J Stat Phys (2015) 160:275–301
DOI 10.1007/s10955-015-1253-6

Statistical Mechanics of the US Supreme Court

Edward D. Lee[1] · **Chase P. Broedersz**[1] · **William Bialek**[1,2]

Received: 14 October 2014 / Accepted: 31 March 2015 / Published online: 10 April 2015

Abstract We build simple models for the distribution of voting patterns in a group, using the Supreme Court of the United States as an example. The maximum entropy model consistent with the observed pairwise correlations among justices' votes, an Ising spin glass, agrees quantitatively with the data. While all correlations (perhaps surprisingly) are positive, the effective pairwise interactions in the spin glass model have both signs, recovering the intuition that ideologically opposite justices negatively influence each another. Despite the competing interactions, a strong tendency toward unanimity emerges from the model, organizing the voting patterns in a relatively simple "energy landscape." Besides unanimity, other energy minima in this landscape, or maxima in probability, correspond to prototypical voting states, such as the ideological split or a tightly correlated, conservative core. The model correctly predicts the correlation of justices with the majority and gives us a measure of their influence on the majority decision. These results suggest that simple models, grounded in statistical physics, can capture essential features of collective decision making quantitatively, even in a complex political context.

Keywords Statistical mechanics · Supreme Court · Maximum entropy

1 Introduction

Social and political systems, almost by definition, generate collective or emergent phenomena. It is natural to try describing these phenomena in the language of statistical mechanics [1–4], but it is not always clear whether this is a metaphor or a real theory within which we can make quantitative predictions. Here, in the spirit of recent work on biological systems

✉ Edward D. Lee
edl56@cornell.edu

1 Joseph Henry Laboratories of Physics, and Lewis–Sigler Institute for Integrative Genomics, Princeton University, Princeton, NJ 08544, USA

2 Initiative for the Theoretical Sciences, The Graduate Center, City University of New York, 365 Fifth Ave., New York, NY 10016, USA

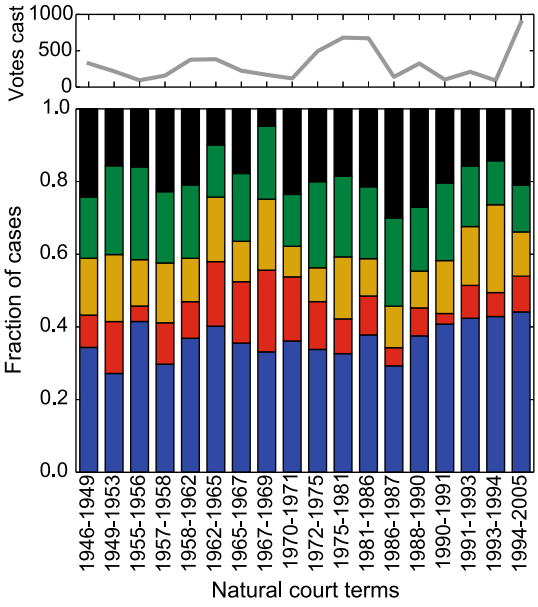

Fig. 1 Distribution of dissenting votes for natural courts that decided more than 100 cases. The *colored* portions represent the number of dissenting votes: *blue* (0), *red* (1), *yellow* (2), *green* (3), *black* (4). On average, 36 % of votes are unanimous over the 18 natural courts shown. Above, the number of votes cast by each natural Court. Data from Reference [10] (Color figure online)

[5–9], we bridge this gap in the context of voting on the Supreme Court of the United States (SCOTUS). While nine justices are not in the thermodynamic limit, we argue that models grounded in statistical physics provide a strikingly accurate, quantitative account of the observed voting patterns.

SCOTUS is the highest court in the US government. We consider natural courts, periods of time with constant membership, and focus on the second Rehnquist Court (1994–2005, 895 votes), the largest data set (Appendix 1). The Court writes majority and minority opinions, sometimes supplemented with other opinions; although these can be nuanced, each justice casts a yes ($\sigma_i = +1$) or no ($\sigma_i = -1$) vote, and the majority decides the outcome.

Popular descriptions of current US politics emphasize strong polarization along party lines, so that consensus and unanimity are seemingly rare. Comments on the nature of decision making on the Supreme Court also point to strong ideological divisions between right and left, with one or two justices providing "swing votes" [11,12]. In reality, unanimous decisions occur more frequently than 5–4 splits [13], as in Fig. 1. This pattern holds for over 50 years, and there is little indication that the unanimous cases are in a special class of "easy" decisions (Appendix 1).

The definition of yes and no in each case is determined by decisions in lower courts and thus is somewhat arbitrary. There are more relevant axes along which votes could be labeled, as with ideology, but it is not clear exactly how this underlying intuition corresponds to quantitative description. As a start, we imagine that the opposite definition was also possible, so that the voting patterns $\{\sigma_i\}$ and $\{-\sigma_i\}$ are equally likely, and we return to this problem below. With this symmetry, the average vote is neutral, $\langle \sigma_i \rangle = 0$ for all justices. Then, the first nontrivial voting statistic is the matrix of correlations, $C_{ij} = \langle \sigma_i \sigma_j \rangle$, shown in Fig. 2. One might have expected that justices known to have opposite ideological positions would tend to cast opposing votes, but all correlations are positive. We would like to understand not just pairwise correlations, but the entire distribution of voting patterns $P(\{\sigma_i\})$.

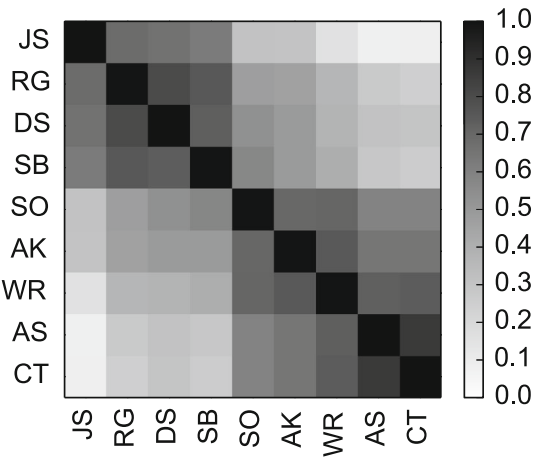

Fig. 2 Correlation matrix of votes in the Rehnquist court (1994–2005, 895 votes). Justices are identified by their initials: *JS* John Paul Stevens, *RG* Ruth Bader Ginsberg, *DS* David Souter, *SB* Stephen Breyer, *SO* Sandra Day O'Connor, *AK* Anthony Kennedy, *WR* William Rehnquist, *AS* Antonin Scalia, *CT* Clarence Thomas. They are ordered roughly from ideological left (JS) to right (CT) [10]. All correlations are positive, despite ideological differences. The *standard error* in estimating C_{ij} is given by $\delta C_{ij} = [(1 - C_{ij}^2)/K]^{1/2}$; with $K = 895$ we have $\delta C_{ij} < 0.034$ for all ij

2 Maximum Entropy Approach

An infinite number of distributions $P(\{\sigma_i\})$ are consistent with the observed correlations C_{ij}, but we search for the least structured distribution, or equivalently the one that generates the most random voting patterns (Appendix 2). Shannon showed that the unique measure of randomness or disorder is the entropy [14–16],

$$S[P(\{\sigma_i\})] \equiv -\sum_{\{\sigma_i\}} P(\{\sigma_i\}) \ln P(\{\sigma_i\}), \tag{1}$$

and so we maximize the entropy while matching the measured correlations C_{ij},

$$\sum_{\{\sigma_k\}} P(\{\sigma_k\})\sigma_i\sigma_j = C_{ij}. \tag{2}$$

The solution to this constrained optimization problem is a Boltzmann–like distribution,

$$P(\{\sigma_i\}) = \frac{1}{Z} e^{-E(\{\sigma_i\})}, \tag{3}$$

where the effective energy of each state is given by

$$E(\{\sigma_i\}) = -\frac{1}{2}\sum_{i\neq j} J_{ij}\sigma_i\sigma_j, \tag{4}$$

and the parameters J_{ij} are adjusted to satisfy the constraints in Eq. (2); as usual the partition function Z serves to normalize the distribution. This distribution embodies the *minimal* implications of the pairwise correlations, and involves no further assumptions. We recognize Eqs. (3, 4) as mathematically equivalent to the Ising model of a magnet, with "spins" σ_i interacting through "couplings" J_{ij}.

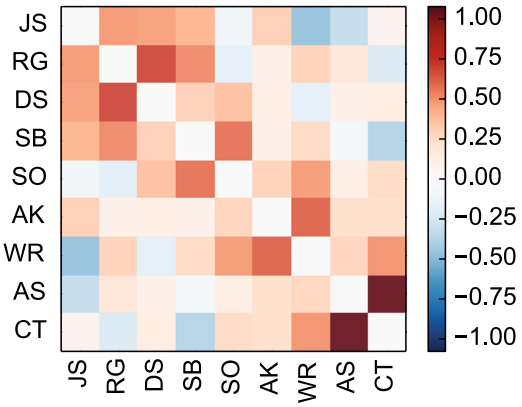

Fig. 3 Effective interactions in the Rehnquist court. We show the couplings J_{ij}, as in Eq. (4). Some J_{ij} are negative despite all positive C_{ij} in Fig. 2 (Color figure online)

With $N = 9$ justices, Eq. (2) provides 36 simultaneous nonlinear equations for the J_{ij} that can be solved numerically. The result for the Rehnquist Court (Fig. 2) is shown in Fig. 3. We see that the interactions J_{ij}, in contrast to the correlations C_{ij}, are both positive and negative.

The correlation matrix tells us that a positive vote by the most conservative justice (CT) increases the probability of a positive vote by the most liberal justice (JS) by $C_{ij}/2 \sim 4\,\%$, but this includes all the indirect paths through other members of the court to influence one another. In the context of the joint distribution for all votes, a positive vote by CT, with all other votes held fixed, contributes a factor $e^{J_{ij}} \sim 1.1$ to the probability of a positive vote by JS, surprisingly pulling JS further in the same direction, and at odds with ideological intuition. But another ideological opposite like AS, with a very similar voting record to CT, contributes a factor of $e^{J_{ij}} \sim 0.7$, *decreasing* the probability by 30 %. Thus, this model constructed only from (measured) positive correlations unmasks the hidden negative interactions, but shows that these do not conform fully to the binary intuition of negative influence across the ideological divide and positive influence within blocs.

