosen uniformly at random. Values of the number of interactions M and number of interaction types T are as listed and values of q_t are chosen uniformly at random in the intervals shown. The results are Spearman R^2 values between the inferred ranking of the individuals and the ground-truth ranking calculated from the parameters used to generate the data, averaged over 1000 random instances. In each entry x/y the first number x is the result from the method of this paper and the second y is from the traditional ranking calculation. Standard errors are less than ± 1 in the final digit in all cases.

		Spearman R^2 (this paper)/(traditional)		
M	T	$0.5 \le q_t \le 1$	$0.25 \le q_t \le 1$	$0 \le q_t \le 1$
5000	5	0.88/0.83	0.83/0.53	0.88/0.42
5000	10	0.89/0.85	0.85/0.52	0.89/0.29
1000	5	0.53/0.50	0.43/0.24	0.54/0.17
1000	10	0.54/0.52	0.41/0.22	0.54/0.11

margin in all cases because it is able to use the fact that some interactions are more strongly indicative of dominance than others.

The difference becomes more pronounced in the remaining columns of the table. In the last column we choose values of q_t in the interval $[0, 1]$, so that interactions are equally likely to indicate dominance or subordination. In this case our method has a large advantage, since it can estimate the valence of the interactions from the data while the standard method cannot, and we see that there is indeed a large difference between the results for the two methods in this regime—our method continues to do well at recovering the true ranking while the standard method performs poorly.

The middle column gives results for the intermediate case where q_t falls in the interval $[0.25, 1]$, which represents a situation in which most interactions indicate dominance but a few do not. Again we see in this situation that the method of this paper significantly outperforms the standard analysis because it is able to discern the valence of the interactions.

The effect on the results of the number of interaction types T is small for both our method and the traditional ranking. The effect of the number of observations M on the other hand is more pronounced: our method returns excellent results for all cases with $M = 5000$, no matter the values of the other parameters, but is noticeably poorer for the sparser case of $M = 1000$. The traditional method also performs poorly for $M = 1000$, although in most cases it does poorly for $M = 5000$ too.

To summarize, our method gives modestly improved rankings in situations where all interactions are indicative of dominance and substantially improved rankings in situations where they are not. Even in cases where the difference between the methods is small, however, the results may be interesting for other reasons. In particular, our method also gives an estimate of the valence probability q_t for each interaction type, which can be of interest in its own right, as we see in the following sections.

Dominance hierarchy in vervet monkeys

We now turn to applications of our method to real-world data. Our first example application is to a classic animal dominance hierarchy. We analyse observations reported by Vilette et al. [27] of 66 wild vervet monkeys in the Samara Private Game Reserve in South Africa between January 2015 and December 2017. The authors report a total of 11 664 agonistic encounters between pairs of monkeys, divided into eight types: charge, chase, displace, facial, lunge, physical, supplant, and vocal. A few other types, such as "scream" and "grab", were recorded but were rare—less than 1%

rspa.royalsocietypublishing.org Proc R Soc A 0000000

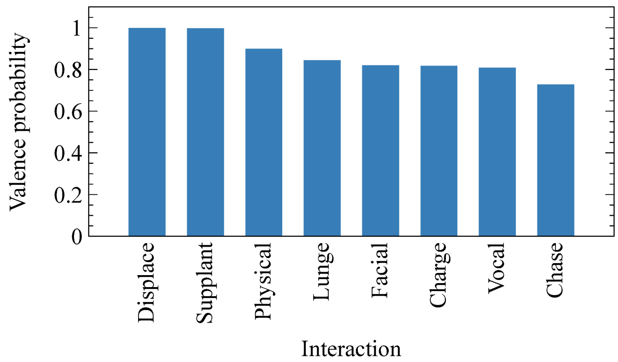

Figure 1. The valence probabilities q_t for the vervet monkey interactions of Vilette et al. [27] as analysed using the methods of this paper. These represent the probability that the given interaction will be instigated by a dominant individual, so that values approaching one indicate dominant behaviours and values approaching zero indicate subordinate ones. In this case all behaviours are dominant on balance, but some more so than others.

of the total—and were removed from the data for our analysis, as were interactions whose type was unknown. The remaining data were analysed using the MAP estimation method of Section 4.

