

Luck, skill, and depth of competition in games and social hierarchies

Maximilian Jerdee¹ and M. E. J. Newman^{1,2}

¹*Department of Physics, University of Michigan, Ann Arbor, Michigan 48109, USA*

²*Center for the Study of Complex Systems, University of Michigan, Ann Arbor, Michigan 48109, USA*

Patterns of wins and losses in pairwise contests, such as occur in sports and games, consumer research and paired comparison studies, and human and animal social hierarchies, are commonly analyzed using probabilistic models that allow one to quantify the strength of competitors or predict the outcome of future contests. Here we generalize this approach to incorporate two additional features: an element of randomness or luck that leads to upset wins, and a “depth of competition” variable that measures the complexity of a game or hierarchy. Fitting the resulting model to a large collection of data sets we estimate depth and luck in a range of games, sports, and social situations. In general, we find that social competition tends to be “deep,” meaning it has a pronounced hierarchy with many distinct levels, but also that there is often a nonzero chance of an upset victory, meaning that dominance challenges can be won even by significant underdogs. Competition in sports and games, by contrast, tends to be shallow and in most cases there is little evidence of upset wins, beyond those already implied by the shallowness of the hierarchy.

I. INTRODUCTION

One of the oldest and best-studied problems in data science is the ranking of a set of items, individuals, or teams based on the results of pairwise comparisons between them. Such problems arise in sports, games, and other competitive human interactions, in paired comparison surveys in market research and consumer choice, in revealed-preference studies of human behavior, and in studies of social hierarchies in both humans and animals. In each of these settings, one has a set of comparisons between pairs of items or competitors, with outcomes of the form “A beats B” or “A is preferred to B,” and the goal is to determine a ranking of competitors from best to worst, allowing for the fact that the data may be sparse (there may be no data for many pairs) or contradictory (e.g., A beats B beats C beats A). A group of chess players might play in a tournament, for example, and record wins and losses against each other. Consumers might express preferences between pairs of competing products, either directly in a survey or implicitly through their purchases or other actions. A flock of chickens might peck each other as a researcher records who pecked whom in order to establish the classic “pecking order.”

A large number of methods have been proposed for solving ranking problems of this kind—see Refs. [1–3] for reviews. In this paper we consider one of the most common, which uses a statistical model for wins and losses and then fits that model to observed win/loss data. In the most widely adopted version one considers a population of n competitors labeled by $i = 1 \dots n$ and assigns to each a real score parameter $s_i \in [-\infty, \infty]$. Then the probability that i beats j in a single pairwise match or contest is assumed to be some function of the difference of their scores: $p_{ij} = f(s_i - s_j)$. The function $f(s)$ satisfies the following axioms:

1. It is increasing in s , since by definition a better competitor has a higher probability of winning than a worse one.

2. It tends to 1 as $s \rightarrow \infty$ and to 0 as $s \rightarrow -\infty$, meaning that an infinitely good player always wins and an infinitely poor one always loses.
3. It is antisymmetric about its mid-point at $s = 0$, with the form

$$f(-s) = 1 - f(s), \quad (1)$$

because the probability of losing is, by definition, one minus the probability of winning. As a corollary, this also implies that the probability $f(0)$ of beating an evenly matched opponent is always $\frac{1}{2}$.

Subject to these constraints the function can still take a wide variety of forms, but the most popular choice by far is the logistic function $f(s) = 1/(1 + e^{-s})$ —shown as the bold curve in Fig. 1a—which gives

$$f(s_i - s_j) = \frac{e^{s_i}}{e^{s_i} + e^{s_j}}. \quad (2)$$

The resulting model is known as the Bradley-Terry model, after R. Bradley and M. Terry who described it in 1952 [4], although it was (unbeknown to them) first introduced much earlier, by Zermelo in 1929 [5].

Given the model, one can estimate the values of the score parameters s_i by a number of standard methods, including maximum likelihood estimation [4–8], maximum a posteriori estimation [9], or Bayesian methods [10, 11], then rank competitors from best to worst in order of their scores. The fitted model can also be used to predict the outcome of future contests between any of the competitors, even if they have never directly competed in the past.

This approach is effective and widely used, but the standard Bradley-Terry model is a simplistic representation of the patterns of actual competition and omits many important elements found in real-world interactions. Generalizations of the model have been proposed that incorporate some of these elements, such as the possibility of ties or draws between competitors [12, 13],

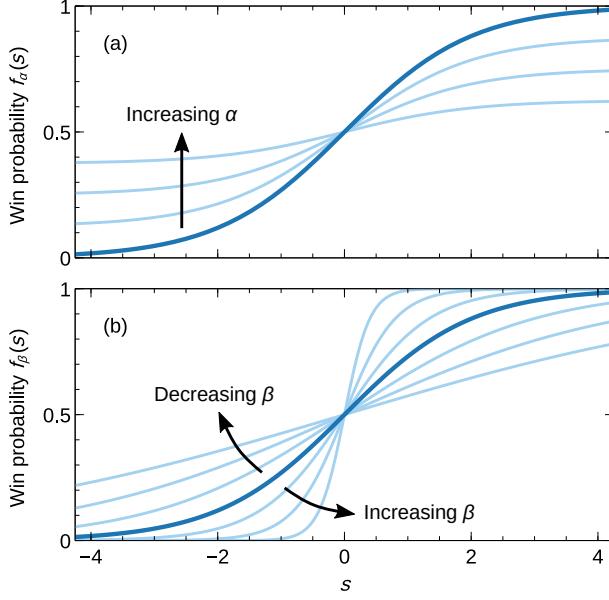


FIG. 1. Score functions $f(s)$. (a) The bold curve represents the standard logistic function $f(s) = 1/(1 + e^{-s})$ used in the Bradley-Terry model. The remaining curves show the function f_α of Eq. (8) for increasing values of the luck parameter α . (b) The score function f_β of Eq. (9) for different values of the depth of competition β , both greater than 1 (steeper) and less than 1 (shallower).

multiway competition as in a horse race [14, 15], or the “home-field advantage” of playing on your own turf [16]. In this paper we consider a further extension of the model that incorporates two additional features of particular interest, which have received comparatively little previous attention: the element of luck inherent for instance in games of chance, and the notion of “depth of competition,” which captures the complexity of games or the number of distinct levels in a social hierarchy. In the remainder of the paper we define and motivate this model and then describe a Bayesian approach for fitting it to data, which we use to infer the values of the luck and depth variables for a variety of real-world data sets drawn from different arenas of human and animal competition. Our results suggest that social hierarchies are in general deeper and may have a larger element of luck to their dynamics than recreational games and sports, which tend to be shallower and show little evidence of a luck component.

Software implementations of the various methods described in this paper are available at <https://github.com/maxjerdee/pairwise-ranking>.

II. THE MODEL

Suppose we observe m matches between n players. The outcomes of the matches can be represented by an $n \times n$

matrix \mathbf{A} with element A_{ij} equal to the number of times player i beats player j . Within the standard Bradley-Terry model the probability of a win is given by Eq. (2) and, assuming the matches to be statistically independent, the probability or likelihood of the complete set of observed outcomes is

$$P(\mathbf{A}|\mathbf{s}) = \prod_{ij} f(s_i - s_j)^{A_{ij}} = \prod_{ij} \left(\frac{e^{s_i}}{e^{s_i} + e^{s_j}} \right)^{A_{ij}}, \quad (3)$$

where \mathbf{s} is the vector with elements s_i . (We assume that the structure of the tournament—who plays whom—is determined separately, so that (3) is a distribution over the directions of the wins and losses only and not over which pairs of players competed.)

The scores are traditionally estimated by the method of maximum likelihood, maximizing (3) with respect to all s_i simultaneously to give estimates

$$\hat{\mathbf{s}} = \operatorname{argmax}_{\mathbf{s}} P(\mathbf{A}|\mathbf{s}). \quad (4)$$

These maximum likelihood estimates (MLEs) can then be sorted in order to give a ranking of the competitors, or simply reported as measures of strength in their own right. The widely used Elo ranking system for chess players, for example, is essentially a version of this approach, but extended to allow for dynamic updates as new matches are added to the data set.

The maximum likelihood approach unfortunately has some drawbacks. For one, the likelihood is invariant under a uniform additive shift of all scores s_i and hence the scores are not strictly identifiable, though this issue can easily be fixed by normalization. A more serious problem is that the likelihood maximum does not exist at all unless the network of interactions—the directed network with adjacency matrix \mathbf{A} —is strongly connected (meaning there is a directed chain of victories from any player to any other), and the maximum likelihood estimation procedure fails, with the divergence of some or all of the scores, unless this relatively stringent condition is met.

This issue can be addressed by introducing a prior on the scores and adopting a Bayesian perspective. A variety of potential priors for this purpose have been systematically examined by Whelan [9], who, after careful consideration, recommends a Gaussian prior with mean zero. The variance is arbitrary—it merely sets the scale on which the score s is measured—but for subsequent convenience we here choose a variance of $\frac{1}{2}$ so that the prior on \mathbf{s} takes the form

$$P(\mathbf{s}) = \prod_{i=1}^n \frac{1}{\sqrt{\pi}} e^{-s_i^2}. \quad (5)$$

An alternative prior, also recommended by Whelan, is the logistic distribution

$$P_L(\mathbf{s}) = \prod_{i=1}^n \frac{1}{(1 + e^{s_i})(1 + e^{-s_i})}. \quad (6)$$

In practice the Gaussian and logistic distributions are similar in shape and the choice of one or the other does not make a great deal of difference. The logistic distribution is perhaps the less natural of the two and we primarily use the Gaussian distribution in this paper, but the logistic distribution does have the advantage of leading to faster numerical algorithms and we have used it in previous work for this reason [8, 17]. We also include it in the basket of models that we compare in Section V.

Once we have defined a prior on the scores we can calculate a maximum a posteriori (MAP) estimate of their values as

$$\hat{\mathbf{s}} = \text{argmax}_{\mathbf{s}} P(\mathbf{s}|\mathbf{A}) = \text{argmax}_{\mathbf{s}} P(\mathbf{A}|\mathbf{s})P(\mathbf{s}). \quad (7)$$

The MAP estimate always exists regardless of whether the interaction network is strongly connected, and using a prior also eliminates the invariance of the probability under an additive shift and hence the need for normalization. As an alternative to computing a MAP estimate we can also simply return the full posterior distribution $P(\mathbf{s}|\mathbf{A})$, which gives us complete information on the expected values and uncertainty of the scores given the observed data.