Before proceeding, we address the errors in estimating J_{ij}. Individual J_{ij} are determined with standard deviations $\Sigma_J = 0.07 - 0.2$, but these errors are correlated. What really matters is predicting the "energy" of each state through Eq. (4), and we find that for low energy states the errors in the energy are $\Sigma_E = 0.2 - 0.3$ (Appendix 3). Thus, we determine the parameters well enough to predict probabilities of common states (those which occur several times in the data) with an accuracy of $\sim$20 %.

The pairwise model might paint an incomplete portrait since some interactions among groups of justices would not be captured by measurements on pairs [17]. The model, however, predicts the full distribution over voting patterns, and thus can be tested in various ways. First, we calculate the probability that the vote is split $(k, 9-k)$, with $k = 5, 6, \cdots, 9$ votes in the majority, shown in Fig. 4a. While there are small quantitative discrepancies, the model correctly predicts that unanimous votes are twice as likely as 5–4 splits, reproducing all probabilities with $\sim$10 % accuracy. Note that if the votes of individuals were independent, then unanimity would occur only $\sim$ 1 % of the time, while 5–4 splits would be most common. The observed tendency toward unanimity is described by the model as a truly emergent phenomenon, the minimal consequence of the observed correlations among pairs of justices.

Although error bars are larger, a second test is to estimate the probability of every voting pattern and compare these estimates with the model's predictions. We show this in Fig. 4b and see that theory and experiment agree within error bars for almost all patterns that occur more than once in the data.

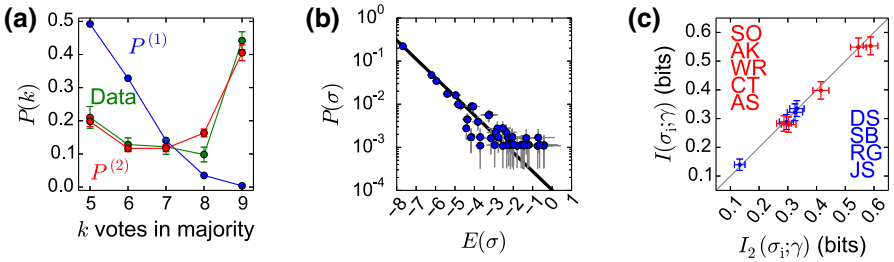

Fig. 4 Testing the maximum entropy model for the Rehnquist court. **a** Probability of k votes in the majority. We compare the data (*green*) with the predictions of the pairwise maximum entropy model $P^{(2)}$ (*red*), and with independent voters $P^{(1)}$ (*blue*). **b** Probability of each of the 102 observed voting patterns $\sigma \equiv \{\sigma_i\}$ versus the "energy" in Eq. (4); line is Eq. (3). Errors in probability arise, as usual, from counting; errors in the energy are propagated from errors in estimating the parameters J_{ij}. Only states that appear more than once are shown, setting a floor for $P(\sigma)$. **c** Mutual information $I(\sigma_i;\gamma)$ between individual votes σ_i and the decision γ of the majority, compared with $I_2(\sigma_i;\gamma)$ from the model. Conservatives are *red* and liberals *blue*, from highest $I(\sigma_i;\gamma)$ to lowest according to data (Table 1) (Color figure online)

Deviations between the model and the data are small, but could add up to significant effects. A third test, then, is to compare mutual information between the votes of individual justices and the majority vote γ. It is a nonlinear function of the sum of the votes, and large errors in predictions would signal important and unconsidered higher order interactions. We show the mutual information in Fig. 4c. The values of the mutual information $I(\sigma_i;\gamma)$ range over a factor of four, so that SO's vote provides 0.55 ± 0.03 bits of information about the decision of the court as a whole, while JS's vote provides only 0.14 ± 0.02 bits. This pattern, related to previous observations [11,18], is reproduced very accurately by the maximum entropy model.

The most direct test of maximum entropy models is to measure the entropy itself. If we build maximum entropy models that capture correlations of order n, then we generate a sequence of models with strictly decreasing entropy, $S_1 > S_2 > S_3 > \cdots > S_n$ [19]. In this sequence, $n = 1$ corresponds to a model of independent voting by each justice, while $n = 9$ corresponds to the exact model which reproduces correlations of all orders. The total amount of correlation in the system can be measured by the multi–information, $I_n = S_1 - S_n$.[1] The pairwise maximum entropy models capture a fraction $F = (S_1 - S_2)/I_9 = 0.95 \pm 0.03$ over all the natural courts shown in Fig. 1.

Taken together, the three results in Fig. 4 with the multi–information captured provide strong evidence that our model for the distribution of voting patterns captures the interesting structure in these data. We emphasize that this model is built only from measured pairwise correlations, and that once we have found the maximum entropy model there is no fitting of the data in Fig. 4; instead we have unambiguous, quantitative predictions, with no adjustable parameters.

3 Energy Landscape

The energy function $E(\{\sigma_i\})$ in Eq. (4) includes competing interactions, since the J_{ij} have both positive and negative signs. We expect that this competition generates multiple local

[1] Note that S_9 is the *actual* entropy of the voting patterns. For details on the technical problems of entropy estimation, see Appendix 5.

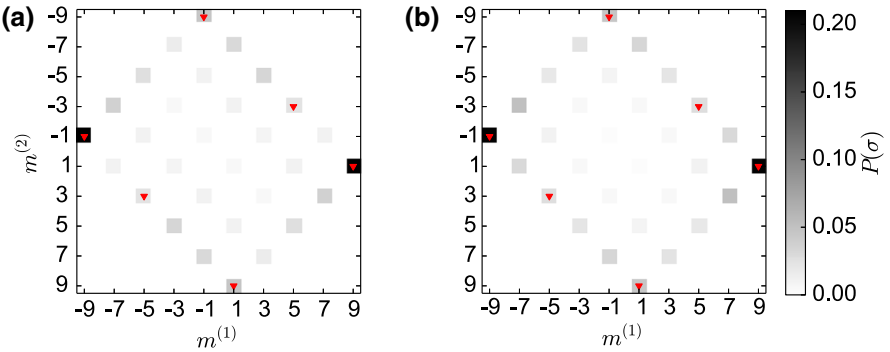

Fig. 5 Projection of the energy landscape in the data (**a**) and the predictions of the pairwise maximum entropy model (**b**). The *horizontal axis* shows the projection $m^{(1)}$ onto the unanimous +1 basin, and the *vertical axis* shows the projection $m^{(2)}$ onto 5–4 basin oriented so the majority voters are +1. The 7–2 basin lies in between as expected. Local energy minima are marked with *red triangles*. Note that points are separated by at least one *empty block* because a single vote flip corresponds to a change of 2 along either dimension. The space is highly structured, with density almost exclusively on the periphery and with a nearly empty center (Color figure online)

minima in the energy landscape [20], or local maxima in the probability distribution, at which flipping the vote of any single justice lowers the probability of the voting pattern. The model defines the energy for all possible states, not only those which are observed to occur in our finite sample of data. Furthermore, in the approach we have taken here, the entire landscape is determined by the *measured* correlations among the votes of pairs of justices. Thus, the landscape here is not a metaphor, but something that we can construct explicitly and quantitatively, with no free parameters.

For the Rehnquist court, with J_{ij} in Fig. 3, we find that more than 99 % (508/512) of the patterns fall into just 2×3 "valleys" in the energy landscape. The most populated pair of valleys are built around the two possible unanimous votes. A second pair of valleys are built around 5–4 splits that occur precisely along ideological lines (WR, SO, AS, AK, and CT versus JS, DS, RG, and SB). The third pair of valleys have at their base 7–2 splits, in which the most conservative and tightly correlated justices (AS and CT) dissent. Essentially all possible votes thus are organized around intuitive, prototypical patterns. Importantly, this bloc structure among multiple justices emerges from the pairwise maximum entropy model with no additional assumptions.

Leaning on the equivalence to statistical physics, we can think of the local minima in energy as the states into which the system is "trying" to order. If $\{\xi_i^{(n)}\}$ is the n^{th} local minimum, we can measure how close the system has come to this state by the overlap

$$m^{(n)} = \sum_{i=1}^{N} \xi_i^{(n)} \sigma_i. \tag{5}$$

If the system is very deep in the valley defined by $\{\xi_i^{(n)}\}$, then we will have $m^{(n)} \approx N$. With two dominant valleys in the energy landscape, the unanimous vote and the 5–4 ideological split, there are two natural "order parameters" $m^{(1)}$ and $m^{(2)}$, respectively [13]. In Fig. 5, we show the probability distribution projected onto these two dimensions, both for the real data and as predicted by the maximum entropy model.

In the projections along $(m^{(1)}, m^{(2)})$, we can see the clear local maxima of probability when $m = \pm N$, both in the data and in the predictions of the model. More surprising is that

Table 1 Measures of influence

	JS	RG	DS	SB	SO	AK	WR	AS	CT
$I(\sigma_i, \gamma)$	0.14	0.29	0.33	0.32	0.55	0.55	0.40	0.28	0.28
Γ_i	0.13	0.25	0.30	0.28	0.34	0.35	0.23	0.20	0.14

Mutual information $I(\sigma_i, \gamma)$ between the vote of Justice σ_i and of the Court γ, as in Fig. 4c. Influence increases as we move from ideological extremes into the medians. Susceptibility Γ_i of the majority to a signal from justice i, as defined in the text. Errors on $I(\sigma_i, \gamma)$ and Γ_i are similar across individuals, about 0.03 and 0.02, respectively

the distribution is almost confined to the edge of the allowed space. This feature of the data, which is predicted clearly by the model, means that the full distribution is in effect dominated by the competition between the tendencies toward unanimity and ideological division, and this is not just a qualitative statement but a quantitative one. Importantly, all of this structure is predicted by the maximum entropy model using only the observed pairwise correlations among votes as inputs.