Figure 1 shows the inferred values of the valence probabilities q_t for each of the interaction types. Recall that these values tell us, for each interaction type, the probability that the interaction will be instigated by the dominant individual of a pair. Values $q_t > \frac{1}{2}$ indicate interactions that are dominant on average and in this case we find (not surprisingly) that all interaction types are dominant, but that they vary in their degree of dominance. "Displace" and "supplant", for instance, while being arguably the least physical of the interactions, are the most indicative of dominance, while "chase" is the least indicative. "Chase" has a valence probability of only 0.728, implying that more than a quarter of the time it is not the chaser but the chasee who is the dominant individual in a pair. Such interactions are thus less reliable indicators of dominance than "displace" and "supplant".

In addition to being informative in their own right, these values now allow us to make a more accurate estimate of the dominance hierarchy: our method automatically weights more (or less) heavily those interactions that strongly (or weakly) indicate dominance. To illustrate this effect, we show in Fig. 2 a comparison of the rankings of the monkeys, from 1 to 66, computed first as above and second assuming that all interactions are equivalent and equally indicative of dominance. If the two sets of rankings agreed perfectly all points in the figure would lie on the dashed line, but, as the figure shows, there are some significant differences between them—monkeys who are ranked either higher or lower by our method, up to 17 places different in the most extreme case.

Student social network

For our second example we analyse social interactions among the students in a 7th grade class in Victoria, Australia, as compiled by Vickers and Chan [28]. In this study the authors interviewed 29 students in a single school class and asked them three questions: (1) who do you get on with in the class, (2) who are your best friends in the class, and (3) who do you prefer to work with? The answers to these three questions define three different types of directed relations between the students. These particular relations are not intrinsically competitive and hence we might not necessarily expect them to define a hierarchy. However, in other work on friendship among school students it has been found that claims of friendship define a clear hierarchy [29,30], and the same turns out to be true in the present case: using the methods described in this paper we find a strong hierarchy hidden in the data of Vickers and Chan, as shown in Fig. 3. In this network

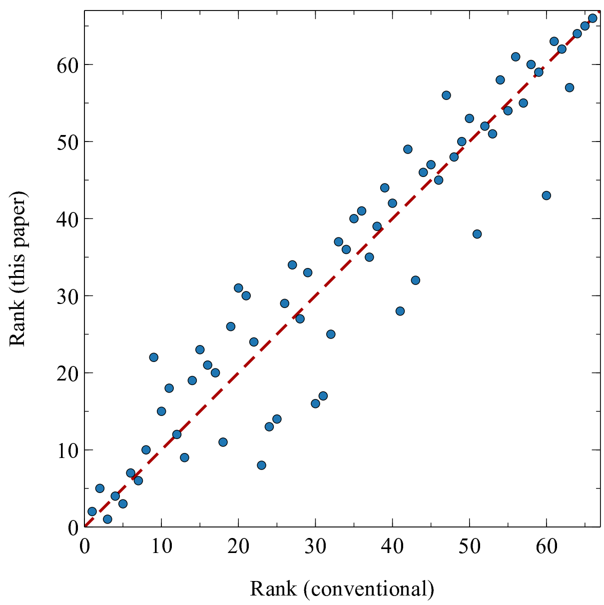

Figure 2. Scatter plot of the rankings of the vervet monkeys. Rankings were computed in two ways: using the method of this paper and by conventional methods that assume all interactions to be equally informative (equivalent to assuming that the valence probability q_t is 1 for all interaction types). The strength parameters λ_u were computed in both ways then the monkeys were sorted in order of their strengths from highest to lowest to give the rankings plotted here.

visualization the nodes represent the students, the directed connections indicate answers to the survey questions, and the vertical position of each node on the page indicates the rank score s_u assigned to it by the analysis, so that nodes higher up the page are more highly ranked. As we can see, most edges in the network run in an upward direction, meaning that students say they get on with, are friends with, or prefer to work with others who are above them in the hierarchy.