III. EXTENSIONS OF THE MODEL

In this section we define generalizations of the Bradley-Terry model that extend the score function f in two useful ways, while keeping other aspects of the model fixed, including the normal prior. The specific generalizations we consider involve dilation or contraction of the score function in the vertical and horizontal directions. Vertical variation controls the element of luck that allows a weak player to sometimes beat a strong one; horizontal variation controls the “depth of competition,” a measure of the complexity of a game or contest.

A. Upset wins and luck

The first generalization of the Bradley-Terry model that we consider is one where the function f is contracted in the vertical direction, as shown in Fig. 1a. We parametrize this function in the form

$$f_\alpha(s) = \frac{1}{2}\alpha + (1 - \alpha)\frac{1}{1 + e^{-s}}, \quad (8)$$

with $\alpha \in [0, 1]$. In the traditional Bradley-Terry model $f(s)$ tends to 0 and 1 as $s \rightarrow \pm\infty$, as discussed in the introduction, but in the modified model with $\alpha > 0$ this is no longer the case. One can think of the parameter α as controlling the probability of an “upset win” in which an infinitely good player loses or an infinitely bad player wins. (The probabilities of these two events must be the same because of the antisymmetry condition, Eq. (1).)

For some games or competitions it is reasonable that $f(s)$ tends to zero and one at the limits. In a game like chess that has no element of randomness, an infinitely good player may indeed win every time. In a game of pure luck like roulette, on the other hand, both players have equal probability $\frac{1}{2}$ of winning, regardless of skill. These two cases correspond to the extreme values $\alpha = 0$ and $\alpha = 1$ respectively in Eq. (8). Values in between represent games that combine both luck and skill, like poker or backgammon, with the precise value of α representing the proportion of luck. For this reason we refer to α as the luck parameter, or simply the “luck.”

One could also consider the chance of the weaker player winning in the standard Bradley-Terry model to be an example of luck or an upset win, but that is not how we use these words here. In the present context the “luck” α describes the probability of winning the game even if one’s opponent is infinitely good, which is zero in the standard model but nonzero in the model of Eq. (8) with $\alpha > 0$.)

Another way to think about α is to imagine a game as a mixture of a luck portion and a skill portion. With probability α the players play a game of pure chance in which the winner is chosen at random, for instance by the toss of a coin. Alternatively, with probability $1 - \alpha$, they play a game of skill, such as chess, and the winner is chosen with the standard Bradley-Terry probability. The overall probability of winning is then given by Eq. (8) and the parameter α represents the fraction of time the game is decided by pure luck. By fitting (8) to observed win-loss data we can learn the luck inherent in a competition or hierarchy. We do this for a variety of data sets in Section IV.

B. Depth of competition

The second generalization we consider is one where the function f is dilated or contracted in the horizontal direction, as shown in Fig. 1b, by a uniform factor $\beta > 0$ thus:

$$f_\beta(s) = \frac{1}{1 + e^{-\beta s}}. \quad (9)$$

The slope of this function at $s = 0$ is given by

$$f'_\beta(0) = \left[\frac{\beta e^{-\beta s}}{(1 + e^{-\beta s})^2} \right]_{s=0} = \frac{1}{4}\beta, \quad (10)$$

so β is simply proportional to the slope. A more functional way of thinking about β is in terms of the probability that the stronger of a typical pair of competitors will win. With a normal prior on s of variance $\frac{1}{2}$ as described in Section II, the difference $s_i - s_j$ between the scores of a randomly chosen pair of competitors will be a priori normally distributed with variance 1, meaning the scores will be separated by an average (root-mean-square) distance of 1. Consider two players separated by this average distance. If β is small, making f_β a relatively flat function

(the shallowest curve in Fig. 1b), the probability p_{ij} of the stronger player winning will be close to $\frac{1}{2}$ and there is a substantial chance that the weaker player will win. Conversely, if β is large then p_{ij} will be close to 1 (the steepest curve in Fig. 1b) and the stronger player is very likely to prevail.

Thus one way to understand the parameter β is as a measure of the imbalance in strength or skill between the average pair of players. When β is large the contestants in the average game are very unevenly matched. As we will shortly see, this is a common situation in social hierarchies, but not in sports and games, perhaps because contests between unevenly matched opponents are less rewarding both for spectators and for the competitors themselves.

Another way to think about β is in terms of the number of levels of skill or strength in a competition. Suppose we define one “level” as the distance $\Delta s = s_i - s_j$ between scores such that i beats j with a certain probability q . For a win probability of the form of Eq. (9) we have $q = 1/(1 + e^{-\beta \Delta s})$ and hence

$$\Delta s = \frac{1}{\beta} \log \frac{q}{1-q}. \quad (11)$$

Considering again the typical pair of players a distance 1 apart, the number of levels between them is

$$\frac{1}{\Delta s} = \frac{\beta}{\log[q/(1-q)]}. \quad (12)$$

Thus the number of levels is simply proportional to β . Let us choose the probability q such that the constant of proportionality is 1, meaning $\log[q/(1-q)] = 1$ or

$$q = \frac{1}{1 + e^{-1}} = 0.731\dots \quad (13)$$

With this definition, a “level” is the skill difference Δs between two players such that the better one wins 73% of the time and our parameter β is simply equal to the number of such levels between the average pair of players.

In this interpretation, β can be thought of as a measure of the complexity or depth of a game or competition. A “deep” game, in this sense, is one that can be played at many levels, with players at each level markedly better than those at the level below. Chess, which is played at a wide range of skill levels from beginner to grandmaster, might be an example.

This concept of depth has a long history. For example, in an article in the trade publication *Inside Backgammon* in 1980 [18], world backgammon champion William Robertie defined a “skill differential” as the strength difference between two players that results in the better one winning 70 to 75% of the time—precisely our definition of a “level”—and the “complexity number” of a sport or game as the number of such skill differentials that separate the best player from the worst. Cauwet *et al.* [19] have defined a similar but more formal measure of game depth that they call “playing-level complexity.” There

has also been discussion in the animal behavior literature of the “steepness” of animal dominance hierarchies [20], which appears to correspond to roughly the same idea.

One should be careful about the details. Robertie and Cauwet *et al.* both define their measures in terms of the skill range between the best and worst players, but this could be problematic in that the range will depend on the particular sample of players one has and will tend to increase as the sample size gets larger, which seems undesirable. Our definition avoids this by considering not the best and worst players in a competition but the average pair of players, which gives a depth measure that is asymptotically independent of sample size.

Even when defined in this way, however, the number of levels is not solely about the intrinsic complexity of the game, but does also depend on who is competing. For example, if a certain competition is restricted to contestants who all fall in a narrow skill range, then β will be small even for a complex game. In a world-class chess tournament, for instance, where every player is an international master or better, the number of levels of play will be relatively small even though chess as a whole has many levels. Thus empirical values of β combine aspects of the complexity of the game with aspects of the competing population.

For this reason we avoid terms such as “complexity number” and “depth of game” that imply a focus on the game alone and refer to β instead as the “depth of competition,” which we feel better reflects its meaning.

C. Combined model

Combining both the luck and depth of competition variables into a single model gives us the score function

$$f_{\alpha\beta}(s) = \frac{1}{2}\alpha + (1-\alpha)\frac{1}{1+e^{-\beta s}}. \quad (14)$$

In Section IV we fit this form to observed data from a range of different areas of study in order to infer the values of α and β . In the process one can also infer the scores s_i , which can be used to rank the participants or predict the outcome of unobserved contests, and we explore this angle in Section V. In this section, however, our primary focus is on α and β and on understanding the varying levels of luck and depth in different kinds of competition.

To perform the fit we consider again a data set represented by its adjacency matrix \mathbf{A} and write the data likelihood in the form of Eq. (3):

$$P(\mathbf{A}|\mathbf{s}, \alpha, \beta) = \prod_{ij} f_{\alpha\beta}(s_i - s_j)^{A_{ij}}. \quad (15)$$

The scores \mathbf{s} are assumed to have the Gaussian prior of Eq. (5), and we assume a uniform (least informative) prior on α , which means $P(\alpha) = 1$. We cannot use a uniform prior on β , since it has infinite support, so instead we use a prior that is approximately uniform over

	Data set	$\hat{\beta}$	n	m	Description	Ref.
Sports/games	Scrabble	0.68	587	23477	<i>Scrabble</i> tournament matches 2004–2008	[21]
	Basketball	1.01	240	10002	National Basketball Association games 2015–2022	[22]
	Chess	1.17	917	7007	Online chess games on lichess.com in 2016	[23]
	Tennis	1.44	1272	29397	Association of Tennis Professionals matches 2010–2019	[24]
	Soccer	1.73	1976	7208	Men’s international association football matches 2010–2019	[25]
	Video games	1.77	125	1951	<i>Super Smash Bros Melee</i> tournament matches in 2022	[26]
Human	Friends	3.54	774	2799	High-school friend nominations	[27]
	CS departments	4.25	205	4388	Doctoral graduates of one department hired as faculty in another	[28]
	Business depts.	4.36	112	7856	Doctoral graduates of one department hired as faculty in another	[28]
Animal	Vervet monkeys	6.01	41	2930	Dominance interactions among a group of wild vervet monkeys	[29]
	Dogs	8.74	27	1143	Aggressive behaviors in a group of domestic dogs	[30]
	Baboons	13.19	53	4464	Dominance interactions among a group of captive baboons	[31]
	Sparrows	22.92	26	1238	Attacks and avoidances among sparrows in captivity	[32]
	Mice	26.48	30	1230	Dominance interactions among mice in captivity	[33]
	Hyenas	100.58	29	1913	Dominance interactions among hyenas in captivity	[34]

TABLE I. Data sets analyzed in Section IV, in order of increasing depth of competition β . Here n is the number of participants and m is the number of matches/interactions. Further information on the data sets is given in Appendix 1.

