Science Advances
AAAS

Supplementary Materials for

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

Maximilian Jerdee and M. E. J. Newman

Corresponding author: M. E. J. Newman, mejn@umich.edu

Sci. Adv. **10**, eadn2654 (2024)
DOI: 10.1126/sciadv.adn2654

This PDF file includes:

S1 Data sets

The example data sets used in this paper are summarized in Table 1 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 forms 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 substantial fraction of the nominations are reciprocal—two individuals each nominate the other as a friend (*40, 41*). 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.

For the faculty hiring data sets, the original source (*33*) included three data sets, for business schools, computer science departments, and history departments. We include only the first two of these in our analysis, purely to avoid cluttering the presentation, since the results for history departments very similar to the other two: we find $(\hat{\beta}, \hat{\alpha}) = (4.38, 0.01)$ for history, compared to $(4.36, 0.01)$ for business and $(4.25, 0.01)$ for computer science.

Animal hierarchies: Data on animal dominance hierarchies are 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.* (*42*). 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.

S2 Other measures of model performance

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 $\hat{\boldsymbol{s}}$ 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:

$$C = \frac{\sum_{ij} W_{ij} \mathbf{1}_{\hat{s}_i > \hat{s}_j}}{\sum_{ij} W_{ij}}, \tag{S1}$$

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. S1A 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 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.

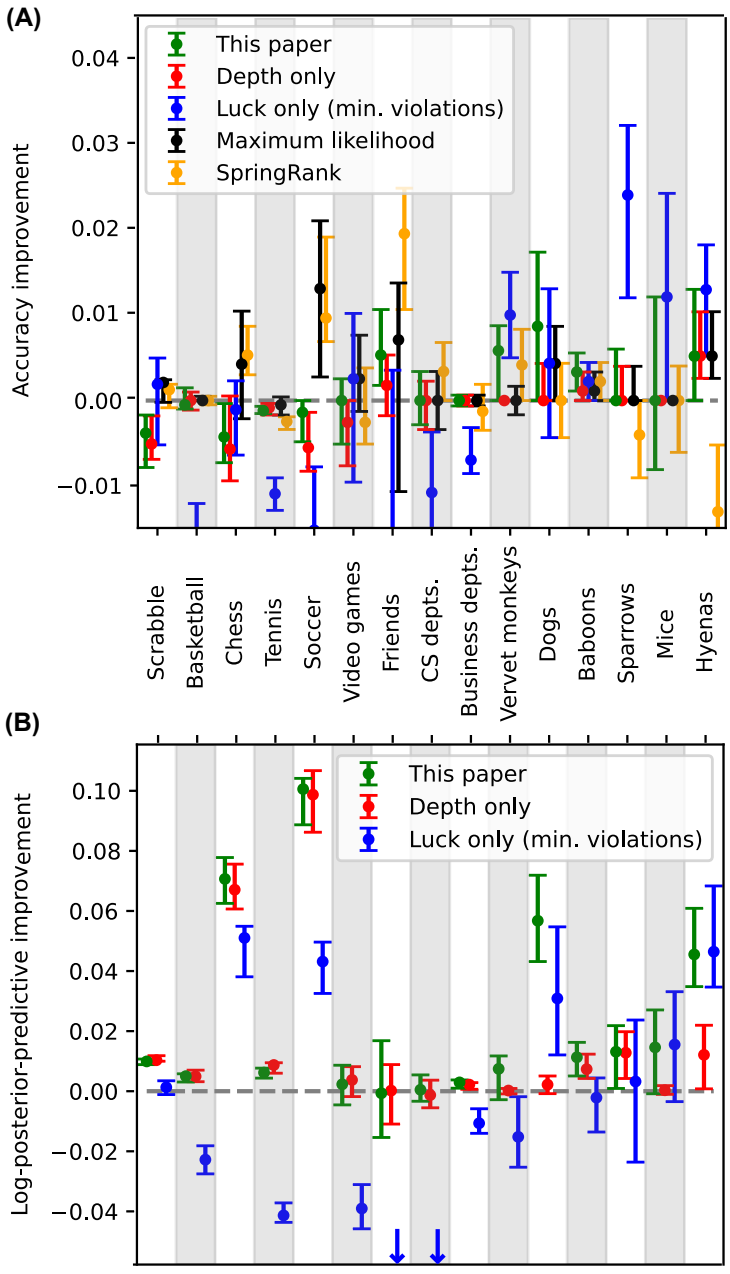

Figure S1: Results from the same set of cross-validation tests shown in Fig. 3, but quantified using (A) accuracy and (B) log posterior-predictive probability, instead of log-likelihood. All results are measured relative to the Bradley-Terry model with a logistic prior, which is represented as the dashed 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. Arrows indicate results that are off the scale.