On the one hand, this may seem counterintuitive, since one might imagine that friendship and co-working relations should be symmetric: if A is friends with B then surely B is friends with A as well? On the other hand, it is common for children (and perhaps adults too) to make or claim "aspirational" connections: they want to be friends with higher status others [29–31]. These effects lead to asymmetries in reported social networks that can be used, as here, to infer hierarchy.

Figure 4 shows the values of the parameters q_t for the three types of connection in this case and here we also find something interesting. The first type ("who do you get on with?") is relatively weakly indicative of hierarchy, but for the second and third types ("best friends" and "work with") the most probable value of q_t is 1, implying that these types of connections are maximally indicative of hierarchy. Moreover, because the value of q_t is the same for these types the algorithm treats them in an identical way, which means effectively that there are only two types of connections in these data. The "best friends" and "work with" interactions are flattened into a single combined interaction type by the analysis. We have encountered similar flattening of the data in some other examples we have analysed, including animal hierarchies, professional networks, and some online social networks.

American football competition

Ranking methods are often applied in sport and athletic competition to rank players or teams. As a demonstration of this type of application we apply our method to professional American football. American football provides an interesting example because, unlike association football,

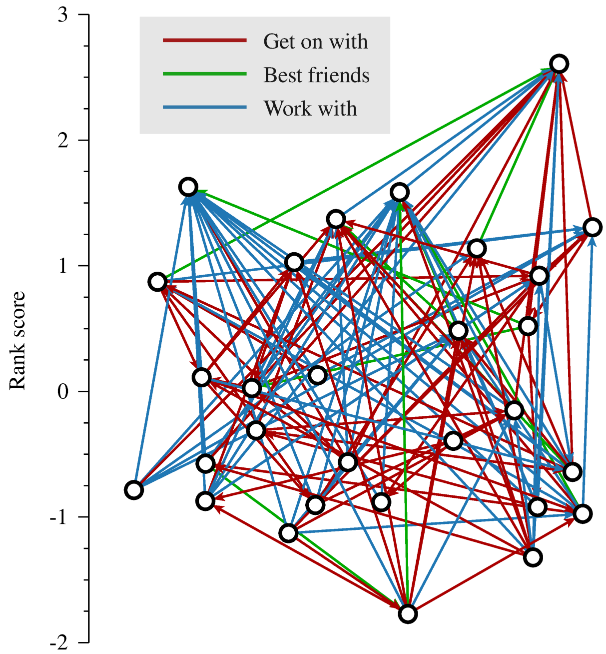

Figure 3. Network of 7th grade students in the study of Vickers and Chan [28]. Nodes represent the students, edges denote the three types of relations as indicated, and the vertical position of each node represents its calculated rank score s_u on the scale shown at the left. Following [30] we show only the one-way connections between nodes for clarity and we show at most one edge between each pair of nodes. Edges between nodes connected in both directions are omitted.

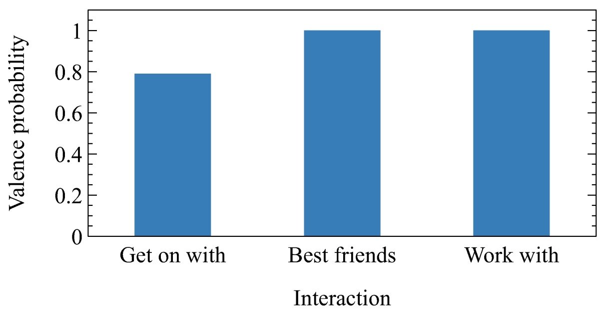

Figure 4. Values of the valence probabilities q_t for the three types of interactions in the student data of Vickers and Chan [28].

it consists of discrete "plays", in which one team (the offense) has possession of the ball and tries to advance it up the field against the other team (the defense). There are different types of plays teams use to do this, including running plays (the ball is carried by a runner), passing plays (the ball is thrown), and punts (the ball is kicked). Here we analyse these individual plays as interactions between the teams and compute a ranking of teams based on the pattern of interactions in all (regular season) games in a given playing season.