“reasonable” values of β and decays in some slow but integrable manner outside this range. A suitable choice in the present case is (the positive half of) a Cauchy distribution centered at zero:

$$P(\beta) = \frac{2w/\pi}{\beta^2 + w^2}, \quad (16)$$

where w controls the scale on which the function decays. In this paper we use $w = 4$, which roughly corresponds to the range of variation in β that we see in real-world data sets, and has the convenient property of giving a uniform prior on the angle of $f_\beta(s)$ at the origin.

It is worth mentioning that the choice of prior on β does have an effect on the results in some cases. When data sets are large and dense, priors tend to have relatively little impact because the posterior distribution is narrowly peaked around the same set of values no matter what choice we make. But some of the data sets we study here are quite sparse and for these the results can vary with the choice of prior. Our qualitative conclusions remain the same in all cases, but it is worth bearing in mind that the quantitative details can change.

Combining the likelihood and priors, we now have

$$P(\mathbf{s}, \alpha, \beta | \mathbf{A}) = P(\mathbf{A} | \mathbf{s}, \alpha, \beta) \frac{P(\alpha)P(\beta)P(\mathbf{s})}{P(\mathbf{A})}. \quad (17)$$

The prior on \mathbf{A} is unknown but constant, so it can be ignored. We now draw from the distribution $P(\mathbf{s}, \alpha, \beta | \mathbf{A})$ to obtain a representative sample of values $\mathbf{s}, \alpha, \beta$. In our calculations we generate the samples using the Hamiltonian Monte Carlo method [35] as implemented in the probabilistic programming language Stan [36], which is ideal for sampling from continuous parameter spaces such as this. A few thousand samples are typically sufficient to get a good picture of the distribution of α and β .

D. Minimum violations ranking

One special case of our model worth mentioning is the limit $\beta \rightarrow \infty$ for fixed $\alpha > 0$. In this limit the function $f_{\alpha\beta}(s)$ becomes a step function with value

$$f_{\alpha,\infty}(s) = \begin{cases} \frac{1}{2}\alpha & \text{if } s < 0, \\ \frac{1}{2} & \text{if } s = 0, \\ 1 - \frac{1}{2}\alpha & \text{if } s > 0. \end{cases} \quad (18)$$

For this choice the data likelihood becomes

$$P(\mathbf{A} | \mathbf{s}, \alpha, \beta) = \left(\frac{1}{2}\alpha\right)^v (1 - \frac{1}{2}\alpha)^{m-v}, \quad (19)$$

where m is the total number of games/interactions/comparisons and v is the number of “violations,” meaning games where the weaker player won. Then the log-likelihood is

$$\begin{aligned} \log P(\mathbf{A} | \mathbf{s}, \alpha, \beta) &= -v \log \frac{1 - \frac{1}{2}\alpha}{\frac{1}{2}\alpha} + m \log(1 - \frac{1}{2}\alpha) \\ &= -Av - B, \end{aligned} \quad (20)$$

where A and B are positive constants. This log-likelihood is maximized when the number of violations v is minimized, which leads to the so-called *minimum violations ranking*, the ranking such that the minimum number of games are won by the weaker player. Thus the minimum violations ranking can be thought of as the limit of our model in the special case where $\beta \rightarrow \infty$.

IV. RESULTS

We have applied these methods to a range of data sets representing competition in sports and games as well as social hierarchies in both humans and animals. The data sets we study are listed in Table I.

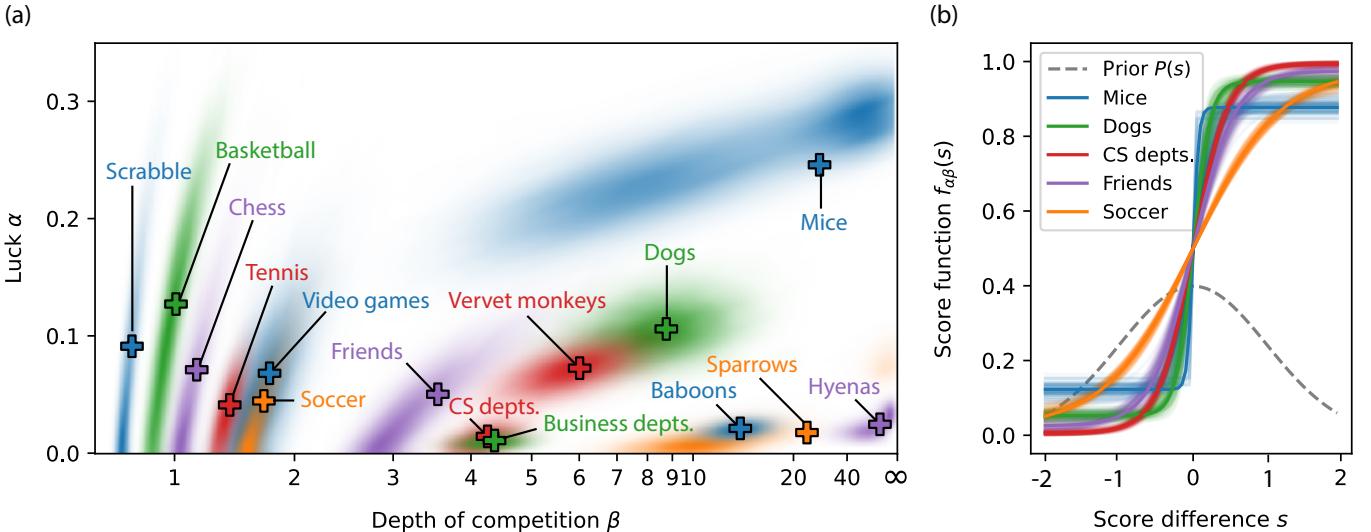


FIG. 2. (a) Each cloud represents the posterior distribution $P(\alpha, \beta | \mathbf{A})$ of the luck and depth parameters for a single data set, calculated from the Monte Carlo sampled values of α and β using a Gaussian kernel density estimate. The + signs indicate the expected values $\hat{\alpha}, \hat{\beta}$ of the parameters for each data set. (b) Fitted functions $f_{\alpha\beta}(s)$ for a selection of the data sets. The bold curve in each case corresponds to the expected values $\hat{\alpha}, \hat{\beta}$, while the other surrounding curves are for a selection of values sampled from the posterior distribution, to give an idea of the variation around the average.

Figure 2a summarizes our results for the posterior probability density of the luck and depth parameters. The axes of the figure indicate the values of α and β and each cloud is a Gaussian kernel density estimate of $P(\alpha, \beta | \mathbf{A})$ computed from the sampled values of α and β for a single data set. The + signs in the figure represent the mean values of α and β for each data set computed directly by averaging the samples.

The figure reveals some interesting trends. Note first that all of the sports and games—chess, basketball, video games, etc.—appear on the left-hand side of the plot in the region of low depth of competition, while all the social hierarchies are on the right with higher depth. We conjecture that the low depth of the sports and games is a result of a preference for matches to be between roughly evenly matched opponents, as discussed in Section III B. For a game to be entertaining to play or watch the outcome of matches should not be too predictable, but in a sport or league with high depth the average pairing is very uneven, with the stronger player very likely to win. Low depth of competition ensures that matches are unpredictable and hence entertaining. In games such as chess, which have high intrinsic depth, the depth can be reduced by restricting tournaments to players in a narrow skill range, such as world-class players, and this is commonly done in many sports and games. We explore this interpretation further in Appendix 5.

There are no such considerations at play in social hierarchies. Such hierarchies are not, by and large, spectator sports, and there is nothing to stop them having high depth of competition. The results in Fig. 2a indicate that in general they do, though the animal hierarchies are deeper than the human ones. A high depth in this

context indicates a hierarchy in which the order of dominance between the typical pair of competitors is clear. This accords with the conventional wisdom concerning hierarchies of both humans and animals, where it appears that participants are in general clear about the rank ordering.

Another distinction that emerges from Fig. 2a is that the results for sports and games generally do not give strong support to a nonzero luck parameter. The expected values, indicated by the + signs, are nonzero in most cases, but the clouds representing the posterior distributions give significant weight to points close to the $\alpha = 0$ line, indicating that we cannot rule out the possibility that $\alpha = 0$ in these competitions. For many of the social hierarchies, on the other hand, there is strong evidence for a nonzero amount of luck, with the posterior distribution having most of its weight well away from $\alpha = 0$.

In part this observation is constrained by the data we have available. It is difficult to distinguish the value of α in a competition with low depth because most matches are fairly evenly balanced—neither player is strongly favored to win. We can also achieve the same outcome by making the luck parameter α large, so that high luck and low depth both give good fits to the data and hence are confounded in the results. This is reflected by the tall shapes of the posterior clouds on the left of Fig. 2a, indicating a high uncertainty about the value of α . In the high-depth region on the right of the figure it is much easier to discern the value of α , and in this region there are many data sets for which we can be quite certain that α is nonzero. This finding of nonzero α also accords with our intuition about social hierarchies. There would be lit-

tle point in having any competition at all within a social hierarchy if the outcomes of all contests were foregone. If all participants knew that every competitive interaction was going to end with the higher-ranked individual winning and the lower-ranked one backing down, then there would be no reason to compete. It is only because there is a significant chance of a win that competition occurs at all.

An interesting counter-example to this observation comes from the two faculty hierarchies, which represent hiring practices at US universities and colleges. The interactions in this data set indicate when one university hires a faculty candidate who received their doctoral training at another university, which is considered a win for the university where the candidate trained. The high depth of competition and low luck parameter for these data sets indicates that there is a pronounced hierarchy of hiring with a clear pecking order and that the pecking order is rarely violated. Lower-ranked universities hire the graduates of higher-ranked ones, but the reverse rarely happens.

Figure 2b shows a selection of the fitted functions $f_{\alpha\beta}(s)$ for five of the data sets. For each data set we show in bold the curve for the expected values $\hat{\alpha}, \hat{\beta}$ along with ten other curves for values of α, β sampled from the posterior distribution, to give an indication of the amount of variation around the average. We see for example that the curve for the soccer data set has a shallow slope (low depth of competition) but is close to zero and one at the limits (low luck). The curve for the mice data set, by contrast, is steep (high depth) but clearly has limits well away from zero and one (nonzero luck).

V. PREDICTING WINS AND LOSSES

In addition to allowing us to infer the luck and depth parameters and rank competitors, our model can also be used to predict the outcomes of unobserved matches. If we fit the model to data from a group of competitors, we can use the fitted model to predict the winner of a new contest between two of those same competitors. The ability to accurately perform such predictions can form the basis for consumer product recommendations and marketing, algorithms for guiding competitive strategies in sports and games, and the setting of odds for betting, among other things.