Both the likelihood and accuracy measures are based on point estimates of model parameters $\hat{\boldsymbol{s}}, \hat{\alpha}$, and $\hat{\beta}$ but, as shown in Fig. 2, point estimates do not always do a good job of capturing the full posterior distribution $P(\boldsymbol{s}, \alpha, \beta | \boldsymbol{A}_{\text{train}})$, 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:

$$P(\boldsymbol{A}_{\text{test}} | \boldsymbol{A}_{\text{train}}) = \int P(\boldsymbol{A}_{\text{test}} | \boldsymbol{s}, \alpha, \beta) P(\boldsymbol{s}, \alpha, \beta | \boldsymbol{A}_{\text{train}}) \, d^n\boldsymbol{s} \, d\alpha \, d\beta. \tag{S2}$$

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

$$P(\boldsymbol{A}_{\text{test}} | \boldsymbol{A}_{\text{train}}) \simeq \frac{1}{N} \sum_{k=1}^{N} P(\boldsymbol{A}_{\text{test}} | \boldsymbol{s}_k, \alpha_k, \beta_k). \tag{S3}$$

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

$$R = \frac{\log P(\boldsymbol{A}_{\text{test}} | \boldsymbol{A}_{\text{train}})}{\sum_{ij} W_{ij}}, \tag{S4}$$

a fully Bayesian performance measure.

We plot this measure for a number of our models and data sets in Fig. S1B. 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 point estimates of the scores only, 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.

S3 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 $\hat{\boldsymbol{s}}$ 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 $P(\boldsymbol{s}, \alpha, \beta|\boldsymbol{A})$ over both the scores and parameters to define estimates

$$(\boldsymbol{s}^*, \alpha^*, \beta^*) \equiv \text{argmax}_{\boldsymbol{s},\alpha,\beta} P(\boldsymbol{s}, \alpha, \beta|\boldsymbol{A}). \tag{S5}$$

We find, however, that this can give biased results by artificially inflating the value of the depth parameter β. This happens because the likelihood $P(\boldsymbol{A}|\boldsymbol{s}, \alpha, \beta)$ is a function of the product $\beta\boldsymbol{s}$ (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 $\boldsymbol{s}$ in this way increases the prior $P(\boldsymbol{s})$ (which is peaked at $\boldsymbol{s} = 0$) and so increases the posterior $P(\boldsymbol{s}, \alpha, \beta|\boldsymbol{A})$. Unchecked, this effect would send the joint maximum to $\beta^* \to \infty$, $\boldsymbol{s} \to 0$. 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 S1.

S4 Bayesian model selection

As mentioned in the main paper, our luck-plus-depth model can be regarded as a generalization of a number of other popular ranking models. Setting $\alpha = 0$ fixing β, for example, recovers the standard Bradley-Terry model with a fixed-width prior, while letting $\beta \to \infty$ recovers the minimum violations ranking. This means that the value of $P(\alpha, \beta|A)$ on points or surfaces within the α, β space gives the corresponding distribution for these other models and hence allows us to compare model performance and perform true Bayesian model selection.

Within the Bayesian paradigm, competing models are often compared by computing the Bayes factor between them, the ratio of their Bayesian evidences. The Bayesian evidence is calculated by integrating the model likelihood over the space of all possible parameter (or score) values, weighted by their prior probabilities. Model selection using the Bayes factor is analogous to a classical likelihood ratio test, although by integrating over parameters Bayes factors naturally penalize over-parameterized models. Applied to our setting, we can use the Bayes factor to adjudicate whether our full luck-plus-depth model is preferred to the various models it generalizes, for each of our data sets.