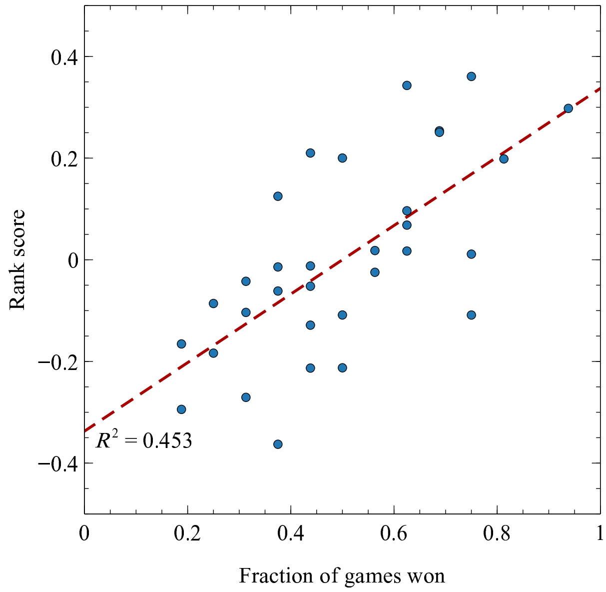

Figure 5. The rank scores s_u of the 32 American football teams in the analysis described here for the 2015 NFL season, plotted against the fraction of games they won in the same season.

A key aspect of this analysis is that we use no information about the actual success of the plays—whether they advance the ball, for instance, or whether any points are scored. Moreover, we specifically remove from the data compulsory plays such as kickoffs and conversions that implicitly signal point scoring, so the only information available to the algorithm is *which* types of plays the teams choose to run. (We do include field goals, which score points, because these are optional and hence are revealing from a strategic point of view.)

Even without explicit indicators of success, however, we can extract a meaningful ranking of the teams. Figure 5 shows an example for the 2015 season of the US National Football League using data from Yurko et al. [32]. With $36\,030$ interactions in total, this example is the largest in this paper, but our algorithm nonetheless runs quickly. Total running time for the calculation was 11 seconds on a standard laptop computer (*circa* 2022). The figure shows the inferred rank score s_u of each of the 32 teams in the league plotted against their actual success during the season, represented by the fraction of games they won. Although the correlation between the two measures is not perfect, it is substantial and significant ($R^2 = 0.453$, $p < 0.0001$). Note that we should not expect perfect correlation even if our rankings were perfectly accurate, since it is an important aspect of commercially successful spectator sports that they contain an element of randomness. If the higher ranked team always won there would be little suspense about the outcome of a game and correspondingly little motivation to watch, so the existence of any method that could reliably predict game winners would be a clear sign of an unsuccessful sport (which American football certainly is not).

One might suppose our success at ranking the teams to be a result of the simple fact that winning teams run more plays than losing teams, because they are in possession of the ball more often, but this is not the case. In fact there is hardly any difference between the number of plays run by the best and worst teams: the top ten teams in our ranking for the 2015 season, for instance, ran an average of 1134 plays in total during the regular season while the bottom ten actually ran a slightly larger number of 1169.

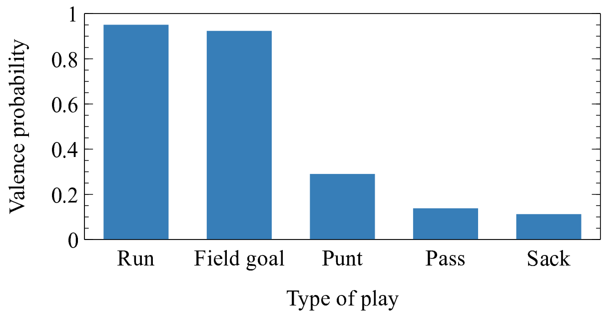

Figure 6. The valence probabilities q_t for each of the five play types used in the ranking of American football teams.