We can test the performance of our model in this prediction task using a cross-validation approach. For any data set \mathbf{A} we randomly remove or “hold out” a small portion of the matches or interactions and then fit the model to the remaining “training” data set. Then we use the fitted model to predict the outcome of the held-out matches and compare the results with the actual outcomes of those same matches.

The simplest version of this calculation involves fitting our model to the training data by making point estimates of the parameters and scores. We first estimate the ex-

pected posterior values $\hat{\alpha}, \hat{\beta}$ of the parameters given the training data. Then, given these parameter values, we maximize the posterior probability as a function of s to obtain MAP estimates \hat{s} of the scores. Finally, we use the combined parameter values and scores to calculate the probability $\hat{p}_{ij} = f_{\hat{\alpha}\hat{\beta}}(\hat{s}_i - \hat{s}_j)$ that a held-out match between i and j was won by i , with $f_{\alpha\beta}(s)$ as in Eq. (14). Further discussion of the procedure is given in Appendix 3.

We can quantify the performance of our predictions by computing the log-likelihood of the actual outcomes of the held-out matches under the predicted probabilities \hat{p}_{ij} . If W_{ij} is the number of times that i actually won against j then the log-likelihood per game is

$$Q = \frac{\sum_{ij} W_{ij} \log \hat{p}_{ij}}{\sum_{ij} W_{ij}}. \quad (21)$$

This measure naturally rewards cases where the model is confident in the correct answer (\hat{p}_{ij} close to 1) and heavily penalizes cases where the model is confident in the wrong answer (\hat{p}_{ij} close to 0). Note that the log-likelihood is equal to minus the description length of the data—the amount of information needed to describe the true sequence of wins and losses in the held-out data given the estimated probabilities \hat{p}_{ij} —so models with high log-likelihood are more parsimonious in describing the true pattern of wins and losses.

To place the performance of our proposed model in context, we compare it against a basket of other ranking models and methods, including widely used standards, some recently proposed approaches, and some variants of the approach proposed in this paper. As a baseline we compare performance against the standard Bradley-Terry model with a logistic prior, which is commonly used in many ranking tasks, particularly in sports, and which we have ourselves used and recommended in the past [8]. We measure the performance of all other models against this one by calculating the difference in the log-likelihood per match, Eq. (21). The other models we test are:

1. The luck-plus-depth model of this paper.
2. A depth-only variant in which the parameter α is set to zero.
3. A luck-only variant in which the parameter β is set to ∞ , which is equivalent to minimum violations ranking as described in Section III C.
4. The Bradley-Terry model under maximum-likelihood estimation, which is equivalent to imposing an improper uniform prior.
5. The “SpringRank” model of De Bacco *et al.* [37], which ranks competitors using a physically motivated mass-and-spring model.

The proportion of data held out in the cross-validation was 20% in all cases, chosen uniformly at random, and at least 50 random repetitions of the complete process were performed for each model for each of the data sets listed in Table I.

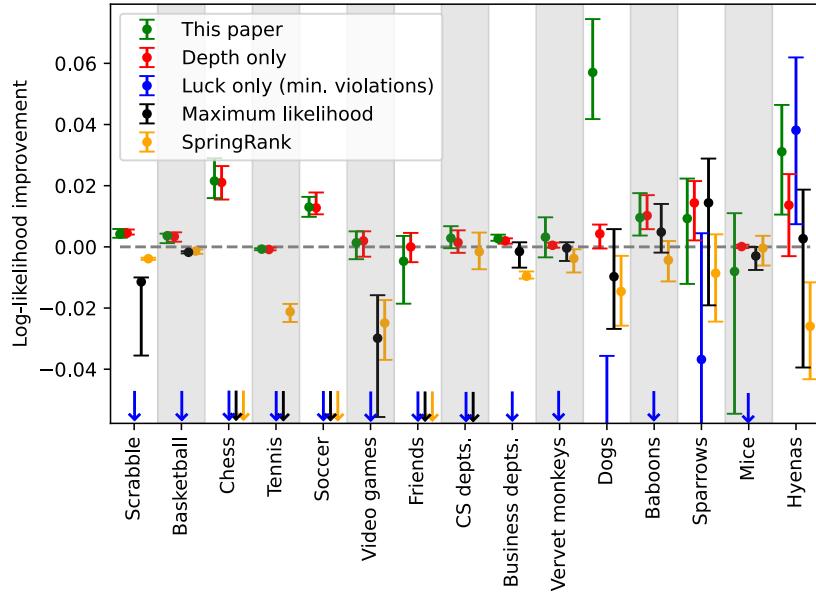


FIG. 3. Comparative performance of the model of this paper and a selection of competing models and methods, in the task of predicting the outcome of unobserved matches in a cross-validation experiment. Performance is measured in terms of the log-likelihood (base 2) of the actual outcomes of matches within the fitted model, which is also equal to minus the description length in bits required to transmit the win/loss data given the fitted model. Log-likelihoods are plotted relative to that of the standard Bradley-Terry model with a logistic prior (the horizontal dashed line). Error bars represent upper and lower quartiles over at least 50 random repetitions of the cross-validation procedure in each case. The arrows along the bottom of the plot indicate cases where the log-likelihood is outside the range of the plot.

The results are summarized in Fig. 3. The horizontal dashed line in the figure represents the baseline set by the Bradley-Terry model and the points with error bars represent the increase (or decrease) in log-likelihood relative to this level for each model and data set. The error bars represent the upper and lower quartiles of variation of the results over the random repetitions. (We use quartiles rather than standard deviations because the distributions are highly non-normal in some cases.)

We note a number of things about these results. First, the model of this paper performs best for every data set without exception, within the statistical uncertainty, although the depth-only version of the model is also competitive in many cases, particularly for the sports and games. The latter observation is unsurprising, since, as we have said, there is little evidence for $\alpha > 0$ in the games. For the particular case of the dominance hierarchy of hyenas, the minimum violations ranking is competitive, which is also unsurprising: as shown in Fig. 2 this hierarchy is very deep—the value of β is over 100—and hence our model and the minimum violations ranking are essentially equivalent. In all the other networks the minimum violations ranking performs worse—usually much worse—than our model. (Arrows at the bottom of the figure indicate results so poor they fall off the bottom of the scale.) The maximum likelihood fit to the Bradley-Terry model also performs quite poorly, a notable observation given that this is one of the most popular ranking algorithms in many settings. It even performs markedly

worse than the same Bradley-Terry model with a logistic prior. Finally, we note that the SpringRank algorithm of [37] is relatively competitive in these tests, though it still falls short of the model of this paper and the standard Bradley-Terry model with logistic prior.

VI. CONCLUSIONS

In this paper we have studied the ranking of competitors based on pairwise comparisons between them, as happens for instance in sports, games, and social hierarchies. Building on the standard Bradley-Terry ranking model, we have extended the model to include two additional features: an element of luck that allows weak competitors to occasionally beat strong ones, and a “depth of competition” parameter that captures the number of distinguishable levels of play in a hierarchy. Deep hierarchies with many levels correspond to complex games or social structures. We have fitted the proposed model to data sets representing social hierarchies among both humans and animals and a range of sports and games, including chess, basketball, soccer, and video games. The fits give us estimates of the luck and depth of competition in each of these examples and we find a clear pattern in the results: sports and games tend to have shallow depth and little evidence of a luck component, while social hierarchies are significantly deeper and more often have an element of luck, with the animal hierarchies being deeper

than the human ones.

We also test our model’s ability to predict the outcome of contests. Using a cross-validation approach we find that the model performs as well as or better than every other model tested in predictive tasks and very significantly better than the most common previous methods such as maximum likelihood fits to the Bradley-Terry model or minimum violations rankings.

ACKNOWLEDGMENTS

The authors thank Elizabeth Bruch, Fred Feinberg, and Dan Larremore for useful conversations. This work was funded in part by the US National Science Foundation under grant DMS–2005899. All empirical data used in this paper are freely available online or from their original authors.

APPENDICES

1. Data sets

The example data sets used in this paper are summarized in Table I of the main paper and divide into three broad categories: sports and games (six data sets), human social hierarchies (three data sets), and animal social hierarchies (six data sets). Here we provide some additional details on these data.

Sports and games: We consider both team competition (basketball, soccer) and individual competition (chess, Scrabble, tennis, video games). For the team sports we treat each team in each year as a different entity with its own assigned score s_i . Thus, for example, the England soccer team in 2015 is considered a different entity from the England soccer team in 2014. This reflects the fact that the composition of teams can change from season to season and with it the ranking of the team in comparison to others.

Two of the game data sets, for chess and Scrabble, were too large in their original form to perform our full Bayesian analysis in a reasonable amount of time, so they were subsampled to reduce them to manageable size. We limited the chess data set to only those players who had participated in at least 200 games and then randomly selected 5% of those players. All others were removed from the data set. The scrabble data set was similarly pared down by limiting it to players who had at least 100 games and then choosing a random 20% of those who remained.

Another issue with some of the game data is the presence of ties, which occur with moderate frequency in both chess and soccer. Although there do exist ranking models that allow for ties [12, 13], we avoid these in the present work for the sake of simplicity, and all our models assume that the only possible outcomes of a match are a win or a loss. To accommodate the chess and soccer

data within this setting we remove all ties from the data, which amounts to 10–30% of matches in those data sets.

Human social hierarchies: A related issue arises in the “friends” data set, which details friend nominations among students in a US middle/high school. A significant fraction of such nominations are reciprocal—two individuals each nominate the other as a friend [38, 39]. Such reciprocated nominations have been treated as ties in some previous analyses [8], but here again we simply remove them. Only unreciprocated friendships are recorded as a win for the person who receives the nomination.

Animal hierarchies: Data on animal dominance hierarchies is copious: this has been an active field of research for at least sixty years. The data sets studied in this paper come from a variety of sources, but particularly from DomArchive, a collection of 436 dominance interaction data sets compiled by Strauss *et al.* [40]. Data sets in the archive vary widely in size, but the sets we focus on are ones with a relatively large number of interactions per individual, which improves the statistics and helps reduce uncertainty on the fitted values of the model parameters.