4 Measuring Influence

A basic question about the dynamics of a court concerns individual influence on the majority decision. One measure is the mutual information $I(\sigma_i; \gamma)$ (Fig. 4c); because the votes are symmetric binary variables, this is equivalent to measuring the correlation $c_{\gamma,i} = \langle \gamma \sigma_i \rangle - \langle \gamma \rangle \langle \sigma_i \rangle$.[2] Alternatively, we can exploit the mapping of our model onto a system of spins and bias each justice's vote by a small "magnetic field" h_i. But our model is equivalent to an equilibrium statistical mechanics problem, so that $\chi_{\gamma i} = \partial \langle \gamma \rangle / \partial h_i = c_{\gamma,i}$. Thus, seemingly different ways of measuring influence are the same.

Continuing with the analogy to magnets, each justice experiences an effective field $h_i^{\rm eff} = \sum_{j \neq i} J_{ij} \sigma_j$ from others. If we imagine that justice i can "lean" in a positive direction by an amount ϵ, this creates fields $\Delta h_{j \neq i}^{\rm eff}(i) = J_{ji} \epsilon$. But through feedback, these fields will also bias the vote of justice i. To isolate the influence of this one justice, we add an additional field to fix the average vote of justice i, $\Delta h_i^{\rm eff}(i) = -(\epsilon / \chi_{ii}) \sum_j \chi_{ji} J_{ij}$, where $\chi_{ji} = \partial \langle \sigma_j \rangle / \partial h_i$. Now we can ask how the average majority vote would change if justice i signals an ϵ tendency toward a positive vote, but does not actually cast this vote. The resulting susceptibilities, $\Gamma_i = (1/\epsilon) \sum_j \chi_{\gamma j} \Delta h_j^{\rm eff}(i)$, are summarized in Table 1.

Using the susceptibility Γ_i as a measure of influence, influence tends to increase moving from the ideological extremes into the medians. The "median Justices" SO and AK are traditionally viewed as swing voters and indeed have maximal influence [11]. At both ideological extremes, we see that Justices JS and CT have minimal—and nearly identical—influence; this similarity is in contrast with their mutual information with the majority conservative court. Another notable observation is that Chief Justice WR's vote is the third most predictive of the majority decision, and interestingly the Chief Justice's position includes special privileges with procedural rules for the agenda and the assignment of written opinions [12]. Yet, he only has a median ranking according to Γ_i. We contrast this with the liberal justices (excepting JS) who are ranked higher by Γ_i than by the mutual information.

[2] $2I(\gamma; \sigma_i) = (1 + c_{\gamma,i}) \log_2(1 + c_{\gamma,i}) + (1 - c_{\gamma,i}) \log_2(1 - c_{\gamma,i})$.

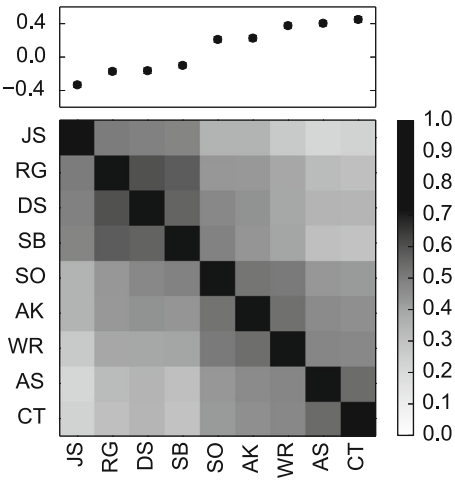

Fig. 6 Data for ideologically labeled votes on the second Rehnquist court. Correlation matrix $\langle \sigma_i \sigma_j \rangle - \langle \sigma_i \rangle \langle \sigma_j \rangle$ of ideological votes in the Rehnquist court, with average votes $\langle \sigma_i \rangle$ plotted above. As explained in the text, conservative/liberal votes are represented as binary variables $\sigma_i = \pm 1$ for each justice i [10]

5 Including Ideological Labels

It is perhaps surprising that we have been able to build accurate models without accounting for justices' ideological biases. On the contrary, structures that embody these biases emerge from the model. So much attention is paid to the right/left split in today's politics that we might imagine such influences are dominant [12]. Indeed, it could be that each justice responds independently to the merits of each case as seen through his or her political biases, and that what we see as correlations reflect nothing more than the fact that we are averaging over cases with different features. It may be useful to make the analogy to the case of sensory neurons. If each neuron in a network responds independently to its sensory inputs, and we average over these inputs, we will see correlations among the responses of different neurons. If we can hold the inputs fixed, however, it is possible that the correlations will vanish because there are no genuine interactions among the cells. It is not clear that we can do a completely analogous experiment with the Supreme Court, but mapping each yes/no vote onto a right/left decision seems like a reasonable start, as emphasized previously [4,11,13,21].

In fact, the raw data of from Spaeth et al. come labeled by the right/left sign of each vote [10]—although we note some difficulties (Appendix 1). But if we take the suggestion of ideological bias seriously, there are established definitions for what constitute the two ends of the ideological axis. Spaeth et al. have used these definitions to classify the votes as liberal (which we assign as $\sigma_i = -1$) or conservative ($\sigma_i = +1$) (Appendix 1, Appendix 6). Figure 6 shows the mean votes and pairwise correlations $\langle \sigma_i \sigma_j \rangle$ in the ideologically calibrated data. From the average ideological vote of his justice, we see strong opposing polarization between ideological extremes with $\langle \sigma_i \rangle_{JS} = -0.33$ and $\langle \sigma_i \rangle_{CT} = 0.45$, yet the correlation matrix remains nonzero and fully positive.

Suppose that we redo the analysis and now identify votes with right/left political positions [10]. Because we have lost the symmetry between yes and no, we want to build a maximum entropy model that matches both the pairwise correlations as before and the expectation values of the votes from individual justices, $\langle \sigma_i \rangle$. This model still has the form of Eq. (3), but now the energy function has explicit fields h_i acting on each spin,

$$E(\sigma) = -\sum_i h_i \sigma_i - \frac{1}{2} \sum_{ij} J_{ij} \sigma_i \sigma_j. \tag{6}$$

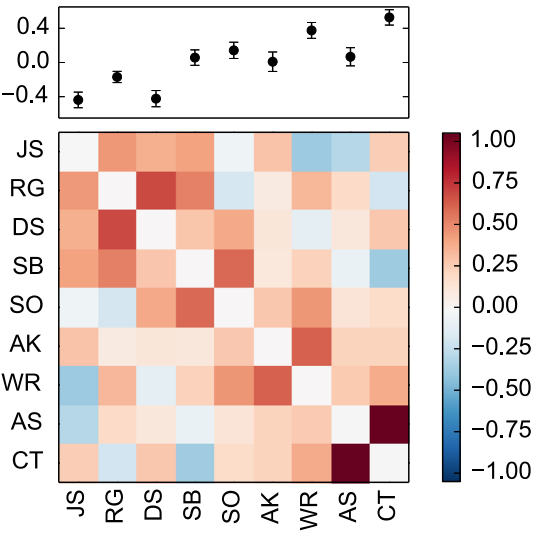

Fig. 7 Model for ideologically labeled votes on the second Rehnquist court. Effective interactions J_{ij}, with the biases h_i plotted above (Color figure online)

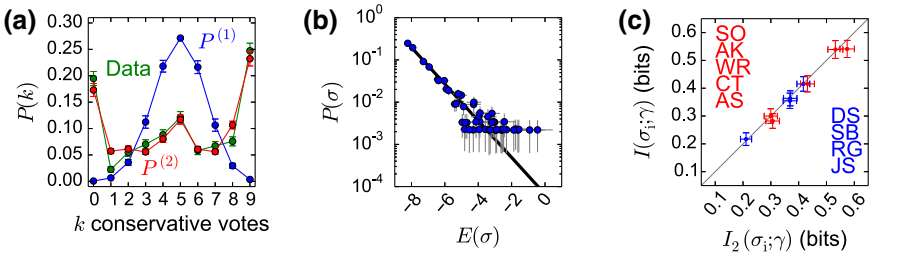

Fig. 8 Testing the ideological, maximum entropy model for the Rehnquist court. **a** Probability of k conservative votes. We compare the data (*green*) with the predictions of the pairwise maximum entropy $P^{(2)}$ (*red*), and with a model of independent votes $P^{(1)}$ (*blue*). **b** Probability of each of the 50 observed voting patterns versus the "energy" in Eq. (6); line is Eq. (3). Errors as in Fig. 4. Only states that appear more than once are shown. **c** Mutual information $I(\sigma_i;\gamma)$ between individual votes σ_i and the decision γ of the majority. Conservatives are *red* and liberals *blue*, from highest $I(\sigma_i;\gamma)$ to lowest according to data. *Error bars* represent standard deviations (Color figure online)

Results for the fields and couplings are shown in Fig. 7, and we see that it again provides a very good account of the data in Fig. 8. Computing entropies as in Appendix 5, we find that this model captures a fraction $F = 0.92 \pm 0.03$ of the multi–information across all natural courts shown in Fig. 1.

Several points about Fig. 8 seem worth noting, especially in relation to the corresponding Fig. 4 which shows the quality of model predictions in the symmetrized data. First, it is clear that the model correctly predicts the emergence of consensus on issues that favor both conservative and liberal positions (Fig. 8a); these consensus votes would be incredibly unlikely if each justice followed his or her biases independently. The probability of each observed voting pattern is predicted, with essentially the same accuracy as in the symmetrized case (Fig. 8b), and the maximum entropy model again captures very precisely the correlation between individual justices and the court majority (Fig. 8c).

The ideological model could have been very different: correlations between the votes of different justices might just reflect their ideological biases, so that if we keep track of these,

all the interactions J_{ij} will vanish. In fact, the couplings J_{ij} in the ideological model are almost the same as in the symmetrized model, with a correlation coefficient of 0.98 (Fig. 17). This means, for example, that the correlations between votes by AS and JS (discussed above) arise not merely because they adhere to opposite biases, but because they genuinely tend to vote against one another [22].

If we map the energy landscape in the case of ideologically labeled votes, we see slight but significant differences from the symmetrized case (not shown). We still have the largest valleys around the unanimous votes, but the conservative basin has $\sim 25\,\%$ more weight, as can be seen from Fig. 8a. There is a valley surrounding the 5–4 split, with conservatives in the majority, and a second smaller valley around a 5–4 split with conservatives voting liberally. We still see the valley around the 7–2 vote against AS and CT, but only one such valley exists since these two justices are so reliably conservative.