The ranking of teams is signalled not by the number of plays, but by which plays the teams run. Figure 6 shows the values of the valence probabilities q_t for each of the five play types included in our analysis. By contrast with our previous examples, not all types of plays indicate dominance. Two types do clearly signal dominance—running plays and field goals—with values of q_t well above $\frac{1}{2}$. The remaining three, however, signal subordination. Of these, the punt is only used to get rid of the ball when a team knows they are likely to lose it anyway, and hence is a naturally subordinate trait. And a sack—the player with the ball gets tackled before they can move it up the field—is a clear sign of team weakness. More surprising is that passing plays, where the ball is thrown, are also a sign of weakness. In general passing plays are some of the most spectacular and successful plays in American football, so one might ask why they are indicative of subordination. The answer may be that passing plays are challenging to execute and often fail, because for instance the thrown ball is not caught or is intercepted by the opposing team. This means that weaker teams have to make more attempts to achieve successful passing plays than stronger teams and hence, on balance, passing plays are indicative of weakness. For instance, during the 2015 season the top ten teams in our ranking made an average of 504 passing plays each while the bottom ten made an average of 620.

6. Conclusions

In this paper we have considered the problem of ranking a set of individuals or teams based on pairwise comparisons when there are multiple types of comparison. Examples include animal dominance hierarchies in which animals use a range of different behaviours to establish or signal dominance, and sporting competitions in which teams use a range of different moves or plays against their opponents. We have shown that even if one does not know in advance either the ranking of the individuals or what information each type of interaction conveys, it is possible to infer both from observed interactions. We have described an efficient method for doing this which combines an expectation-maximization algorithm with a variant of the Bradley-Terry model.

We have presented a number of example applications of the method, including applications to animal and human dominance hierarchies, and an application to the sport of American football. The method provides a way to sensitively infer rankings taking all interaction types into account and weighting each one appropriately given the information it contains. At the same time the results shed light on the interactions themselves, telling us, without need for other input, whether each type of interaction is indicative of dominance or subordination, and to what extent.

Natural extensions of the work reported here include the exploration of alternative models for multimodal comparisons, including generalizations of popular models for the unimodal case such as Thurstonian models [3,33] or models that allow for dependencies between observations [5]. One could also consider goodness of fit measures to assess the success of our model or any other, model selection to choose between alternatives, or more elaborate inference procedures for the current model, including fully Bayesian approaches similar to those applied in the unimodal case [25,34], which would have the advantage of making the ranking independent of the parametrization of the model. These extensions, however, we leave for future work.

Acknowledgements

The author thanks Elizabeth Bruch, Carrie Ferrario, Liza Levina, and Ambuj Tewari for useful conversations. This work was funded in part by the US National Science Foundation under grant DMS–2005899.

Code and data

Complete computer code implementing the method of this paper, along with example data, is available in the supplementary materials. Also included is a program to generate synthetic data as used in the tests in Section 5. The data for the examples on vervet monkeys, 7th grade students, and American football in Section 5 are all previously published and publicly available.

References

1. Zermelo E. 1929 Die Berechnung der Turnier-Ergebnisse als ein Maximumproblem der Wahrscheinlichkeitsrechnung. *Mathematische Zeitschrift* **29**, 436–460.
2. Bradley RA, Terry ME. 1952 Rank analysis of incomplete block designs: I. The method of paired comparisons. *Biometrika* **39**, 324–345.
3. David HA. 1988 *The Method of Paired Comparisons*. London: Griffin 2 edition.
4. Davidson RR, Farquhar PH. 1976 A bibliography on the method of paired comparisons. *Biometrics* **32**, 241–252.
5. Cattelan M. 2012 Models for paired comparison data: A review with emphasis on dependent data. *Statistical Science* **27**, 412–433.
6. Côté SD. 2000 Dominance hierarchies in female mountain goats: Stability, aggressiveness and determinants of rank. *Behaviour* **137**, 1541–1566.
7. Frank LG. 1986 Social organization of the spotted hyaena Crucuta crocuta: II. Dominance and reproduction. *Animal Behaviour* **34**, 1510–1527.
8. Golemiec M, Schneider J, Boyce WT, Bush NR, Adler N, Levine JD. 2016 Layered social network analysis reveals complex relationships in kindergarteners. *Frontiers in Psychology* **7**, 276.
9. Lott DF. 1979 Dominance relations and breeding rate in mature male American bison. *Zeitschrift Tierpsychologie* **49**, 418–432.
10. Davidson RR, Bradley RA. 1969 Multivariate paired comparisons: The extension of a univariate model and associated estimation and test procedures. *Biometrika* **56**, 81–95.
11. Bockenholt U. 1988 A logistic representation of multivariate paired comparison models. *Journal of Mathematical Psychology* **32**, 44–63.
12. Dittrich R, Francis B, Hatzinger R, Katzenbeisser W. 2006 Modelling dependency in multivariate paired comparisons: A log-linear approach. *Mathematical Social Sciences* **52**, 197–209.
13. Lin S. 2010 Rank aggregation methods. *Wiley Interdisciplinary Reviews: Computational Statistics* **2**, 555–570.
14. Rajkumar A, Agarwal S. 2014 A statistical convergence perspective of algorithms for rank aggregation from pairwise data. *Proceedings of Machine Learning Research* **32**, 118–126.
15. Dwork C, Kumar R, Naor M, Sivakumar D. 2001 Rank aggregation methods for the web. In *Proceedings of the 10th International Conference on the World Wide Web (WWW 2001)* pp. 613–622 New York. Association of Computing Machinery.
16. Pritikin JN. 2020 An exploratory factor model for ordinal paired comparison indicators. *Heliyon* **6**, e04821.
17. Newman M. 2018 *Networks*. Oxford: Oxford University Press 2 edition.
18. Bianconi G. 2018 *Multilayer Networks: Structure and Function*. Oxford: Oxford University Press.