2. Cross-validation

In the cross-validation results reported in the main paper we quantify predictive performance of the various models by calculating the log-likelihood of the testing (held-out) data within the fitted model—see Fig. 3. This is not, however, the only way to measure performance; there are a number of other approaches in common use. In this appendix we describe some alternative performance metrics and investigate how our models size up when measured by these metrics. In general the results are similar to those presented in the main paper, but there are some differences in the details.

A simple way to quantify the predictive performance of a model is to count the number of times the model predicts the correct winner in the test data. As before, we start by fitting the model to the training portion of the data to obtain MAP estimates \hat{s} of the scores, then, given those estimates, player i is considered favored to beat player j if $\hat{s}_i > \hat{s}_j$. The *accuracy* C of the model is defined to be the fraction of matches in the testing data where this prediction is born out:

$$C = \frac{\sum_{ij} W_{ij} \mathbf{1}_{\hat{s}_i > \hat{s}_j}}{\sum_{ij} W_{ij}} \quad (22)$$

where W_{ij} is the number of times i beats j in the testing data, as previously, and $\mathbf{1}_x$ is the indicator function which is 1 if x is true and 0 otherwise.

Values of this accuracy measure are shown in Fig. 4a for each of the models considered in this paper for each of our data sets. As with our previous results for log-likelihood, we report performance relative to a baseline

set by the standard Bradley-Terry model with a logistic prior, represented by the horizontal dashed line in the figure. Comparing with our earlier results from Fig. 3, the difference between models is smaller when measured in terms of accuracy than log-likelihood. For example, the minimum violations ranking performs quite poorly according to the log-likelihood, but is comparable and sometimes better than our models in terms of accuracy. This may be because the minimum violations ranking is more directly tuned to solving this specific problem: by minimizing violations we precisely minimize the number of outcomes that are predicted incorrectly. On the other hand, the minimum violations algorithm does not reflect how confident we are in each outcome or any other aspect of the prediction task, and in this sense is inferior to other approaches.

Both the likelihood and accuracy measures are based on point estimates of model parameters \hat{s} , $\hat{\alpha}$, and $\hat{\beta}$ but, as shown in Fig. 2, point estimates do not always do a good job of capturing the full posterior distribution $P(\mathbf{s}, \alpha, \beta | \mathbf{A}_{\text{train}})$, particularly in sparse data sets. To get around this issue, we can calculate the average of the likelihood over the distribution of parameter values thus:

$$\begin{aligned} P(\mathbf{A}_{\text{test}} | \mathbf{A}_{\text{train}}) \\ = \int P(\mathbf{A}_{\text{test}} | \mathbf{s}, \alpha, \beta) P(\mathbf{s}, \alpha, \beta | \mathbf{A}_{\text{train}}) d^n \mathbf{s} d\alpha d\beta. \end{aligned} \quad (23)$$

In practice, this quantity can be estimated from a set of N samples of $(\mathbf{s}_k, \alpha_k, \beta_k)$ (with $k = 1 \dots N$) drawn from the posterior $P(\mathbf{s}, \alpha, \beta | \mathbf{A}_{\text{train}})$, as the average

$$P(\mathbf{A}_{\text{test}} | \mathbf{A}_{\text{train}}) \simeq \frac{1}{N} \sum_{k=1}^N P(\mathbf{A}_{\text{test}} | \mathbf{s}_k, \alpha_k, \beta_k). \quad (24)$$

We can calculate this estimate from the same Monte Carlo samples we already generated, which we used previously to visualize the posterior distribution in Fig. 2. As our measure of performance we then compute the *log-posterior-predictive probability* per game

$$R = \frac{\log P(\mathbf{A}_{\text{test}} | \mathbf{A}_{\text{train}})}{\sum_{ij} W_{ij}}, \quad (25)$$

a fully Bayesian performance measure.

We plot this measure for a number of our models and data sets in Fig. 4b. Note, however, that since the measure involves an integral over the posterior distribution of the scores, we cannot apply it to ranking methods that return only point estimates of the scores rather than a full probability distribution, which in this case means the Bradley-Terry MLE and SpringRank, which are thus excluded from the figure. Among the remaining methods the full luck-plus-depth model of this paper performs best, or equal-best, for every data set, by this measure.

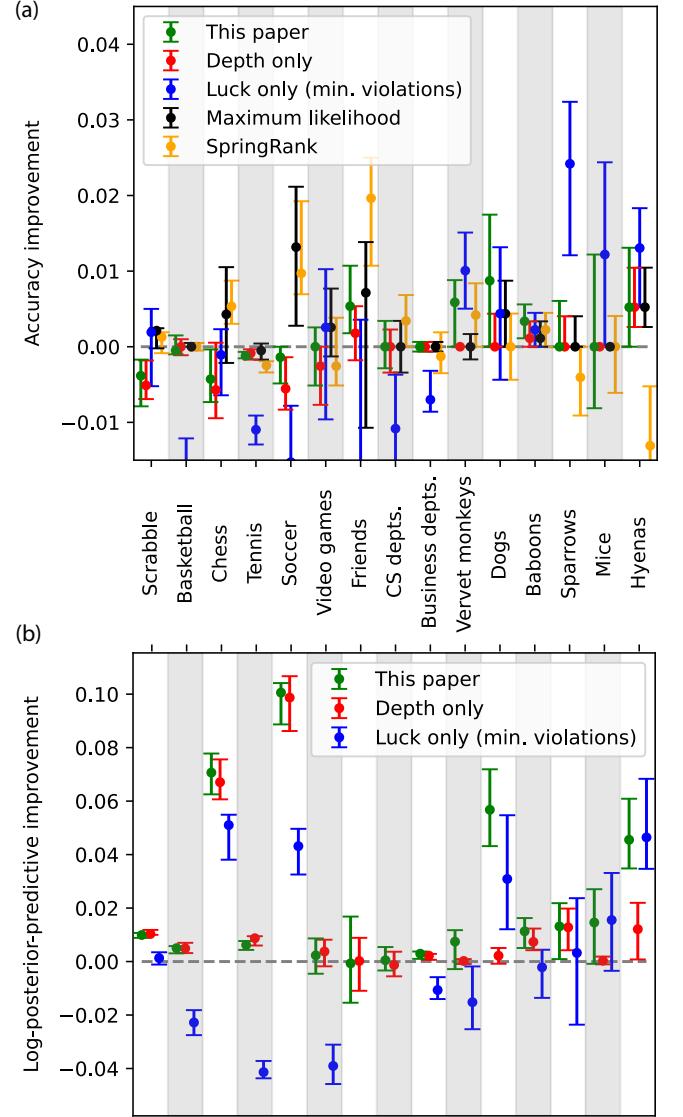


FIG. 4. Results from the same set of cross-validation tests shown in Fig. 3, but quantified using (a) accuracy and (b) log-posterior predictive probability, instead of log-likelihood. All results are measured relative to the Bradley-Terry model with a logistic prior, which is represented as the dashed horizontal line in each panel. Error bars represent upper and lower quartiles, estimated from at least 50 random repetitions of the cross-validation procedure in each case. The maximum likelihood and SpringRank models are not included in the lower comparison, since they are based on point estimates rather than Bayesian methods and hence one cannot calculate a posterior-predictive probability.

3. Point estimates of parameters

To compute the log-likelihood and accuracy measures of predictive success we use point estimates of the model parameters and scores, which we compute one after the other: we estimate the expected posterior values of the parameters $\hat{\alpha}$, $\hat{\beta}$ from a simple average of the Monte

Carlo samples, then we fix these values and compute the MAP values of the scores \hat{s} using a standard numerical optimization method. We could, alternatively, use the expected values of the scores, which would be easy to calculate from the samples, but we prefer MAP values since they give a more appropriate point of comparison with other approaches based on maximum probability estimates, such as the maximum likelihood fit to the Bradley-Terry model or the SpringRank algorithm.

One might imagine one could simplify the calculation by just jointly optimizing the posterior $P(\mathbf{s}, \alpha, \beta | \mathbf{A})$ over both the scores and parameters to define estimates

$$(\mathbf{s}^*, \alpha^*, \beta^*) \equiv \operatorname{argmax}_{\mathbf{s}, \alpha, \beta} P(\mathbf{s}, \alpha, \beta | \mathbf{A}). \quad (26)$$

We find, however, that this can give biased results by artificially inflating the value of the depth parameter β . This happens because the likelihood $P(\mathbf{A}|\mathbf{s}, \alpha, \beta)$ is a function of the product βs (see Eq. (14)), meaning that the value of the likelihood is unchanged if we increase β while simultaneously reducing all the scores by the same factor. Reducing \mathbf{s} in this way increases the prior $P(\mathbf{s})$ (which is peaked at $\mathbf{s} = 0$) and so increases the posterior $P(\mathbf{s}, \alpha, \beta | \mathbf{A})$. Unchecked, this effect would send the joint maximum to $\beta^* \rightarrow \infty, \mathbf{s} \rightarrow 0$. The prior $P(\beta)$ somewhat mitigates this problem, but in practice the jointly fitted value β^* is still unreasonably large: values for each of the data sets are shown in Table II.

4. Other measures of depth

In this paper we measure depth of competition by the parameter β in our joint luck-plus-depth model, Eq. (14). This is not the only possible approach for quantifying depth, however, and in this appendix we discuss some alternative approaches and explain how they relate to similar ideas presented elsewhere.

As discussed in Section III B, our depth measure β counts the number of “levels of skill” between two typical players in a population, who in expectation have a priori score difference $s_i - s_j = 1$ (because of our choice of prior on s). An alternative, and common, way to define depth is as the number of levels between not the typical pair of players but the best and worst players, which is given by

$$\hat{\beta}_{\text{range}} = \hat{\beta}(\hat{s}_{\max} - \hat{s}_{\min}). \quad (27)$$

In the data sets studied here we find that the factor $\hat{s}_{\max} - \hat{s}_{\min}$ varies from about 2.5 to 4. The range tends to be larger when there are more competitors, presumably because outliers are more likely in large samples, and we regard this as downside of this measure, although in practice the depth order of our data sets does not change significantly between this measure and our own. Values of $\hat{\beta}_{\text{range}}$ are reported in Table II for each of the data sets.