For example, as discussed in Section S5, the Bradley-Terry model with unit logistic prior on the scores s_i corresponds approximately to the parameter choices $\alpha_L = 0$ and $\beta_L \simeq 2.56$. The Bayesian evidence of this model on a particular data set is then obtained by integrating over the possible values of the scores while holding these parameters fixed:

$$P(A|\alpha_L, \beta_L) = \int P(A|\boldsymbol{s}, \alpha_L, \beta_L) P(\boldsymbol{s})\, d\boldsymbol{s}. \tag{S6}$$

The Bayesian evidence of the full model, where the parameters α and β run free, is

$$\begin{aligned} P(A) &= \int P(A|\boldsymbol{s},\alpha,\beta)P(\boldsymbol{s})P(\alpha)P(\beta)\, d\boldsymbol{s}\, d\alpha\, d\beta \\ &= \int P(A|\alpha,\beta)P(\alpha)P(\beta)\, d\boldsymbol{s}\, d\alpha\, d\beta. \end{aligned} \tag{S7}$$

Then the Bayes factor between the two models is the ratio

$$K = \frac{P(A|\alpha_L,\beta_L)}{P(A)}, \tag{S8}$$

which can be interpreted in a manner akin to a p-value: if the value is very small then the evidence for the model with zero luck and a fixed logistic prior is much weaker than for our more general model.

In practice the Bayes factor can be calculated from our Monte Carlo samples, which allow us to approximate the posterior distribution $P(\alpha,\beta|A)$ up to a multiplicative constant. The Bayes factor can then be found by applying Bayes law thus:

$$\frac{P(A|\alpha_L,\beta_L)}{P(A)} = \frac{P(\alpha_L,\beta_L|A)}{P(\alpha_L)P(\beta_L)} \tag{S9}$$

This expression is precisely the function of α_L and β_L plotted in Figure 2A. (The inverse prior factor $1/P(\alpha_L)P(\beta_L)$ is implicitly included because of the nonlinear scale we use for the horizontal axis—the scale was deliberately chosen so that the Jacobian of the transformation gives exactly the required factor.) Therefore, the darker regions in Figure 2 indicate choices of parameters that correspond to models with high Bayes factor, while absence of color indicates that the Bayes factor is low and the corresponding model is excluded. Thus it is clear that the conventional Bradley-Terry model at $\alpha_L = 0$ and $\beta_L \simeq 2.56$ is excluded relative to our full model for all but one data set, the "friends" data.

Other models correspond to regions in parameter space, not just points. Minimum violations ranking, for example, is the $\beta \to \infty$ limit of our model for any value of α—the whole of the right edge of Fig. 2A—while the depth-only model corresponds to the lower ($\alpha = 0$) edge. The full model evidence for each of these models can then be found by integrating over the appropriate subspace of parameters. Even without performing the integral, however, it is clear simply from looking at the figure that the minimum violations ranking is only plausible for the hyenas, mice, and sparrows, while the depth-only model is plausible in all *but* the vervet monkeys, dogs, and mice.

(To be fair, for this simple visual analysis to be correct we would have to ensure that the distributions $P(\alpha,\beta|A)$ in Fig. 2A are all normalized to 1, meaning that probability clouds with a large extent, such as the mice data set, would be far dimmer than in fact they are—for

visual clarity the clouds have actually been normalized to maximize contrast. Nonetheless, the results of our rough eyeball test do hold out regarding the minimum violations ranking and the depth-only model.)

S5 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 the section on depth of competition, 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

$$\hat{\beta}_{\text{range}} = \hat{\beta}(\hat{s}_{\text{max}} - \hat{s}_{\text{min}}). \tag{S10}$$

In the data sets studied here we find that the factor $\hat{s}_{\text{max}} - \hat{s}_{\text{min}}$ 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 greatly between this measure and our own. Values of $\hat{\beta}_{\text{range}}$ are reported in Table S1 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 S1 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 substantial 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 the standard deviation, we report figures for each of our data sets in Table S1, 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 substantially easier to calculate than the latter, this could be a useful approach for calculations where accuracy and rigor are not at a premium.

We also note that MAP estimation in our depth-only model is equivalent to fitting the usual Bradley-Terry model with an L2 regularization, equivalent to a Gaussian prior. This correspondence suggests that one could infer a point estimate of the depth by tuning the strength of the L2 regularization by maximizing performance on some held-out validation data set. Like the joint MAP estimation of the depth and the scores discussed in Section S3, however, this method displays a bias towards large depth values. Specifically, since validation performance is only assessed at the MAP point estimate of the scores, if that point scales as $\boldsymbol{s} \sim \beta^{-1}$ for large β the validation log-likelihood is largely unaffected, being proportional to βs. The Bayesian approach, by contrast, penalizes high values of β by effectively assessing performance integrated over the whole posterior distribution of scores, since large β values are sensitive to slight changes in these scores beyond the point estimate.