19. Boccaletti S, Bianconi G, Criado R, del Genio CI, Gomez-Gardenes J, Romance M, Sendina-Nadal I, Wang Z, Zanin M. 2014 The structure and dynamics of multilayer networks. *Physics Reports* **544**, 1–122.
20. Kivelä M, Arenas A, Barthelemy M, Gleeson JP, Moreno Y, Porter MA. 2014 Multilayer networks. *Journal of Complex Networks* **2**, 203–271.
21. Iacovacci J, Rahmede C, Arenas A, Bianconi G. 2016 Functional multiplex pagerank. *Europhys. Lett.* **116**, 28044.
22. Rahmede C, Iacovacci J, Arenas A, Bianconi G. 2018 Centralities of nodes and influences of layers in large multiplex networks. *Journal of Complex Networks* **6**, 733–752.
23. Rao PV, Kupper LL. 1967 Ties in paired-comparison experiments: A generalization of the Bradley-Terry model. *Journal of the American Statistical Association* **62**, 194–204.
24. Davidson RR. 1970 On extending the Bradley-Terry model to accommodate ties in paired comparison experiments. *Journal of the American Statistical Association* **65**, 317–328.
25. Davidson RR, Solomon DL. 1973 A Bayesian approach to paired comparison experimentation. *Biometrika* **60**, 477–487.
26. Whelan JT. 2017 Prior distributions for the Bradley-Terry model of paired comparisons. Preprint arxiv:1712.05311.
27. Vilette C, Bonnell T, Henzi P, Barrett L. 2020 Comparing dominance hierarchy methods using a data-splitting approach with real-world data. *Behavioral Ecology* **31**, 1379–1390.
28. Vickers M, Chan S. 1981 Representing classroom social structure. Technical report Victoria Institute of Secondary Education Melbourne.
29. Hallinan MT, Kubitschek WN. 1988 The effect of individual and structural characteristics on intransitivity in social networks. *Social Psychology Quarterly* **51**, 81–92.
30. Ball B, Newman MEJ. 2013 Friendship networks and social status. *Network Science* **1**, 16–30.
31. Dijkstra JK, Cillessen AHN, Lindenberg S, Veenstra R. 2010 Basking in reflected glory and its limits: Why adolescents hang out with popular peers. *Journal of Research on Adolescents* **20**, 942–958.
32. Yurko R, Ventura S, Horowitz M. 2019 nflWAR: A reproducible method for offensive player evaluation in football. *Journal of Quantitative Analysis in Sports* **15**, 163–183.
33. Thurstone LL. 1927 A law of comparative judgment. *Psychological Review* **34**, 368–389.
34. Caron F, Doucet A. 2012 Efficient Bayesian inference for generalized Bradley-Terry models. *Journal of Computational and Graphical Statistics* **21**, 174–196.