Our depth measure β is defined in the context of our full luck-plus-depth model, but in many cases, particularly for the sports data sets, there is no strong evidence

of a nonzero luck parameter α . An alternative approach for quantifying depth in these cases is to use a depth-only model as in Eq. (9). Depth values calculated by fitting this model are given in Table II and denoted β_0 , which we refer to as “restricted depth.” In practice these figures are not very different for those for β in cases (such as sports) where the value of α is small anyway, or more precisely when the posterior distribution in Figure 2 meaningfully intersects the $\alpha = 0$ axis, so that the zero-luck model is plausible. On the other hand, β and β_0 can differ substantially when the data support a significantly nonzero value of α . For example, the mice data set has an expected value of α around 0.25 with a posterior distribution that has considerable separation from $\alpha = 0$, and in this case we find a large difference between a value of $\hat{\beta} = 26.5$ and $\hat{\beta}_0 = 2.1$, the latter being more akin to the sports data than to the other animal hierarchies.

The restricted depth β_0 is closer in spirit to previous measures of depth that do not consider the element of luck, and the occurrence of large discrepancies with the value of β in some data sets suggests that such previous measures might potentially be in error by a significant margin. For applications where the element of luck is not an issue, however, the restricted depth could be useful as a simplification of our measure. It can be calculated relatively straightforwardly, to a good approximation, using the standard Bradley-Terry model with a logistic prior, a model we have recommended in the past. In our current analysis we have used Gaussian priors, but the logistic prior has some practical advantages in that it enables simple and fast iterative methods for computing MAP scores. In the most common version of this approach, one uses the unit logistic distribution $1/[(1+e^s)(1+e^{-s})]$ as prior with the standard ($\beta = 1$) Bradley-Terry model, which leads to an elegant iterative algorithm for calculating the scores [8]. The logistic prior, however, has variance $\frac{1}{3}\pi^2$, whereas our Gaussian prior has variance $\frac{1}{2}$, so, though the qualitative shape of the two distributions is similar, the logistic distribution has substantially greater width, by a factor of $\pi\sqrt{2/3}$. An alternative way to perform the same calculation is to shrink the width of the prior to be the same as the Gaussian, while simultaneously shrinking the width of the Bradley-Terry score function by the same factor, which is equivalent to choosing $\beta = \pi\sqrt{2/3} = 2.565$. This leaves the algorithm, and the resulting ranking, unchanged, and thus the iterative method with a logistic prior is equivalent to the depth-only model with $\beta = 2.565$.

Happily, this choice of β falls squarely in the middle of the range of values seen in Fig. 2 and in practice this approach has quite competitive performance, as shown in Fig. 3, where it is used as the baseline. On the other hand, there are plenty of cases where the value $\beta = 2.565$ is clearly misspecified, which is signaled by fitted scores whose variance does not match the width of the prior. This observation suggests that we could use the spread of the fitted scores as a heuristic measure of (restricted) depth and in practice this approach seems to work quite

well. Quantifying the spread by its standard deviation, we report figures for each of our data sets in Table II, and we find that there is good correlation between this standard deviation and the restricted depth $\hat{\beta}_0$ as calculated earlier. Given that the former is significantly easier to calculate than the latter, this could be a useful approach for calculations where accuracy and rigor are not at a premium.

A quite different approach to measuring depth has been developed in the animal behavior literature, where the notion of “steepness” has gained currency in discussions of dominance hierarchies [20]. Steepness is most often defined through quantities known as “David’s scores,” which are measures of individual performance analogous to our fitted s_i [1]. The David’s scores are defined as

$$DS_i = w_i + \sum_j w_j P_{ij} - l_i - \sum_j l_j P_{ji} \quad (28)$$

where P_{ij} is the fraction of times that i beats j :

$$P_{ij} = \frac{A_{ij}}{A_{ij} + A_{ji}}, \quad (29)$$

and w_i and l_i are row and column sums of this matrix:

$$w_i = \sum_j P_{ij}, \quad l_i = \sum_j P_{ji}. \quad (30)$$

De Vries *et al.* [20] propose normalizing the David’s scores according to

$$\text{NormDS}_i = \frac{DS_i + \binom{n}{2}}{n}, \quad (31)$$

which vary between 0 and $n - 1$, then the animals are ranked according to the resulting values. With the inferred rank order on the x -axis and the normalized David’s score on the y -axis, the steepness of the hierarchy is then defined to be the slope S_{DS} of the ordinary line of best fit. A nice feature of this formulation is that the steepness runs from 0 to 1, with the value 1 being achieved in any hierarchy where all dominance interactions run from higher ranked to lower ranked individuals (zero violations).

Neumann and Fischer [41] have recently proposed a related measure that considers the slope S_{Elo} of the line of best fit between Elo scores for the competitors and their inferred ordinal ranking. Elo scores are essentially a sequential (time-dependent) version of a maximum likelihood fit to the Bradley-Terry model and so this definition is closer to the ideas considered in this paper. Neumann and Fischer also incorporate Bayesian elements where certain aspects of the fitting process are randomized, such as the sequential order (if the true order is unknown) and the initial values of the ratings.

In Table II we report values for a number of our data sets of S_{DS} (calculated using the R package **steepness** [42]) and S_{Elo} (calculated using the R package

EloSteepness [43]). Overall, we find that the results are clearly correlated with the other measures shown in the table, although S_{DS} has trouble differentiating between the lower depth data sets. The Elo-based steepness S_{Elo} fares better and correlates quite well with the restricted depth $\hat{\beta}_0$, although the calculations are computationally demanding on account of the randomization and prove intractable for our larger data sets (as indicated by “–” in the table).

To complete our collection of measures of depth we also include in Table II the parameter β_S that appears in the SpringRank model [37]. This parameter has not previously been used as a measure of depth but one can make an argument for its use in this way—see Appendix 6.

Finally, we note in passing that there is an analogy between the depth parameter β and a notion of “temperature” for a data set. The form of the score function of Eq. (9) is precisely that of the Fermi-Dirac probability function of many-body physics, the probability of occupation at inverse temperature β of an energy level with energy s above the Fermi level. While we have not directly exploited this analogy here, it is a part of a broader correspondence between noise and unpredictability in statistics and temperature in physics.

5. Depth as predictability

In Section IV we observed that among our data sets the sports and games have lower depth compared to the social hierarchies, and we speculated that this was because a high-depth sport would not be as interesting to watch: at high depth a typical pair of competitors will be very unevenly matched and there will be little suspense about who is going to win. In other words, high depth should result in high predictability of outcomes. In this appendix we test this hypothesis by calculating various measures of predictability.

A natural measure of predictability is the same log-likelihood that we studied in Section V. The log-likelihood of a data set is equal to minus the description length of the outcomes of the matches in that set, given the fitted model. That is, it is equal to the amount of information it would take to communicate the outcomes to a receiver who already knows the fitted model. Higher information (more negative log-likelihood) implies more unpredictable outcomes. Completely random outcomes (matches decided by the toss of a coin) would give a log-likelihood of -1 per match (in log-base-2 units), while completely predictable ones would give zero.

Previously, we plotted the log-likelihood relative to the baseline set by the standard Bradley-Terry model, but in the present context we are interested in the absolute value. Figure 5 shows the absolute value for each of our data sets, arranged in order of increasing depth β . As the figure shows, the low-depth sports on the left are indeed quite unpredictable and none of our models perform much better than chance at predicting outcomes (log-

	Data set	$\hat{\beta}$	β^*	$\hat{\beta}_{\text{range}}$	$\hat{\beta}_0$	$\text{std}(\hat{s}_L)$	S_{DS}	S_{Elo}	$\hat{\beta}_S$	Luck	
										$\hat{\alpha}$	α^*
Sports/games	Scrabble	0.68	3.13	2.43	0.60	0.64	0.00	—	2.24	0.09	0.00
	Basketball	1.01	10.79	3.66	0.83	0.61	0.01	0.48	2.32	0.13	0.02
	Chess	1.17	4.73	4.21	1.04	0.91	0.00	—	2.85	0.07	0.12
	Tennis	1.44	1.98	5.88	1.34	0.72	0.00	—	2.67	0.04	0.00
	Soccer	1.73	6.23	4.97	1.58	1.02	0.00	—	4.00	0.04	0.00
	Video games	1.77	17.53	5.12	1.55	1.10	0.02	0.62	2.95	0.07	0.05
Human	Friends	3.54	10.36	9.88	2.80	1.16	0.00	—	5.23	0.05	0.00
	CS departments	4.25	15.42	12.11	3.88	1.88	0.01	0.78	4.46	0.01	0.00
	Business depts.	4.36	13.72	11.73	4.07	2.25	0.14	0.84	4.07	0.01	0.01
Animal	Vervet monkeys	6.01	30.39	17.07	3.57	2.23	0.40	0.85	4.34	0.07	0.07
	Dogs	8.74	33.29	24.82	3.76	2.03	0.25	0.93	3.65	0.11	0.09
	Baboons	13.19	18.61	39.04	9.37	4.38	0.05	0.95	5.63	0.02	0.02
	Sparrows	22.92	63.89	69.68	8.68	3.62	0.50	0.91	7.72	0.02	0.01
	Mice	26.48	59.48	72.29	2.10	1.35	0.31	0.72	3.22	0.25	0.24
	Hyenas	100.58	168.48	246.42	9.83	4.00	0.30	0.95	8.15	0.02	0.02

TABLE II. Inferred parameter values for the data sets considered in Section IV. From left to right: $\hat{\beta}$ is expected depth, β^* is the jointly optimized MAP depth as in Eq. (26), $\hat{\beta}_{\text{range}}$ is depth between the best and worst player as in Eq. (27), $\hat{\beta}_0$ is restricted depth as inferred in the depth-only ($\alpha = 0$) model, $\text{std}(\hat{s}_L)$ is the standard deviation of the MAP scores within the logistic-prior model, S_{DS} is the steepness measure of de Vries *et al.* [20], S_{Elo} is the steepness measure of Neumann and Fischer [41], $\hat{\beta}_S$ is the maximum likelihood estimate of the parameter β_S in the SpringRank model [37], $\hat{\alpha}$ is the expected luck, and α^* is the jointly optimized MAP estimate of the luck.