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 (*22*). 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

$$\mathrm{DS}_i = w_i + \sum_j w_j P_{ij} - l_i - \sum_j l_j P_{ji} \tag{S11}$$

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

$$P_{ij} = \frac{A_{ij}}{A_{ij} + A_{ji}}, \tag{S12}$$

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

$$w_i = \sum_j P_{ij}, \qquad l_i = \sum_j P_{ji}. \tag{S13}$$

De Vries *et al.* (*22*) propose normalizing the David's scores according to

$$\mathrm{NormDS}_i = \frac{\mathrm{DS}_i + \binom{n}{2}}{n}, \tag{S14}$$

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 (*43*) 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 S1 we report values for a number of our data sets of S_{DS} (calculated using the R package `steepness` (*44*)) and S_{Elo} (calculated using the R package `EloSteepness` (*45*)). 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 S1 the parameter β_S that appears in the SpringRank model (*25*). 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 S7.

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.

		Measures of depth								Luck	
	Data set	$\hat{\beta}$	β^*	$\hat{\beta}_{\text{range}}$	$\hat{\beta}_0$	std($\hat{\boldsymbol{s}}_L$)	S_{DS}	S_{Elo}	$\hat{\beta}_S$	$\hat{\alpha}$	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 S1: Inferred parameter values for the data sets considered in the Results section of the main paper. From left to right: $\hat{\beta}$ is expected depth, β^* is the jointly optimized MAP depth as in Eq. (S5), $\hat{\beta}_{\text{range}}$ is depth between the best and worst player as in Eq. (S10), $\hat{\beta}_0$ is restricted depth as inferred in the depth-only ($\alpha = 0$) model, std($\hat{\boldsymbol{s}}_L$) is the standard deviation of the MAP scores within the logistic-prior model, S_{DS} is the steepness measure of de Vries *et al.* (*22*), S_{Elo} is the steepness measure of Neumann and Fischer (*43*), $\hat{\beta}_S$ is the maximum likelihood estimate of the parameter β_S in the SpringRank model (*25*), $\hat{\alpha}$ is the expected luck, and α^* is the jointly optimized MAP estimate of the luck.

S6 Depth as predictability

In the Results section of the main paper 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 our section on predicting wins and losses. 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 S2A 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-likelihood per match is close to -1). Some of the methods we compare against, notably the maximum likelihood Bradley-Terry and minimum violation ranking, fall well short even of random guesses, as indicated by the arrows at the bottom of the figure. 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