While the ideologically labeled data has somewhat more structure than the symmetrized data, it seems fair to summarize our analysis by saying that keeping track of the ideological biases of the justices in relation to the content of the question before the court adds relatively little to our predictive power, an observation we make precise in Appendix 7. How is this possible?

The essential feature of a maximum entropy model is the predicted energy landscape. For the symmetrized model, this landscape has multiple valleys, corresponding to unanimous votes and 5–4 ideological splits, as well as the smaller valleys in which AS and CT dissent from their seven colleagues (Fig. 5). This organization emerges collectively from the interactions among the judges, and we have seen that these interactions encode the ideological differences on the court even though we did not introduce these explicitly in constructing the model. Once the court is "polarized" along ideological lines, it takes only very small biases to align the polarized vote with the right/left content of the question before the court.

6 Discussion

We find that a pairwise maximum entropy model is sufficient to capture the voting distribution of SCOTUS, capturing over 90 % of the multi-information both when including and excluding ideological labels. We show that the model, built from only two-point correlations, can predict aggregate statistics like the distribution of votes in the majority, the correlation between the vote of a single justice and that of the court, and more detailed features like the entire probability distribution of votes. Although there are small deviations, the overall fit is remarkably good.

We explore the energy landscape of the model and find that prototypical voting blocs manifest as energy minima, or probability maxima, and that over 99 % of observed states belong to the three unanimous, ideological, and conservative voting structures. Thus, important voting bloc patterns arise as natural features in a landscape built only from pairwise interactions. We might consider neighbors of the minima as "noisy" versions of the dominant voting patterns. These fluctuations provide information about correlations in the system, and we can explore how fluctuations around the observed distribution might impact the statistical structure.

We propose one proxy for justice influence as the impact that fluctuations around his or her average vote have on the resulting majority outcome. This question of justice influence is highly debated in the political science literature [11,18], and we find that the final decision is least susceptible to the behavior of ideological extremes and most susceptible to those in the middle, a finding consistent with conventional wisdom. Although this measure is correlated

with the mutual information between the vote of a justice and the majority vote in the shown data, this correlation does not hold for all natural courts we investigated.

The pairwise maximum entropy model's success suggests that simple models, grounded in statistical physics, provide surprisingly accurate descriptions of collective behavior even in a complex, political context. One of the main intuitions behind the use of statistical physics ideas in the description of social dynamics is that the emergence of consensus or polarization is analogous to the emergence of order in physical systems at thermal equilibrium: having everyone in a group agree to vote the same way reminds us of all the spins in a magnet "agreeing" to point in the same direction. Importantly, once all the spins in a magnet agree to point in the same direction, even a very small external magnetic field is sufficient to get the entire magnet pointing north.

Concretely, the energy difference between a single electron spin pointing up or down in the earth's magnetic field is much, much smaller than the energy $k_B T$ that sets the scale of random thermal motion: individual spins *do not* point north reliably, although the collective magnetization of a compass magnet certainly does. Similarly, the biases which couple individual justices' ideological preferences to the merits of individual cases are weak, insufficient to induce unanimity or even to predict correctly the probability of a 5–4 split. What we see in the patterns of Supreme Court votes is dominated by the emergence of collective states, which then align to the particulars of individual cases. This is not a metaphor or analogy, but rather the description of a precise, quantitative model that predicts almost all the structure of these votes from the pattern of pairwise correlations.[3]

Acknowledgments We thank L Amaral, G Berman, M Castellana, B Daniels, J Evans, J Flack, D Krakauer, M Tikhonov, and others for many helpful discussions. Work in Princeton was supported in part by the National Science Foundation through Grants PHY–0957573 and CCF–0939370, by the WM Keck Foundation, and by the Lewis–Sigler Fellowship. Work at CUNY was supported in part by the Burroughs Wellcome Fund and by the Winston Foundation.

Appendices

The goal of this Appendix is twofold. First, there are a variety of technical issues which should be clarified. Second, we would like to make the relevant ideas as accessible as possible, beyond a physics audience. Thus, we give more than the usual background.

Appendix 1: The Data

Just 10 years ago, Sirovich's pioneering analysis of voting patterns on the Supreme Court required the manual entry and coding of data from individual cases [13]. Our task has been made much easier by the efforts of Spaeth and coworkers, who have compiled a large body of data and made it accessible through a web site at Washington University in St Louis [10]. To illustrate, we show in Fig. 9 the raw data on the Rehnquist court that forms the basis for most of our analyses. This example also shows the ideologically labeled liberal versus conservative votes which Spaeth et al. assigned using the criteria they detail on their webpage (see also Appendix 6).

[3] This can be seen as part of a larger effort to use the maximum entropy principle as a way of building statistical mechanics models of complex biological systems directly from data. See, as recent examples, Refs. [8,9], and references therein.

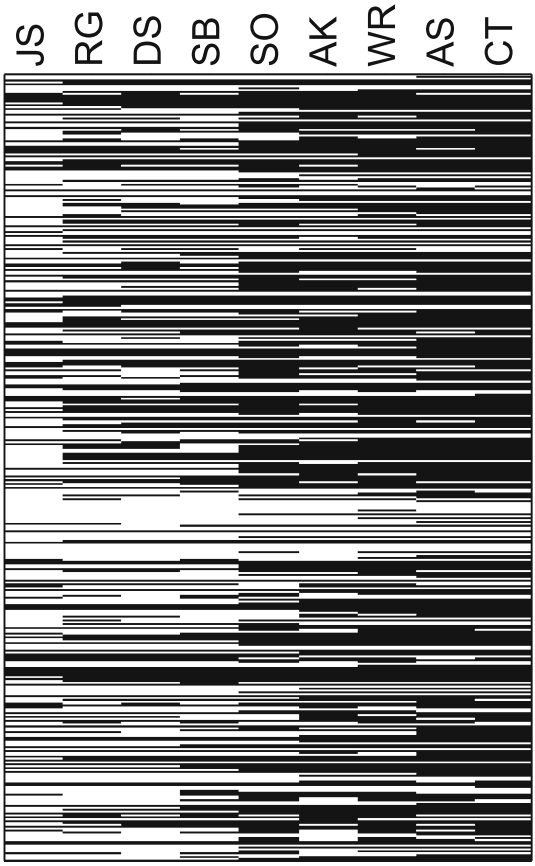

Fig. 9 Raw data on the Rehnquist court [10]. *Black* denotes a positive ($\sigma_i = +1$) vote, and *white* a negative ($\sigma_i = -1$) vote. As explained in the text, the sign is set along an ideological scale so that positive (negative) corresponds to a conservative (liberal) decision as defined by [10]

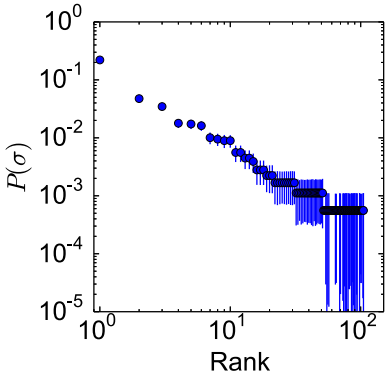

Fig. 10 Zipf plot of symmetrized voting patterns in second Rehnquist Court. *Error bars* are standard deviations over bootstrap samples

As discussed in the main text, the definitions of yes and no are, to some extent, arbitrary. Much of our discussion, then, is for a symmetrized data set in which we take our samples to consist both of the observed voting patterns $\{\sigma_i\}$ and the inverted patterns $\{-\sigma_i\}$. In Fig. 10, we show the probability of these patterns, sorted by their rank. This representation of the data often is called a "Zipf plot," after Zipf's discussion of the distribution of words in English [23]. As with words, we see that the probability has an approximately power–law dependence

on rank (a straight line on the log–log plot of Fig. 10), although in the present case this is true only over a very limited dynamic range. Although we have shown results primarily from the second Rehnquist Court, we have considered a total of 18 natural courts available in the data set; aspects of these different courts are discussed below.

One important point to consider about the data is whether the unanimous decision has uninteresting origins rather than serving to elucidate interactions between justices. In contrast, we note that some papers in the Supreme Court literature consider unanimous votes uninformative [24]. For example, unanimous decisions could be dominated by "fixing" unusual decisions from extreme appellate courts. To test for this, we compare the distribution across appellate courts of the unanimous and non–unanimous decisions. The probability that a case originated from court i is p_i^u for unanimous and p_i^n for non-unanimous votes. For the second Rehnquist court, the Kullback–Leibler divergence, $D_{KL}(\{p_i^u\}||\{p_i^n\}) = 0.11$, is small relative to entropies of the distributions $S[\{p_i^u\}] = 4.27$ and $S[\{p_i^n\}] = 4.54$.

We can check whether the decisions are "easy" in the sense that the justices already know to decide unanimously. Given historical records about justice opinions preceding the public vote, some have shown that many unanimous votes begin with dissents, contrary to the notion of "easy." In the Waite court (1874–1887) that voted unanimously on more than 80 % of cases, 40 % of those votes initially had at least one dissent, but only 9 % had dissents in the final votes [17]. Finally, we exclude votes that have no ideological interpretation in case they are different—only about 2 % of votes in the second Rehnquist Court [10]. These cases can deal with matters like interstate conflict that are not typically associated with ideology.

Appendix 2: Maximum Entropy Models

The concept of entropy has its roots in thermodynamics, roughly 150 years ago. The idea that a maximum entropy principle could be used to build models of systems well outside the domain of thermodynamics is a more recent development, but still is more than 50 years old [14]. The intuition, which has surprising consequences, is that we would like to make models that match certain experimental observations, but we would like to do this in a way that does not introduce any structure beyond that which is necessary to match the data. Here we follow the classical development of this idea, leading to Eqs. (3, 4). For a textbook account see Appendix A.7 of Ref. [25].

To be more concrete, we imagine that the system we are studying has several degrees of freedom, each described by a variable σ_i; here these variables are the votes cast by the $i = 1, 2, \cdots, N = 9$ justices. The state of the entire system then is defined by the set of variables $\{\sigma_i\}$, and in the simplest case what we mean by "making a model" is writing down the probability distribution out of which these states are being drawn, $P(\{\sigma_i\})$.