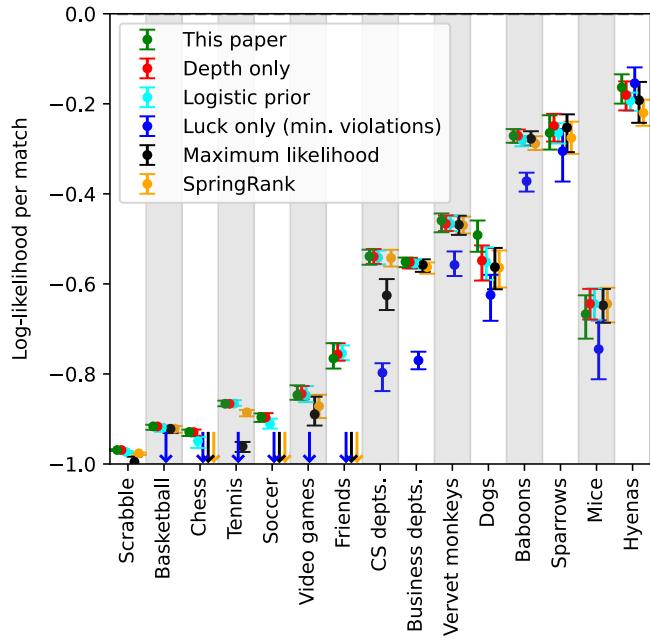


FIG. 5. Absolute log-likelihood values per match in the cross-validation tests of Fig. 3. This figure differs from Fig. 3 in showing absolute values rather than values relative to the Bradley-Terry model with logistic prior.

likelihood per match is close to -1). As depth increases, however, outcomes generally become more predictable, and the deepest animal hierarchies have a log-likelihood approaching zero, meaning outcomes are nearly perfectly predictable.

There are some exceptions to this trend, most notably

the mice data set which, as seen in Fig. 2, has a large element of luck ($\hat{\alpha} \simeq 0.25$). This introduces substantial randomness into the matches, despite the high depth, and greatly decreases predictability.

We can shed further light on predictability by calculating the average amount of information needed to describe matches that are truly drawn from our model. That is, we consider two players whose scores s_i are drawn from our normal prior with variance $\frac{1}{2}$, so that the difference of their scores is normally distributed with variance 1, and we assume that the probability of i beating j is given exactly by $p_{ij} = f_{\alpha\beta}(s_i - s_j)$, Eq. (14), for some values of α and β that we specify. Then the average information needed to describe the outcome of the match is given by the standard entropy function for a Bernoulli random variable

$$H[p_{ij}] = -p_{ij} \log p_{ij} - (1 - p_{ij}) \log(1 - p_{ij}). \quad (32)$$

Then, writing $s = s_i - s_j$ and integrating, the average entropy per match over matches between many random pairs of players is

$$S_{\alpha\beta} = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} H[f_{\alpha\beta}(s)] e^{-s^2/2} ds. \quad (33)$$

Unfortunately, this integral does not seem to have a closed-form solution, but it can be evaluated numerically. Figure 6 shows a modified version of Fig. 2 from the main paper, representing the posterior probability distribution of α, β for our various data sets, with superimposed lines representing the contours of the average entropy. As the figure shows, the entropy is higher for lower depth and for higher luck, as we would expect, since both increase the unpredictability of outcomes. We also note that the

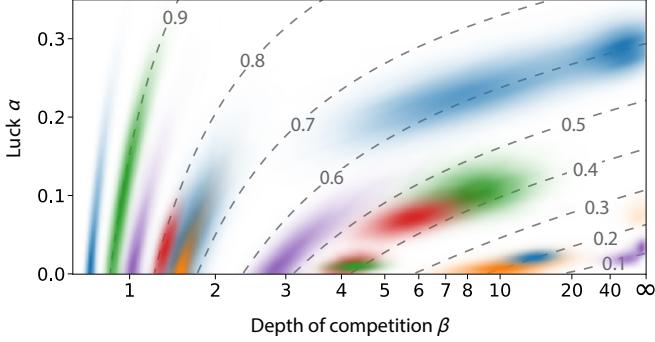


FIG. 6. The data sets of Fig. 2 with dashed lines representing the contours of average entropy per match. Low entropy indicates confidence about the outcome of a match; high entropy indicates unpredictability.

posterior distributions of individual data sets appear to follow the contour lines quite closely, arcing upward and to the right. This occurs because the entropy is by definition equal to minus the log-likelihood, and our prior on α and β is slowly varying by construction, so the posterior is also slowly varying along the contour lines of constant likelihood. The contour lines are calculated as averages over outcomes drawn from the fitted model, whereas the probability clouds in the figure represent real-world data, so the two are not precisely comparable. But to the extent that the data are well described by the model we would expect them to agree and hence for the clouds to follow the contours in the plot. This also means that, while some of the clouds in the figure are quite extended, indicating substantial uncertainty about the values of α and β , they are narrow in the direction perpendicular to the contours, meaning that we have high confidence about the value of the log-likelihood. This is reflected in Fig. 5, where we see that the uncertainty on our estimates of the log-likelihood is quite modest.

6. SpringRank

Among the various approaches to ranking considered in this paper, SpringRank [37] is a recent and novel approach based on a physical analogy to the behavior of a network of masses and springs. In this appendix we make some observations on the method and how it relates to the Bradley-Terry model, which forms the foundation for the other methods we consider.

In SpringRank the likelihood of observing a directed network \mathbf{A} is given by a product of Poisson distributions over all possible directed edges:

$$P(\mathbf{A}|\mathbf{s}, \beta_S, c) = \prod_{ij} \frac{r_{ij}^{A_{ij}}}{A_{ij}!} e^{-r_{ij}}, \quad (34)$$

with the expect number of directed edges $i \rightarrow j$ given by

$$r_{ij} = ce^{-\frac{1}{2}\beta_S(s_i - s_j - 1)^2}, \quad (35)$$

for given scores \mathbf{s} , inverse temperature β_S , and a “sparsity” parameter c . Equation (34) can be rewritten as

$$\begin{aligned} P(\mathbf{A}|\mathbf{s}, \beta_S, c) &= \prod_{i < j} \frac{r_{ij}^{A_{ij}}}{A_{ij}!} e^{-r_{ij}} \frac{r_{ji}^{A_{ji}}}{A_{ji}!} e^{-r_{ji}} \\ &= \prod_{i < j} \frac{(r_{ij} + r_{ji})^{A_{ij} + A_{ji}} e^{-(r_{ij} + r_{ji})}}{(A_{ij} + A_{ji})!} \\ &\quad \times \frac{(A_{ij} + A_{ji})!}{A_{ij}! A_{ji}!} \left(\frac{r_{ij}}{r_{ij} + r_{ji}} \right)^{A_{ij}} \left(\frac{r_{ji}}{r_{ij} + r_{ji}} \right)^{A_{ji}} \\ &= \prod_{i < j} \frac{m_{ij}^{\bar{A}_{ij}} e^{-m_{ij}}}{\bar{A}_{ij}!} \\ &\quad \times \left(\frac{\bar{A}_{ij}}{A_{ij}} \right) \frac{1}{[1 + e^{-2\beta_S(s_i - s_j)}]^{A_{ij}} [1 + e^{-2\beta_S(s_j - s_i)}]^{A_{ji}}}, \end{aligned} \quad (36)$$

where $m_{ij} = r_{ij} + r_{ji}$ and $\bar{A}_{ij} = A_{ij} + A_{ji}$ is an element of the adjacency matrix $\bar{\mathbf{A}}$ of the undirected network of matches.

Equation (36) is equal to the likelihood of generating an undirected network $\bar{\mathbf{A}}$ of matches and then separately choosing the directions of the edges, i.e., the winners of the matches:

$$P(\mathbf{A}|\mathbf{s}, \beta_S, c) = P(\bar{\mathbf{A}}|\mathbf{s}, \beta_S, c) P(\mathbf{A}|\mathbf{s}, \beta_S, \bar{\mathbf{A}}), \quad (37)$$

where the probability of the undirected network is another product of Poisson distributions:

$$P(\bar{\mathbf{A}}|\mathbf{s}, \beta_S, c) = \prod_{i < j} \frac{m_{ij}^{\bar{A}_{ij}} e^{-m_{ij}}}{\bar{A}_{ij}!} \quad (38)$$

and

$$\begin{aligned} P(\mathbf{A}|\mathbf{s}, \beta_S, \bar{\mathbf{A}}) &= \prod_{i < j} \left(\frac{\bar{A}_{ij}}{A_{ij}} \right) \frac{1}{[1 + e^{-2\beta_S(s_i - s_j)}]^{A_{ij}} [1 + e^{-2\beta_S(s_j - s_i)}]^{A_{ji}}}. \end{aligned} \quad (39)$$

(It is straightforward to confirm that the latter is correctly normalized for $A_{ij} = 0 \dots \bar{A}_{ij}$ and $A_{ji} = \bar{A}_{ij} - A_{ij}$.)

But Eq. (39) is identical to the likelihood for the model studied in this paper, Eqs. (3) and (14), with $\alpha = 0$ and $\beta = 2\beta_S$. (The binomial coefficient accounts for the number of ways of assigning directions A_{ij} to the \bar{A}_{ij} undirected edges.) This observation suggests that we might use β_S as a measure of the (restricted) depth of a hierarchy, and indeed we observe a correlation between the maximum likelihood value $\hat{\beta}_S$ and our own restricted depth parameter β_0 , as shown in Table II.