$$H[p_{ij}] = -p_{ij} \log p_{ij} - (1 - p_{ij}) \log(1 - p_{ij}). \tag{S15}$$

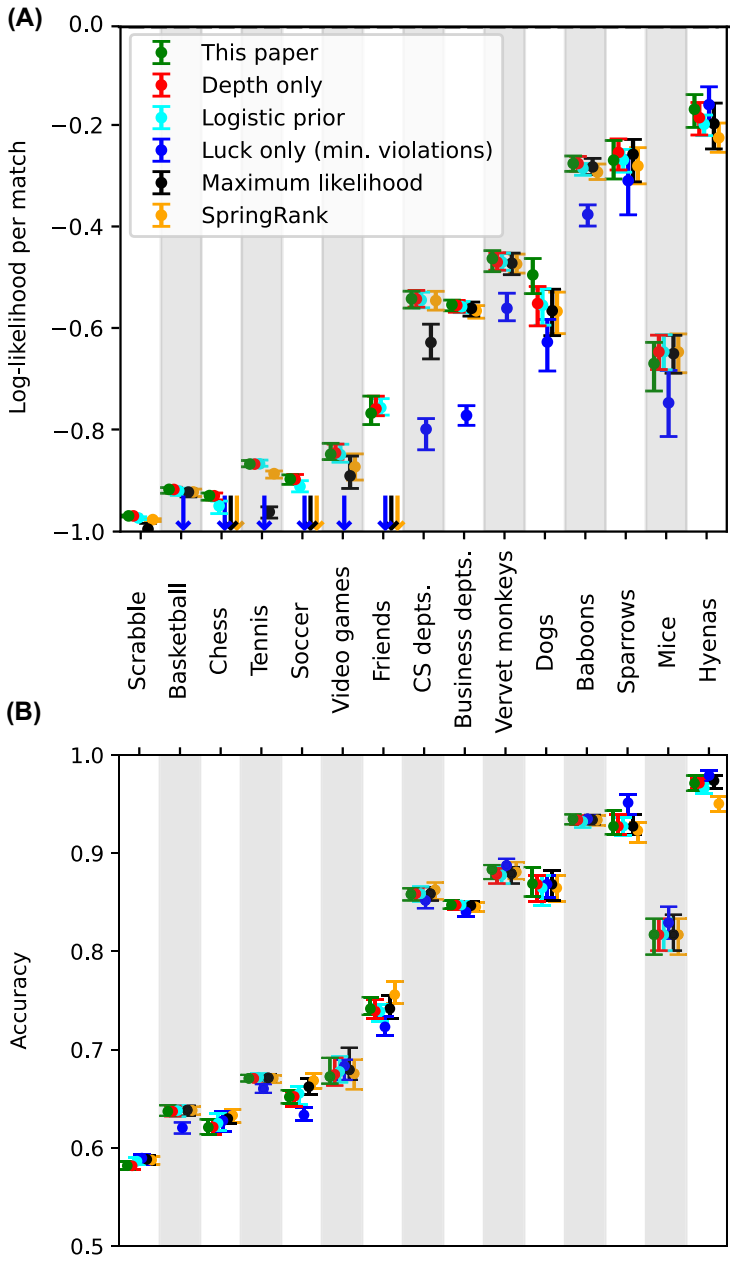

Figure S2: Absolute log-likelihood and accuracy values per match in the cross-validation tests of Fig. 3. This figure differs from Fig. 3 in showing absolute values rather than values relative to the Bradley-Terry model with logistic prior.

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

$$S_{\alpha\beta} = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} H\big[f_{\alpha\beta}(s)\big]\, e^{-s^2/2}\, ds. \tag{S16}$$

Unfortunately, this integral does not seem to have a closed-form solution, but it can be evaluated numerically. Figure S3 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 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. S2A, where we see that the uncertainty on our estimates of the log-likelihood is quite modest.

S7 SpringRank

Among the various approaches to ranking considered in this paper, SpringRank (*25*) 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 $\boldsymbol{A}$ is given by a product of Poisson distributions over all possible directed edges:

$$P(\boldsymbol{A}|\boldsymbol{s}, \beta_S, c) = \prod_{ij} \frac{r_{ij}^{A_{ij}}}{A_{ij}!} e^{-r_{ij}}, \tag{S17}$$

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

$$r_{ij} = c e^{-\frac{1}{2}\beta_S (s_i - s_j - 1)^2}, \tag{S18}$$

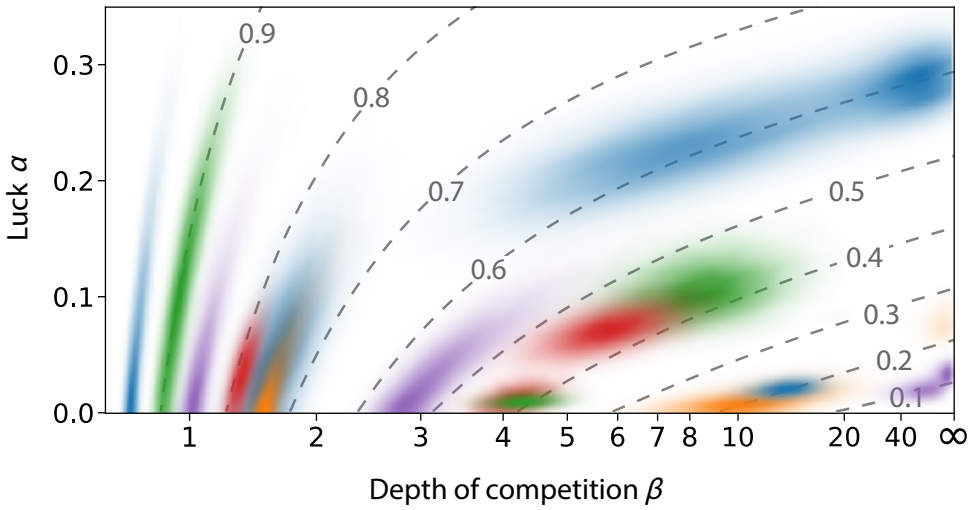

Figure S3: The data sets of Fig. 2 with dashed lines representing the contours of average entropy per match. Low entropy indicates confidence about the outcome of a match; high entropy indicates unpredictability.