Once we adopt a probabilistic description, experimental observations are averages, or expectation values in this distribution. Thus, we might want to know the average vote of each justice, and this is given by

$$\langle \sigma_i \rangle_P \equiv \sum_{\{\sigma_j\}} \sigma_i P(\{\sigma_j\}), \tag{7}$$

where the sum is over all possible patterns of votes by all the justices, and the subscript reminds us that this average is being computed in the distribution P. Similarly, we might want to know the correlations between votes cast by different justices, and the pairwise correlations are given by

$$\langle \sigma_i \sigma_j \rangle_P \equiv \sum_{\{\sigma_k\}} \sigma_i \sigma_j P(\{\sigma_k\}). \tag{8}$$

These restrictions are equivalent to fixing the covariances, which be get by subtracting off the means: $C_{ij} = \langle \sigma_i \sigma_j \rangle - \langle \sigma_i \rangle \langle \sigma_j \rangle$. For the symmetrized data, we have that $\langle \sigma_i \rangle = 0$ so matching the pairwise correlations is the same as matching the covariances.

If we want our model to match experimental observations on these expectation values, we insist that $\langle \sigma_i \rangle$ computed from the distribution $P(\{\sigma_i\})$ be the same as the average computed from the experimental data,

$$\langle \sigma_i \rangle_P = \langle \sigma_i \rangle_{\text{expt}}, \tag{9}$$

and similarly for the correlations,

$$\langle \sigma_i \sigma_j \rangle_P \equiv \langle \sigma_i \sigma_j \rangle_{\text{expt}}. \tag{10}$$

Notice that we could also make other choices, matching different features of the data. The average votes and their pairwise correlations, however, seem like natural choices.

As emphasized in the text, Eqs. (9, 10) do not specify the distribution $P(\{\sigma_i\})$ uniquely. Indeed, there are infinitely many distributions that are consistent with these experimental observations. Out of these infinitely many possibilities, we would like to discipline ourselves, and not introduce any structure that is not actually needed in order to match the data, where "match the data" now has the concrete meaning of satisfying Eqs. (9, 10). Another way of saying this is that we would like a probability distribution such that, when we choose states out of this distribution, these states look as random as possible while still matching the data.

The idea that a distribution has minimal structure, or generates states with maximum randomness, might seem hopelessly qualitative. But in 1948 Shannon proved that there is a unique way to translate this intuition into mathematical terms if we adopt some simple requirements [15,16]. The result is that the only consistent measure of the randomness of states, or the lack of structure, is given by the entropy of the probability distribution,

$$S[P(\{\sigma_i\})] \equiv -\sum_{\{\sigma_i\}} P(\{\sigma_i\}) \ln P(\{\sigma_i\}). \tag{11}$$

There is an ambiguity of units; indeed, chemists and physicists typically choose different units for the entropy even in the thermodynamic context. The ambiguity of units is equivalent to an arbitrariness in choosing the base of the logarithm. Here we choose the natural log, but in other settings it is conventional to choose the logarithm base 2, in which case the units of entropy are bits. Importantly, the entropy which Shannon found as a measure of randomness or disorder in probability distributions is *exactly* the entropy that arises in statistical mechanics, and this is the same as the entropy for these systems in the thermodynamic sense.

In thermodynamics, coming to thermal equilibrium means finding a state with maximal entropy given whatever constraints the system experiences. In the present context, there are no heat flows and there is no notion of temperature or equilibrium. Instead, the maximum entropy probability distribution provides us with a model that is consistent with observed facts—taking Eqs. (9, 10) as constraints—but otherwise has as little structure as possible [14].

To carry out the maximum entropy construction, we need to find the probability distribution that maximizes the entropy subject to the constraints in Eqs. (9, 10). To do this we use the method of Lagrange multipliers [26]. We recall that if we want to maximize a function $f(\vec{x})$ of many variables, $\vec{x} \equiv \{x_1, x_2, \cdots, x_D\}$ subject to the constraint that $g(\vec{x}) = 0$, we can construct a new function $\tilde{f}(\vec{x}; \zeta) = f(\vec{x}) + \zeta g(\vec{x})$, where ζ is a "Lagrange multiplier." If we maximize $\tilde{f}(\vec{x}; \zeta)$ with respect to $\vec{x}$, we find a one parameter family

of solutions, depending on the value of ζ. If we maximize again with respect to ζ we will pick out the one solution in this family that satisfies the constraint $g(\overrightarrow{x}) = 0$. If we have many constraints, we add more Lagrange multipliers, one for each constraint, and sum the corresponding contributions to $\tilde{f}$.

If we want to maximize the entropy of the probability distribution $P(\{\sigma_i\})$ subject to the constraints in Eqs. (9, 10), the method of Lagrange multipliers tells us that we need to introduce a function

$$\begin{aligned}\tilde{S}[P(\{\sigma_j\}); \{h_i, J_{ij}\}] \equiv & -\sum_{\{\sigma_i\}} P(\{\sigma_i\}) \ln P(\{\sigma_i\}) + \sum_i h_i \left[\sum_{\{\sigma_j\}} \sigma_i P(\{\sigma_j\}) - \langle \sigma_i \rangle_{\text{expt}} \right] \\ & + \frac{1}{2} \sum_{ij} J_{ij} \left[\sum_{\{\sigma_k\}} \sigma_i \sigma_j P(\{\sigma_k\}) - \langle \sigma_i \sigma_j \rangle_{\text{expt}} \right] \\ & + \lambda \left[\sum_{\{\sigma_j\}} P(\{\sigma_i\}) - 1 \right]. \end{aligned} \tag{12}$$

Here, h_i is the Lagrange multiplier introduced to enforce the constraint on $\langle \sigma_i \rangle$ in Eq. (9), and J_{ij} is the Lagrange multiplier introduced to enforce the constraint on $\langle \sigma_i \sigma_j \rangle$ in Eq. (10); because the correlation matrix is symmetric we can take $J_{ij} = J_{ji}$, and the factor of $1/2$ reminds us that we are counting each term twice. Finally, λ is the Lagrange multiplier introduced to enforce the normalization of the probability distribution, which allows us formally to treat the variables $P(\{\sigma_i\})$ as independent real numbers, not worrying that they have to sum to one.

To find the maximum of $\tilde{S}$, we take derivatives with respect to the elements of the probability distribution, and set these to zero:

$$\frac{\partial \tilde{S}[P(\{\sigma_j\}); \{h_i, J_{ij}\}]}{\partial P(\{\sigma_i\})} = -\ln P(\{\sigma_i\}) - 1 + \sum_i h_i \sigma_i + \frac{1}{2} \sum_{ij} J_{ij} \sigma_i \sigma_j + \lambda \tag{13}$$

$$= 0 \tag{14}$$

$$\Rightarrow P(\{\sigma_i\}) = \frac{1}{Z} \exp \left[\sum_i h_i \sigma_i + \frac{1}{2} \sum_{ij} J_{ij} \sigma_i \sigma_j \right], \tag{15}$$

where $Z = e^{1-\lambda}$. In addition, we need to maximize $\tilde{S}$ with respect to the Lagrange multipliers. For λ, the condition $\partial \tilde{S}[P(\{\sigma_j\}); \{h_i, J_{ij}\}]/\partial \lambda = 0$ is equivalent to the normalization condition,

$$\sum_{\{\sigma_i\}} P(\{\sigma_i\}) = 1. \tag{16}$$

This sets the value of Z, which is called the partition function in statistical physics,

$$Z = \sum_{\{\sigma_i\}} \exp \left[\sum_i h_i \sigma_i + \frac{1}{2} \sum_{ij} J_{ij} \sigma_i \sigma_j \right]. \tag{17}$$

If we maximize with respect to h_i we find the condition in Eq. (9), and if we maximize with respect to J_{ij} we find the condition in Eq. (10). We have as many experimental measurements as we have Lagrange multipliers, and so there are enough equations to determine all the parameters. Solving these equations is another step, discussed in the next Section. Strictly

speaking, finding the point where derivatives vanish yields an extremum, not necessarily a maximum of the entropy. But the entropy is a convex function of the probability distribution [16], so that relevant second derivatives all are negative; hence any extremum will be a maximum.

To summarize, we have shown that the least structured probability distribution consistent with measured averages and pairwise correlations has the form

$$P(\{\sigma_i\}) = \frac{1}{Z} e^{-E(\{\sigma_i\})} \tag{18}$$

$$Z = \sum_{\{\sigma_i\}} e^{-E(\{\sigma_i\})} \tag{19}$$

$$E(\{\sigma_i\}) = -\sum_i h_i \sigma_i - \frac{1}{2} \sum_{ij} J_{ij} \sigma_i \sigma_j. \tag{20}$$

The parameters $\{h_i, J_{ij}\}$ are not arbitrary, but must be solved such that the predicted averages and pairwise correlations match the measured values, as in Eqs. (9, 10). Equations (18, 19) are exactly the Boltzmann distribution, which describes the distribution of states taken on by a system in thermal equilibrium, where the energy of each state is given by $E(\{\sigma_i\})$. In the physical setting, there is a real temperature T, which determines an energy scale $k_B T$, where k_B is Boltzmann's constant. Then, to be precise, we should write

$$P(\{\sigma_i\}) = \frac{1}{Z} e^{-E(\{\sigma_i\})/k_B T}, \tag{21}$$

but we are free to choose our units of energy so that $k_B T = 1$. Then it is clear that our problem of building minimally structured models leads us exactly to a statistical mechanics model of the system we are studying.

The physical interpretation of our model is that the (yes/no) votes of judges are Ising $(+1/-1)$ spins that each experience a "magnetic field" h_i and interact in pairs through the couplings J_{ij}. We have chosen a sign convention such that $h_i > 0$ favors a yes vote ($\sigma_i = +1$), and $J_{ij} > 0$ favors justices i and j voting in the same way.