However, it is the other term, Eq. (38), that particularly distinguishes SpringRank from the other models we have considered. This term, which measures the likelihood that the set of observed matches occurs at all,

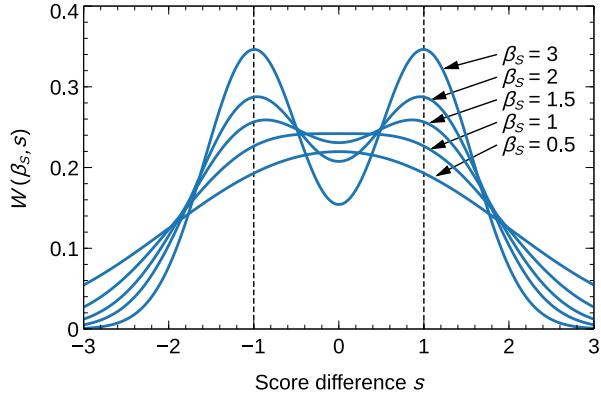


FIG. 7. The function $W(\beta_S, s)$ of Eq. (41) plotted against s , for various values of β_S as indicated.

has no equivalent in the Bradley-Terry model and related models. The quantity m_{ij} , which is the expected number of matches between i and j , can be rewritten in the form

$$m_{ij} = M \frac{W(\beta_S, s_i - s_j)}{\sum_{i < j} W(\beta_S, s_i - s_j)}, \quad (40)$$

where

$$W(\beta_S, s) = \sqrt{\frac{\beta_S}{8\pi}} [e^{-\frac{1}{2}\beta_S(s-1)^2} + e^{-\frac{1}{2}\beta_S(s+1)^2}]. \quad (41)$$

(Note that $W(\beta_S, s)$ is symmetric in s so the sign of the

score difference in Eq. (40) has no effect.) In this formulation the parameter M controls the total number of (undirected) edges in the network and the (properly normalized) probability density $W(\beta_S, s_i - s_j)$ controls how they are distributed given the scores s_i . Figure 7 shows the form of $W(\beta_S, s)$ for various choices of β_S . For $\beta_S \leq 1$ there is a single peak at $s = 0$ so that interactions are preferentially between evenly matched players, but above $\beta_S = 1$ the function becomes bimodal and increasingly peaked around $s = \pm 1$, so that players with a score difference near 1 are more likely to interact.

It is arguably a disadvantage of the SpringRank model that the same parameter β_S controls both the depth of competition via Eq. (39) and the distribution of matches via Eq. (40). Conceptually these are separate processes, and one could make an argument for a model in which they were controlled by separate parameters, although we have not taken that approach here—we use the model as originally defined for the sake of consistency.

In our cross-validation tests we use the maximum likelihood point estimate for the value of β_S , in keeping with the other models we study. We note, however, that De Bacco *et al.* [37], in their original work on SpringRank, used different values of β_S depending on whether the results were scored using log-likelihood or accuracy, choosing in each case the value that gave the best performance according to the measure used.

Finally, we note that the original specification of the SpringRank model also included an optional Gaussian prior on the scores. We have not adopted this prior in our tests, since we find that it tends to diminish the performance of the method.

-
- [1] H. A. David, *The Method of Paired Comparisons*. Griffin, London, 2nd edition (1988).
 - [2] M. Cattelan, Models for paired comparison data: A review with emphasis on dependent data. *Statistical Science* **27**, 412–433 (2012).
 - [3] A. N. Langville and C. D. Meyer, *Who's #1? The Science of Rating and Ranking*. Princeton University Press, Princeton (2013).
 - [4] R. A. Bradley and M. E. Terry, Rank analysis of incomplete block designs: I. The method of paired comparisons. *Biometrika* **39**, 324–345 (1952).
 - [5] E. Zermelo, Die Berechnung der Turnier-Ergebnisse als ein Maximumproblem der Wahrscheinlichkeitsrechnung. *Mathematische Zeitschrift* **29**, 436–460 (1929).
 - [6] L. R. Ford, Jr., Solution of a ranking problem from binary comparisons. *American Mathematical Monthly* **64**(8), 28–33 (1957).
 - [7] D. R. Hunter, MM algorithms for generalized Bradley-Terry models. *Annals of Statistics* **32**, 384–406 (2004).
 - [8] M. E. J. Newman, Efficient computation of rankings from pairwise comparisons. *Journal of Machine Learning Research* **24**, 238 (2023).
 - [9] J. T. Whelan, Prior distributions for the Bradley-Terry model of paired comparisons. Preprint arXiv:1712.05311 (2017).
 - [10] R. R. Davidson and D. L. Solomon, A Bayesian approach to paired comparison experimentation. *Biometrika* **60**, 477–487 (1973).
 - [11] F. Caron and A. Doucet, Efficient Bayesian inference for generalized Bradley-Terry models. *Journal of Computational and Graphical Statistics* **21**, 174–196 (2012).
 - [12] P. V. Rao and L. L. Kupper, Ties in paired-comparison experiments: A generalization of the Bradley-Terry model. *Journal of the American Statistical Association* **62**, 194–204 (1967).
 - [13] R. R. Davidson, On extending the Bradley-Terry model to accommodate ties in paired comparison experiments. *Journal of the American Statistical Association* **65**, 317–328 (1970).
 - [14] R. D. Luce, *Individual Choice Behavior: A Theoretical Analysis*. John Wiley, New York (1959).
 - [15] R. L. Plackett, The analysis of permutations. *Journal of the Royal Statistical Association C* **24**, 193–202 (1975).
 - [16] A. Agresti, *Categorical Data Analysis*. Wiley, New York (1990).
 - [17] M. E. J. Newman, Ranking with multiple types of pairwise comparisons. *Proc. R. Soc. London A* **478**, 20220517 (2022).

- [18] W. Robertie. *Inside Backgammon* **2**(1), 3–4 (1980).
- [19] M.-L. Cauwet, O. Teytaud, H.-M. Liang, S.-J. Yen, H.-H. Lin, I.-C. Wu, T. Cazenave, and A. Saffidine, Depth, balancing, and limits of the Elo model. *Proceedings of the 2015 IEEE Conference on Computational Intelligence and Games 2015* (2015).
- [20] H. de Vries, J. M. G. Stevens, and H. Vervaecke, Measuring and testing the steepness of dominance hierarchies. *Animal Behaviour* **55**, 585–592 (2006).
- [21] Scrabble tournament records. <https://www.cross-tables.com/>. Accessed: 2023-10-07.
- [22] N. Lauga, NBA games data. <https://www.kaggle.com/datasets/nathanlauga/nba-games/data>. Accessed: 2023-10-07.
- [23] Online chess match data from lichess.com. <https://www.kaggle.com/datasets/arevel/chess-games>. Accessed: 2023-10-07.
- [24] J. Sackmann, ATP tennis data. https://github.com/JeffSackmann/tennis_atp. Accessed: 2023-10-07.
- [25] M. Jürisoo, International men’s football results from 1872 to 2023. <https://www.kaggle.com/datasets/martj42/international-football-results-from-1872-to-2017>. Accessed: 2023-10-07.
- [26] Super Smash Bros. Melee head to head records. <https://etossed.github.io/rankings.html>. Accessed: 2023-10-07.
- [27] J. R. Udry, P. S. Bearman, and K. M. Harris, National Longitudinal Study of Adolescent Health (1997). This research uses data from Add Health, a program project directed by Kathleen Mullan Harris and designed by J. Richard Udry, Peter S. Bearman, and Kathleen Mullan Harris at the University of North Carolina at Chapel Hill, and funded by grant P01-HD31921 from the Eunice Kennedy Shriver National Institute of Child Health and Human Development, with cooperative funding from 23 other federal agencies and foundations. Special acknowledgment is due Ronald R. Rindfuss and Barbara Entwistle for assistance in the original design. Information on how to obtain the Add Health data files is available on the Add Health website (<https://www.cpc.unc.edu/addhealth>). No direct support was received from grant P01-HD31921 for this analysis.
- [28] A. Clauset, S. Arbesman, and D. B. Larremore, Systematic inequality and hierarchy in faculty hiring networks. *Science Advances* **1**, e1400005 (2015).
- [29] C. Vilette, T. Bonnell, P. Henzi, and L. Barrett, Comparing dominance hierarchy methods using a data-splitting approach with real-world data. *Behavioral Ecology* **31**, 1379–1390 (2020).
- [30] M. J. Silk, M. A. Cant, S. Cafazzo, E. Natoli, and R. A. McDonald, Elevated aggression is associated with uncertainty in a network of dog dominance interactions. *Proceedings of the Royal Society B* **286**, 20190536 (2019).
- [31] M. Franz, E. McLean, J. Tung, J. Altmann, and S. C. Alberts, Self-organizing dominance hierarchies in a wild primate population. *Proceedings of the Royal Society B* **282**, 20151512 (2015).
- [32] D. J. Watt, Relationship of plumage variability, size and sex to social dominance in Harris’ sparrows. *Animal Behaviour* **34**, 16–27 (1986).
- [33] C. M. Williamson, B. Franks, and J. P. Curley, Mouse social network dynamics and community structure are associated with plasticity-related brain gene expression. *Frontiers in Behavioral Neuroscience* **10**, 152 (2016).
- [34] E. D. Strauss and K. E. Holekamp, Social alliances improve rank and fitness in convention-based societies. *Proceedings of the National Academy of Sciences* **116**, 8919–8924 (2019).
- [35] R. M. Neal, MCMC using Hamiltonian dynamics. In S. Brooks, A. Gelman, G. Jones, and X.-L. Meng (eds.), *Handbook of Markov Chain Monte Carlo*, pp. 113–162, Chapman and Hall, New York (2011).
- [36] M. Betancourt, A conceptual introduction to Hamiltonian Monte Carlo. Preprint arXiv:1701.02434 (2017).
- [37] C. De Bacco, D. B. Larremore, and C. Moore, A physical model for efficient ranking in networks. *Science Advances* **4**, eaar8260 (2018).
- [38] M. T. Hallinan and W. N. Kubitschek, The effect of individual and structural characteristics on intransitivity in social networks. *Social Psychology Quarterly* **51**, 81–92 (1988).
- [39] B. Ball and M. E. J. Newman, Friendship networks and social status. *Network Science* **1**, 16–30 (2013).
- [40] E. D. Strauss, A. R. DeCasien, G. Galindo, E. A. Hobson, D. Shizuka, and J. P. Curley, DomArchive: A century of published dominance data. *Philosophical Transactions of the Royal Society B* **337**, 20200436 (2022).
- [41] C. Neumann and J. Fischer, Extending Bayesian Elo-rating to quantify the steepness of dominance hierarchies. *Methods in Ecology and Evolution* **14**, 669–682 (2023).
- [42] D. Leiva and H. de Vries, *Testing steepness of dominance hierarchies* (2022), URL <https://CRAN.R-project.org/package=steepness>. R package, version 0.3-0.
- [43] C. Neumann, *EloSteepness: Bayesian dominance hierarchy steepness via Elo rating and David’s scores* (2023), URL <https://CRAN.R-project.org/package=EloSteepness>. R package, version 0.5.0.