for given scores $\boldsymbol{s}$, inverse temperature β_S, and a "sparsity" parameter c. Equation (S17) can be rewritten as

$$\begin{aligned} P(\boldsymbol{A}|\boldsymbol{s},\beta_S,c) &= \prod_{i<j} \frac{r_{ij}^{A_{ij}}}{A_{ij}!} e^{-r_{ij}} \, \frac{r_{ji}^{A_{ji}}}{A_{ji}!} e^{-r_{ji}} \\ &= \prod_{i<j} \frac{(r_{ij}+r_{ji})^{A_{ij}+A_{ji}} e^{-(r_{ij}+r_{ji})}}{(A_{ij}+A_{ji})!} \frac{(A_{ij}+A_{ji})!}{A_{ij}!A_{ji}!} \left(\frac{r_{ij}}{r_{ij}+r_{ji}}\right)^{A_{ij}} \left(\frac{r_{ji}}{r_{ij}+r_{ji}}\right)^{A_{ji}} \\ &= \prod_{i<j} \frac{m_{ij}^{\bar{A}_{ij}} e^{-m_{ij}}}{\bar{A}_{ij}!} \binom{\bar{A}_{ij}}{A_{ij}} \frac{1}{[1+e^{-2\beta_S(s_i-s_j)}]^{A_{ij}}[1+e^{-2\beta_S(s_j-s_i)}]^{A_{ji}}}, \end{aligned} \tag{S19}$$

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

Equation (S19) is equal to the likelihood of generating an undirected network $\bar{\boldsymbol{A}}$ of matches and then separately choosing the directions of the edges, i.e., the winners of the matches:

$$P(\boldsymbol{A}|\boldsymbol{s},\beta_S,c) = P(\bar{\boldsymbol{A}}|\boldsymbol{s},\beta_S,c)\, P(\boldsymbol{A}|\boldsymbol{s},\beta_S,\bar{\boldsymbol{A}}), \tag{S20}$$

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

$$P(\bar{\boldsymbol{A}}|\boldsymbol{s},\beta_S,c) = \prod_{i<j} \frac{m_{ij}^{\bar{A}_{ij}} e^{-m_{ij}}}{\bar{A}_{ij}!} \tag{S21}$$

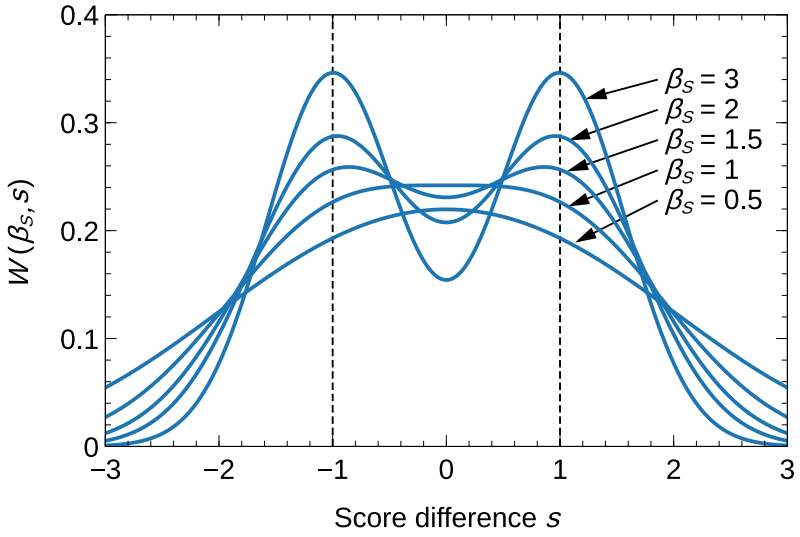

Figure S4: The function $W(\beta_S, s)$ of Eq. (S25) plotted against s, for various values of β_S as indicated.