Because we are asking to match both the averages $\langle \sigma_i \rangle$ and the pairwise correlations $\langle \sigma_i \sigma_j \rangle$, there are two sets of terms in the "energy" $E(\{\sigma_i\})$. In our initial formulation of the voting problem on the US Supreme Court, as described in the main text, yes and no votes are symmetric, and we automatically have $\langle \sigma_i \rangle = 0$ for every justice i. Then we need to match only the correlations between the votes of pairs of justices, and hence the energy function simplifies to

$$E(\{\sigma_i\}) = -\frac{1}{2} \sum_{ij} J_{ij} \sigma_i \sigma_j. \tag{22}$$

We see that Eqs. (18, 22) are the same as Eqs. (3, 4) of the main text. In a later discussion, we will break the symmetry between yes and no votes, and the fields h_i will then be important (Appendix 6).

It is important to emphasize that the maximum entropy method is *not* a model. It is a framework for building models that capture particular aspects of the data while making no additional assumptions. Thus, there are no free parameters to be "fit," and we are able to make unambiguous, quantitative predictions, as in Fig. 4.

Appendix 3: Solving the Inverse Problem

The maximum entropy construction arrives at Eqs. (18, 19, 22) analytically. To complete the construction, we actually have to find the numerical values of the coupling parameters J_{ij} that allow the model to match the observed correlations. That is, we have to solve Eq. (10), which can be written more explicitly as

$$\sum_{\{\sigma_k\}} \sigma_a \sigma_b \frac{1}{Z} \exp\left[\frac{1}{2}\sum_{ij} \sigma_i J_{ij} \sigma_j\right] = \langle \sigma_a \sigma_b \rangle_{expt}. \tag{23}$$

We note that both the correlations and the couplings define symmetric matrices. Thus, with $i = 1, 2, \cdots, N = 9$ justices, these are $N(N-1)/2 = 36$ simultaneous equations for the 36 independent parameters J_{ij}. These equations are relatively straightforward to solve numerically, for example using MATLAB's `fsolve` routine. Results for the second Rehnquist court are shown in Fig. 3.

The maximum entropy problem is formulated on the assumption that we know the correlation functions from experiment. In fact, these measurements come with error bars, since our sample is finite. We would like to convince ourselves that these errors have only a small impact on our ability to construct the model and make predictions. As a first test, we ask what happens when we choose random fractions of the full data set. Figure 11 shows that, for selected elements of the matrix J_{ij}, our best estimate has a very weak systematic dependence on the size of the data set, and that these individual parameters can be determined with reasonable precision. Figure 12 surveys these errors in the entire matrix, showing that

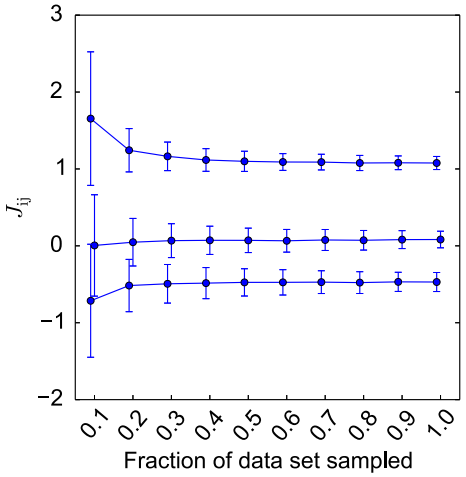

Fig. 11 Convergence of parameters with larger sample sizes. As examples, we show the positive and negative couplings with largest absolute values, as well as the coupling that has the smallest absolute value. We take bootstrap samples with replacement, of size $n = \text{Fraction} \times K$ with $K = 895$ the total number of votes recorded for the second Rehnquist Court. We construct independent maximum entropy models for each sample, and plot the mean couplings versus fraction of K; *error bars* are standard deviations of the couplings over multiple bootstrap samples at each Fraction. Points have been slightly displaced along the x-axis to minimize overlap between *error bars*

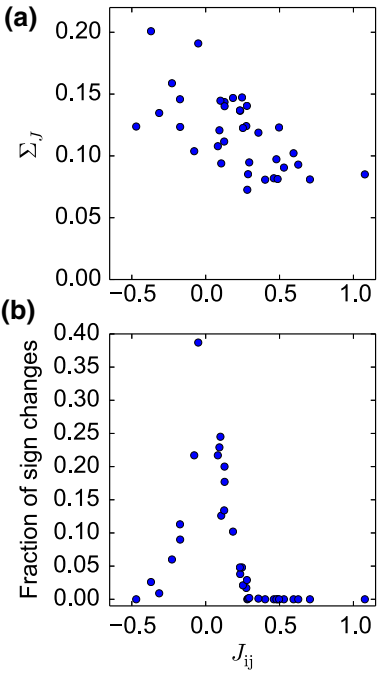

Fig. 12 **a** Standard deviation Σ_J versus J_{ij}. **b** Probability of a sign flip in J_{ij} over bootstrap samples. The strongest 2/3 of couplings have fixed sign with 95 % confidence

all elements are determined within ± 0.2, and many within ± 0.1; these random errors are estimated, as in Fig. 11, across bootstrap resamplings of the data. Importantly, for most of the J_{ij}, these resamplings have a low probability of changing our estimate of the sign of the interaction, and uncertainty about the sign is confined to the J_{ij} that have the smallest magnitude.

While our model is parameterized by the J_{ij}, the fundamental prediction of the model is the probability of each voting pattern; the logarithm of this probability is the "energy" of the state, from Eq. (4). If the errors in all the J_{ij} were independent, then the errors in the energy would typically be six times larger than the errors in the individual J_{ij}, and this would be quite bad. In fact the errors are unlikely to be independent, and they are not. To get some intuition, we know that if the J_{ij} themselves are drawn at random, then the correlations C_{ij} have a complicated structure [20]. Conversely, we expect that independent random additions to the C_{ij} would produce a structured change in the J_{ij}.

When we draw random samples of the data to generate C_{ij} and construct the corresponding J_{ij}, we get the whole matrix of J_{ij} and hence a whole set of predictions for the energies of individual states. We can look at the standard deviations of these energies, as a function of the means, as shown in Fig. 13. Low energy (more likely) states have errors $\Sigma_E \sim 0.2$, which means that we can predict the probability of these states with $\sim$20 % accuracy. This is possible only because the errors in the J_{ij} are correlated. Once we reach $\Sigma_E \sim \ln 2$, we can predict probabilities only within a factor of two. But this level of error is reached only for states with energy E roughly 8 units above the lowest energy state, and hence relative probability $\sim e^{-8} < 10^{-3}$. With only 895 samples, we thus should be able to predict, with reasonable reliability, the probabilities of all voting combinations that actually occur in the data, and this is borne out in Fig. 4.

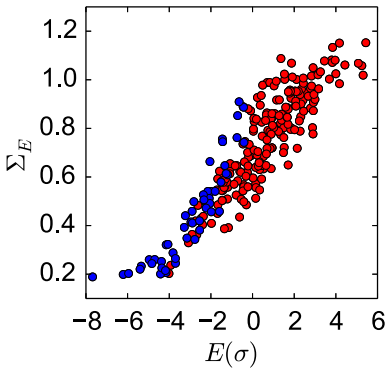

Fig. 13 Standard deviation Σ_E versus mean of energies $E(\sigma)$, for each state observed in the data. States that appear only once are shown in *red*, and states that appear more than once are shown in *blue* (Color figure online)

Appendix 4: The Energy Landscape

Maximum entropy models express the probability of a system being in a certain state (here, the pattern of votes by the nine justices) in terms of an "energy" for that state, and this energy in turn is built out of terms that express "interactions" among the elements of the system (here, the votes of the individual justices). It is useful to think about the energy as a landscape on the space of states, with deep valleys corresponding to states of high probability and mountains corresponding to states of low probability; mountain passes provided the most likely paths that connect one highly likely valley to another. The model defines the energy for all possible states, not only those which are observed to occur in our finite sample of data. Further, in the approach we have taken here, the entire landscape is determined by the *measured* correlations among the votes of pairs of justices. Thus, the landscape here is not a metaphor, but something that we can construct explicitly and quantitatively, with no free parameters.

In the case of the Supreme Court, the space of states is discrete: each justice votes yes or no ($\sigma_i = +1$ or $\sigma_i = -1$), and so while there are nine dimensions the allowed states live only on the corners of a (hyper)cube in these nine dimensions. Nonetheless we can identify two states as being neighbors if the jump from one state to the other involves changing the vote of only one justice. The bottom of a valley is a place where all moves take us uphill, and correspondingly we can ask for local minima of the energy function such that flipping the vote of any one justice always increases the energy. These local minima of the energy are predicted to be local maxima of the probability, and hence provide us with anchor points for thinking about the voting patterns. A local minimum of energy defines a prototypical vote, and the neighboring patterns—which are predicted to be less likely—can be thought of as "noisy versions" of this prototype.

As discussed in the main text, the model that we find for the patterns of votes on the US Supreme Court belongs to a class of models known in the physics literature as spin glasses [20], and a signature of these models is frustration of which a common consequence is the existence of multiple local minima (e.g. in the Sherrington-Kirkpatrick model). Thus, we expect that even our small ($N = 9$) system will have several local minima or prototypical voting patterns, and this is what we find.

Concretely, the energy landscape for the second Rehnquist court has three major valleys, plus the symmetric mirror of these valleys obtained by exchanging the definitions of yes and no. Together, the states in these valleys account for over 99 % (508/512) of the possible voting states, corresponding to almost the full mass of the probability distribution. As noted in the

main text, the prototypical states at the bottoms of the valleys are the unanimous vote, the 5–4 ideological split, and the 7–2 vote against Scalia and Thomas. We emphasize that these breaks along ideological lines emerge from the model even though we make no reference to ideology in our construction; these structures are encoded in the pairwise correlations, and the maximum entropy method allows us to make these structures explicit.

As we discuss in the main text, we find two dominant valleys in the energy landscape, and we show the probability distribution of overlap along these two natural "order parameters" $m^{(1)}$ and $m^{(2)}$ in Fig. 5. The probability distribution is confined nearly completely to the borders of this projection, showing that it is dominated forces competing for unanimity and ideological division.

We also note here that the two order parameters, the unanimous and ideological votes, are special in the sense that they are maximally different. In the space of symmetrized votes, true orthogonality is impossible because we have an odd number of voters, and the maximal distance is when $m^{(n)} = \pm - 1$ as is the case with unanimous and ideological votes. This observation might remind us of Principle Component Analysis, where we find a set of complete, orthogonal basis vectors to span the space. If we restrict ourselves to just a few components, we choose the ones with the largest eigenvalues that allow us to explain as much variation as possible. With the Supreme Court, we find that the primary energy minima correspond to maximally different states and allow us to represent the other states as linear combinations of those as is manifest in the concentration at the periphery of Fig. 5.