and

$$P(\boldsymbol{A}|\boldsymbol{s}, \beta_S, \bar{\boldsymbol{A}}) = \prod_{i<j} \binom{\bar{A}_{ij}}{A_{ij}} \frac{1}{[1+e^{-2\beta_S(s_i-s_j)}]^{A_{ij}}[1+e^{-2\beta_S(s_j-s_i)}]^{A_{ji}}}. \tag{S22}$$

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

But Eq. (S22) is identical to the likelihood for the model studied in this paper, Eqs. (3) and (14),

$$P(\boldsymbol{A}|\boldsymbol{s}, \alpha, \beta, \bar{\boldsymbol{A}}) = \prod_{i<j} \binom{\bar{A}_{ij}}{A_{ij}} f_{\alpha\beta}(s_{ij})^{A_{ij}} f_{\alpha\beta}(s_{ji})^{A_{ji}}, \tag{S23}$$

if we choose $\alpha = 0$ 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 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 S1.

However, it is the other term, Eq. (S21), 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, 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

$$m_{ij} = M \frac{W(\beta_S, s_i - s_j)}{\sum_{i<j} W(\beta_S, s_i - s_j)}, \tag{S24}$$

where

$$W(\beta_S, s) = \sqrt{\frac{\beta_S}{8\pi}} \left[e^{-\frac{1}{2}\beta_S (s-1)^2} + e^{-\frac{1}{2}\beta_S (s+1)^2} \right]. \tag{S25}$$

(Note that $W(\beta_S, s)$ is symmetric in s so the sign of the score difference in Eq. (S24) 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 S4 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. (S22) and the distribution of matches via Eq. (S24). 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.* (*25*), 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. However, unlike the definition given in (*25*), our definition of the accuracy is independent of the choice of β_S.

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.

REFERENCES AND NOTES

1. H. A. David, *The Method of Paired Comparisons* (Griffin, ed. 2, 1988).

2. M. Cattelan, Models for paired comparison data: A review with emphasis on dependent data. *Stat. Sci.* **27**, 412–433 (2012).

3. A. N. Langville, C. D. Meyer, *Who's #1? The Science of Rating and Ranking* (Princeton Univ. Press, 2013).

4. R. A. Bradley, 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. *Math. Z.* **29**, 436–460 (1929).

6. L. R. Ford Jr., Solution of a ranking problem from binary comparisons. *Am. Math Mon.* **64**, 28–33 (1957).

7. D. R. Hunter, MM algorithms for generalized Bradley-Terry models. *Ann. Stat.* **32**, 384–406 (2004).

8. M. E. J. Newman, Efficient computation of rankings from pairwise comparisons. *J. Mach. Learn. Res.* **24**, 1–25 (2023).

9. J. T. Whelan, Prior distributions for the Bradley-Terry model of paired comparisons. arXiv:1712.05311 [math.ST] (2017).

10. R. R. Davidson, 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. *J. Comput. Graph. Stat.* **21**, 174–196 (2012).

12. P. V. Rao, L. L. Kupper, Ties in paired-comparison experiments: A generalization of the Bradley-Terry model. *J. Am. Stat. Assoc.* **62**, 194–204 (1967).

13. R. R. Davidson, On extending the Bradley-Terry model to accommodate ties in paired comparison experiments. *J. Am. Stat. Assoc.* **65**, 317–328 (1970).

14. R. D. Luce, *Individual Choice Behavior: A Theoretical Analysis* (John Wiley, 1959).

15. R. L. Plackett, The analysis of permutations. *J. R. Stat. Assoc. C* **24**, 193–202 (1975).

16. A. Agresti, *Categorical Data Analysis* (Wiley, 1990).

17. S. Chen, T. Joachims, Modeling intransitivity in matchup and comparison data, in *Proceedings of the Ninth ACM International Conference On Web Search and Data Mining* (2016), pp. 227–236.

18. R. Makhijani, J. Ugander, Parametric models for intransitivity in pairwise rankings, in *The World Wide Web Conference* (2019), pp. 3056–3062.

19. M. E. J. Newman, Ranking with multiple types of pairwise comparisons. *Proc. R. Soc. A* **478**, 20220517 (2022).

20. W. Robertie. *Inside Backgammon* **2**, 3–4 (1980).

21. M.-L. Cauwet, O. Teytaud, H.-M. Liang, S.-J. Yen, H.-H. Lin, I.-C. Wu, T. Cazenave, A. Saffidine, Depth, balancing, and limits of the Elo model, in *Proceedings of the 2015 IEEE Conference on Computational Intelligence and Games 2015* (IEEE, 2015).

22. H. de Vries, J. M. G. Stevens, and H. Vervaecke, Measuring and testing the steepness of dominance hierarchies. *Anim. Behav.* **71**, 585–592 (2006).

23. R. M. Neal, MCMC using Hamiltonian dynamics, in *Handbook of Markov Chain Monte Carlo*, S. Brooks, A. Gelman, G. Jones, X.-L. Meng, Eds. (Chapman and Hall, 2011), pp. 113–162.

24. M. Betancourt, A conceptual introduction to Hamiltonian Monte Carlo. arXiv:1701.02434 [stat.ME] (2017).

25. C. De Bacco, D. B. Larremore, C. Moore, A physical model for efficient ranking in networks. *Sci. Adv.* **4**, eaar8260 (2018).

26. Scrabble tournament records; https://cross-tables.com/.

27. N. Lauga, NBA games data; https://kaggle.com/datasets/nathanlauga/nba-games/data.

28. Online chess match data from lichess.com; https://kaggle.com/datasets/arevel/chess-games.

29. J. Sackmann, ATP tennis data; https://github.com/JeffSackmann/tennis_atp.

30. M. Jürisoo, International men's football results from 1872 to 2023; https://kaggle.com/datasets/martj42/international-football-results-from-1872-to-2017.

31. Super Smash Bros. Melee head to head records; https://etossed.github.io/rankings.html.

32. 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 Entwisle for assistance in the original design. Information on how to obtain the Add Health data files is available on the Add Health website (https://cpc.unc.edu/addhealth). No direct support was received from grant P01-HD31921 for this analysis.

33. A. Clauset, S. Arbesman, D. B. Larremore, Systematic inequality and hierarchy in faculty hiring networks. *Sci. Adv.* **1**, e1400005 (2015).

34. C. Vilette, T. Bonnell, P. Henzi, L. Barrett, Comparing dominance hierarchy methods using a data-splitting approach with real-world data. *Behav. Ecol.* **31**, 1379–1390 (2020).

35. M. J. Silk, M. A. Cant, S. Cafazzo, E. Natoli, R. A. McDonald, Elevated aggression is associated with uncertainty in a network of dog dominance interactions. *Proc. R. Soc. B* **286**, 20190536 (2019).

36. M. Franz, E. McLean, J. Tung, J. Altmann, S. C. Alberts, Self-organizing dominance hierarchies in a wild primate population. *Proc. R. Soc. B* **282**, 20151512 (2015).

37. D. J. Watt, Relationship of plumage variability, size and sex to social dominance in Harris' sparrows. *Anim. Behav.* **34**, 16–27 (1986).

38. C. M. Williamson, B. Franks, J. P. Curley, Mouse social network dynamics and community structure are associated with plasticity-related brain gene expression. *Front. Behav. Neurosci.* **10**, 152 (2016).

39. E. D. Strauss, K. E. Holekamp, Social alliances improve rank and fitness in convention-based societies. *Proc. Natl. Acad. Sci. U.S.A.* **116**, 8919–8924 (2019).

40. M. T. Hallinan, W. N. Kubitschek, The effect of individual and structural characteristics on intransitivity in social networks. *Soc. Psychol. Q.* **51**, 81–92 (1988).

41. B. Ball, M. E. J. Newman, Friendship networks and social status. *Netw. Sci.* **1**, 16–30 (2013).

42. E. D. Strauss, A. R. DeCasien, G. Galindo, E. A. Hobson, D. Shizuka, J. P. Curley, DomArchive: A century of published dominance data. *Philos. Trans. R. Soc. B* **377**, 20200436 (2022).

43. C. Neumann, J. Fischer, Extending Bayesian Elo-rating to quantify the steepness of dominance hierarchies. *Methods Ecol. Evol.* **14**, 669–682 (2023).

44. D. Leiva, H. de Vries, Testing steepness of dominance hierarchies (2022); https://CRAN.R-project.org/package=steepness. R package, version 0.3-0.

45. C. Neumann, EloSteepness: Bayesian dominance hierarchy steepness via Elo rating and David's scores (2023); https://CRAN.R-project.org/package=EloSteepness. R package, version 0.5.0.