The idea of projecting the pattern votes onto two dimensions is not new. About 10 years ago, Sirovich [13] noted that the covariance of justices' votes is dominated by two principal components, which are very close to the dimensions defined by $m^{(1)}$ and $m^{(2)}$ here. Although we start with the same covariance matrix, the maximum entropy approach does more than identify dominant dimensions, since it makes quantitative predictions about the entire probability distribution. This is possible because the models we build respect the discrete nature of the votes—we are constructing a joint probability distribution for binary variables—while the covariance matrix itself could have arisen from a set of continuous variables, and the geometric interpretation of principal components analysis does not make reference to the "corners of the hypercube" structure in the space of voting patterns. By respecting the discreteness of votes, even the least structured model that is consistent with the covariance matrix exhibits a very rich structure, and one that is in detailed agreement with the data.

Appendix 5: Estimating Entropies and Information

At various points in our analyses, we estimate information theoretic quantities such as the entropy of a distribution or the mutual information between different variables. It is well known that such estimates can be systematically biased in small data sets [27]. This problem received considerable attention in the analysis of experiments on neural coding [28–31], and here we explain how we use what was learned in that context to be sure that our estimates are reliable. For a more pedagogical discussion, see Appendix A.8 of Ref. [25].

For the second Rehnquist Court, it is plausible that all relevant information theoretic quantities will be well determined: there are $\Omega = 512$ states, or really 256 independent probabilities in the symmetrized data, and we have $N = 2 \times 895$ samples (again, doubled because of symmetry). For other natural courts (Fig. 1), however, the number of samples is highly variable, down to as few as 91 votes. We would like to use all of the available examples, and thus we need to understand how the limited data set sizes can bias our estimates.

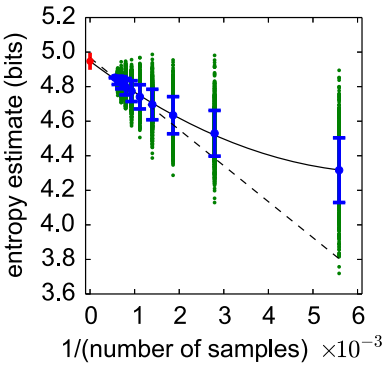

Fig. 14 Estimate of entropy voting patterns in the second Rehnquist Court. Given an initial data set with N samples, we draw multiple bootstrap samples of size n and form the "naive" estimate of the entropy from Eq. (24); results are shown as *green* points; means and standard deviations at each n plotted in *blue*. Extrapolations based on Eq. (25): linear fit to samples $n > N/2$ is the *dashed line*, and the quadratic fit to all plotted points is shown as a *solid line*. *Red* point is our best estimate, with errors (Color figure online)

The hardest quantity to estimate is the entropy of the distribution of voting patterns, since this depends on the probability of every single state. The naïve approach is to identify the observed frequency of occurrence of each state with its probability, and then plug these estimates into the definition of the entropy, here measured in bits,

$$S_{\text{naive}}(n) = -\sum_{\{\sigma_i\}} \hat{P}_n(\{\sigma_i\}) \log_2 \hat{P}_n(\{\sigma_i\}), \tag{24}$$

where $\hat{P}_n$ is the frequentist estimate of probability based on n samples. The differences between frequencies and probabilities are random—they average to zero, and as the sample size becomes larger their variance decreases uniformly. But the entropy is a nonlinear function of the probabilities, and so these random errors become systematic [27,28]. If the number of samples n is large enough, these systematic errors in the naive estimate take a simple form,

$$S_{\text{naive}}(n) = S_\infty + \frac{A}{n} + \frac{B}{n^2} + \cdots, \tag{25}$$

where S_∞ is the true entropy that we would find with an infinite number of samples, and A and B are constants. If we can convince ourselves that we are in the regime where this formula describes our systematic errors, then we are safe in taking the extrapolated S_∞ as an estimate of the entropy, as shown in Fig. 14. Notice that the "finite size correction," $S_\infty - S_{\text{naive}}(N)$ in Fig. 14, is less than 10 % of the total entropy, and that the difference between extrapolations where we include or ignore B in Eq (25) is even smaller; this is true, consistently, for the data on all the natural courts that we consider.

The extrapolation procedure in Fig. 14 should generate an unbiased estimate of the actual entropy. The maximum entropy models that we have constructed should, by definition, generate an upper bound on the entropy. This upper bound is based on measurements of the pairwise correlations, and since there are only 36 independent correlation matrix elements, even small data sets give a fairly reliable basis from which to construct these models. Happily, there is also a lower bound on the entropy that we can construct, and this too is rather robust to small sample sizes.

In a uniform distribution over Ω states, the probability that two states chosen at random are the same is $P_c = 1/\Omega$ and the entropy is $S = \log_2 \Omega$. Thus, in this case, we can estimate

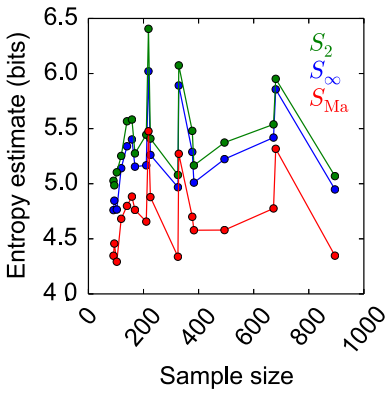

Fig. 15 Entropy estimates for all the natural courts, as a function of sample size. *Statistical errors*, as shown in Fig. 14, are small (Color figure online)

the entropy if we can estimate the probability of a coincidence, where two states are the same. Notice that if we have n samples, we can test $n(n-1)/2$ independent pairs, and so we start to get a reliable estimate of P_c as soon as $n \gg \sqrt{\Omega}$, which is much less than the naive expectation that we need to see all the states ($n \sim \Omega$) in order to say something about the distribution from which they are drawn. As an aside, this is the basis for the "birthday problem," where the number of people needed to make it likely that two of them share the same birthday is much less than the number of possible birthdays; a discussion appears in Feller's classic text [32].

To go beyond the uniform distribution, we note that the probability of a coincidence among two randomly chosen states is

$$P_c = \sum_{\{\sigma_i\}} [P(\{\sigma_i\})]^2 = \langle P(\{\sigma_i\})\rangle, \tag{26}$$

where $\langle \cdots \rangle$ stands for an expectation value over the distribution $P(\{\sigma_i\})$. But for any positive random variable x, we have

$$\log_2\langle x\rangle \geq \langle \log_2 x\rangle. \tag{27}$$

Applying this inequality to Eq 26, we have

$$\log_2 P_c = \log_2\langle P(\{\sigma_i\})\rangle \geq \langle \log_2 P(\{\sigma_i\})\rangle, \tag{28}$$

$$\Rightarrow -\log_2 P_c \leq -\langle \log_2 P(\{\sigma_i\})\rangle. \tag{29}$$

But

$$-\langle \log_2 P(\{\sigma_i\})\rangle = -\sum P(\{\sigma_i\}) \log_2 P(\{\sigma_i\}) = S, \tag{30}$$

and so we have a lower bound on the entropy,

$$S \geq S_{\rm Ma} = -\log_2 P_c. \tag{31}$$

We will refer to this as the "Ma bound," after Ref. [33].

In Fig. 15 we show the various entropy estimates for all the natural courts we consider, ordered by the number of votes recorded for each court (sample size). We see that entropy estimates based on extrapolation, as in Fig. 14, are consistently $\sim$10 % above the Ma bound, and this is true across the full range of sample sizes. Indeed, the entropies (and the Ma Bounds) for the different natural courts themselves vary by only $\pm$10 %, suggesting that the structure of voting patterns is quite stable across the decades. Although there are $2^9 = 512$ possible voting patterns, the fact that the entropy is consistently $S \sim 5$ bits

indicates that, effectively, the court uses only $2^S \sim 32$ of these patterns. But with these few patterns, even one hundred votes is enough to generate a reasonably good sampling. It should be noted that entropy estimates based on extrapolation can easily violate the Ma bound if the number of samples we have in the data is genuinely too small (see, for example, Ref. [29]). Taken together, these results suggest strongly that our entropy estimates are reliable for all of the natural courts, and hence we can compute the multi–information and assess the fraction of this which is captured by the maximum entropy model, as discussed in the text.

For the symmetrized data, each justice is equally likely to vote yes or no, and hence if they voted independently the entropy would be exactly $S_1 = 9$ bits. Then we can read from Fig. 15 our estimate of the multi–information, $I_K^{\rm est} = S_1 - S_\infty$, as well as the multi–information captured by the pairwise maximum entropy model, $S_1 - S_2$. The resulting fraction $F = (S_1 - S_2)/I_K^{\rm est} = 0.95 \pm 0.03$, where the error bar is the standard deviation across the set of natural courts. Alternatively, we know that the multi–information must be greater than $I_K^{\rm Ma} = S_1 - S_{\rm Ma}$, and we thus can conclude that $F \geq (S_1 - S_2)/I_K^{\rm Ma} = 0.83 \pm 0.03$; this is a true bound, and hence a conservative estimate.

We also need to estimate other information theoretic quantities, such as the mutual information between individual justices' votes and the court majority, $I(\sigma_i; \gamma)$. But these quantities involve probability distributions over many fewer states, and thus the sampling issues discussed here are negligible.

Appendix 6: Ideologically Defined Votes

As noted in the main text, the definition of "yes" and "no" votes in Supreme Court decisions has an element of arbitrariness, since the question before the Court is always to affirm or overturn a previous decision. In our initial approach to the data, we elevated this arbitrariness to a symmetry, imagining that every case could have come with the opposite definition of yes and no, so that voting patterns $\{\sigma_i\}$ and $\{-\sigma_i\}$ should be equally probable. An alternative is to note that each case presents an issue that can be mapped to the state of national politics, and there is (except for rare cases) a reasonable consensus that someone with leftist tendencies would vote one way and someone with rightist tendencies the other. This is, of course, only one dimension along which cases may vary.

As we discuss in the text, we map each yes/no vote onto a right/left decision, an asymmetry noted previously [11,13,21]. Here, we note some difficulties with labeling the ideology of cases as also mentioned in Ref. [10]. First, there is a problem of circularity, since ideologies are defined partly by the actors themselves. Thus, it is not truly an external, fixed measure along which we can consider the votes of SCOTUS. Instead, the axis is partially defined by the internal dynamics of the system. The problem of circularity then implies a second problem of non-stationarity since the definition of liberal and conservative positions evolve over time. Spaeth et al. have adjusted for these changes over time, but it is difficult to gauge what the association between ideological ideals and votes may be at a given time. Overall, there is evidence for an important unidimensional space similar to our intuitive concepts of conservative versus liberals, but how this axis overlaps with our intuitions seems inexact.

As we note in the main text, our new model in Eq. (6) includes ideological biases (Fig. 16), but does not seem to break the symmetry encoded by the couplings very strongly. We compare the couplings from the symmetrized and ideological models in Fig. 17.

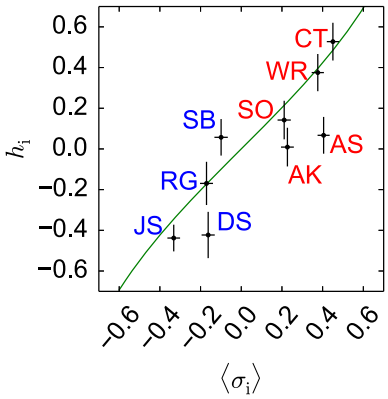

Fig. 16 Bias h_i against mean votes $\langle \sigma_i \rangle$. *Green line* traces the function $\langle \sigma \rangle = \tanh(h)$, as would be expected if each justice voted independently with bias h_i. *Error bars* on h_i are the standard deviation across independent constructions of maximum entropy models from multiple bootstrap samples, as for Σ_J in Fig. 12, while errors in $\langle \sigma_i \rangle$ arise as usual from counting statistics (Color figure online)

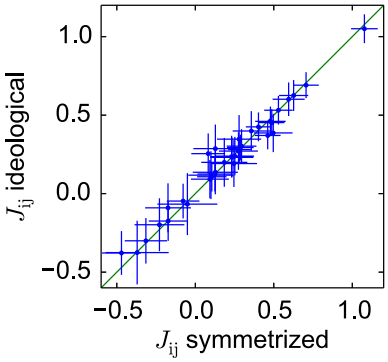

Fig. 17 Comparison of couplings from ideological against symmetrized data. The correlation coefficient between average couplings over bootstrap samples is 0.98. *Error bars* are the standard deviations of our estimates, as in Fig. 12A. *Green line* is 1:1 (Color figure online)

Appendix 7: Model Complexity

The maximum entropy models that we consider assign a probability to every pattern of votes on the court, $P(\{\sigma_i\})$, and we can thus compute the likelihood of observing the data set as

$$\mathcal{L} = \prod_{n=1}^{K} P\left(\left\{\sigma_i^{(n)}\right\}\right), \tag{32}$$

where $\{\sigma_i^{(n)}\}$ is the pattern of votes in decision n, and we have data on K decisions ($K = 895$ for the Second Rehnquist court on which we focus most of our attention). Table G1 shows the log-likelihood results for various models.

In the main text, we have contrasted models that provide ideological labels on the cases with those that treat yes and no votes symmetrically. But we can also consider, within each class, models that match different expectation values in the data. With the pairwise symmetric model, we match the pairwise correlations, but we could also match higher order correlations. We show in Table 2 that matching fourth order correlations alone does not do better than matching the pairwise correlations, but matching *both* the pairs and the quadruplets does help. The improvement, however, is small, $\Delta \log_2 \mathcal{L}/K \sim 0.2$ bits. At the same time, adding fourth order correlations means including 126 more parameters. Is the gain in descriptive power sufficient to justify this added complexity?

In a nested sequence as with adding higher order correlations to maximum entropy models [19], the more complex models always give a better account of the data. Thus, we must

Table 2 Log-likelihoods for the symmetrized and ideological models when constraining different combinations of correlations

The rightmost column is the log-likelihood per data point with $K = 895$ for the Second Rehnquist court. The models we discuss in the text are the symmetrized model matching correlations of order 2 and the ideological model matching correlations of orders 1 and 2

Correlations of order	Number of parameters	$-\log_2 \mathcal{L}$	$-(\log_2 \mathcal{L})/K$
Symmetrized model			
2	36	4534	5.1
4	126	4698	5.2
2,4	162	4343	4.9
6,8	92	5385	6.0
8	9	6999	7.8
Ideological model			
1	9	7539	8.4
2	36	4534	5.1
1,2	45	4233	4.7
3	84	7009	7.8
1,2,3	129	4120	4.6
7,8,9	46	6847	7.7
8,9	10	6997	7.8

penalize the increased complexity although some argue that the penalty should be imposed externally and others that the process of model learning can be defined with sufficient generality that such a penalty emerges naturally. Here, we take the second view following a stream of work on Bayesian methods that is described at length in Sect. 6.5 of Ref. [25].

When we build a model to describe data, we are choosing from the family of models with different possible parameter values. We should imagine that, *a priori*, there is some measure, or probability distribution, on this parameter space. As we accumulate data, our estimate of the parameters narrows around the best values; roughly, we expect that the size of the region consistent with the data is of length $\sim 1/\sqrt{K}$ along each parameter. In a model with l parameters, this corresponds to a volume $\Omega \sim K^{-l/2}$. If we compute the likelihood of the data not by focusing on the very best fit parameters, but by averaging over the possible values of parameters—relying on the data to concentrate the average on the small neighborhood around the best fit—then the likelihood picks up a phase space factor $\sim \Omega$. Thus the log-likelihood has a term $\log_2(\mathcal{L})_p \sim -(l/2)\log_2 K$. While increasing l increases the likelihood of the data in the best fit model, this penalty on complexity reduces the total likelihood.

Another point of view is that the model provides a literal description, or code for the data. The usual computation of the log-likelihood gives, by Shannon's classical arguments, the minimum number of bits needed to represent the data, but this does not account for the space needed to represent the parameters. Since we start with the possible parameters in some range, with linear dimensions of order one, and we end up with parameters confined to a volume Ω, the reduction in entropy is $\log_2(1/\Omega)$ bits, the minimum amount of space needed to represent our knowledge of the parameters. This is the same as $-\log_2(\mathcal{L})_p$ above.

Thus, whether we think about computing total probabilities in the family of models, or using the models as a code for the data, there is a natural "penalty for complexity" that grows linearly with the number of parameters and logarithmically with the number of samples. When we add 126 parameters to a model, and have only $K = 895$ samples, this is a huge effect, $-(\log_2(\mathcal{L})_p)/K \sim 0.69$ bits. This reduction in the volume of parameter space far outweighs the 0.2 bit gain in descriptive power. We conclude that, within the class of symmetrized models, explicitly matching higher order correlations is not worth the added complexity, and our simple pairwise model is favored. A similar argument can be applied within the class of ideological models, where the tiny gains in descriptive power from matching higher order correlations are not sufficient to overcome the natural penalty for added complexity.

We also compare the symmetric model to the ideological. The pairwise ideological model is an improvement by $\Delta \log_2 \mathcal{L}/K \sim 0.3$ bits, and this gain outweighs the added complexity from nine additional parameters of only 0.05 bits. But in the ideological models, we have extra information because every case is labeled as a conservative or liberal vote. This extra information that we need in order to fully describe the data is the Kullback–Leibler divergence between the distribution $P(\{\sigma_i\})$ and a model in which the voting patterns are symmetrized, or 0.3 bits per sample. Including both the cost of nine extra parameters and the ideological information, the ideological model provides almost the same total description length for the data as does the symmetrized model. Thus, in accord with the qualitative impressions discussed in the main text, there is no statistical evidence that keeping track of ideological biases provides a better account of the distribution of voting patterns than that which emerges from our symmetrized model.

References

1. Fortunato, S., Castellano, C.: Scaling and universality in proportional elections. Phys. Rev. Lett. **99**, 138701 (2007)
2. Castellano, C., Fortunato, S., Loreto, V.: Statistical physics of social dynamics. Rev. Mod. Phys. **81**, 591–646 (2009)
3. Fortunato, S., Macy, M., Redner, R.: Editorial. J. Stat. Phys. **151**, 1–8 (2013). This is an introduction to a special issue of J Stat Phys on "The application of statistical mechanics to social phenomena."
4. Guimerà, R., Sales-Pardo, M.: Justice blocks and predictability of US Supreme Court votes. PLoS One **6**, e27188 (2011)
5. Schneidman, E., Berry II, M.J., Segev, R., Bialek, W.: Weak pairwise correlations imply strongly correlated network states in a neural population. Nature **440**, 1007–1012 (2006)
6. Lezon, T.R., Banavar, J.R., Cieplak, M., Maritan, A., Federoff, N.V.: Using the principle of entropy maximization to infer genetic interaction networks from gene expression patterns. Proc. Natl. Acad. Sci. (USA) **103**, 19033–19038 (2006)
7. Seno, F., Trovato, A., Banavar, J.R., Maritan, A.: Maximum entropy approach for deducing amino acid interactions in proteins. Phys. Rev. Lett. **100**, 078102 (2008)
8. Bialek, W., Cavagna, A., Giardina, I., Mora, T., Silvestri, E., Viale, M., Walczak, A.: Statistical mechanics for natural flocks of birds. Proc. Natl. Acad. Sci. (USA) **109**, 4786–4791 (2012)
9. Tkačik, G., Marre, O., Amodei, D., Schneidman, E., Bialek, W., Berry, M.J. II.:Searching for Collective Behavior in a Large Network of Sensory Neurons. PLoS Comput. Biol. **10**, e1003408 (2014)
10. Spaeth, H.J., Epstein, L., Ruger, T.W., Whittington, K., Segal, J.A., Martin, A.D.: Supreme Court database. Version 2011 Release 3 (2011). http://scdb.wustl.edu/index.php. Accessed 3 April 2012
11. Martin, A.D., Quinn, K.M., Epstein, L.: The median justice on the United States Supreme Court. NC Law. Rev. **83**, 1275–1322 (2004)
12. Segal, J.A., Spaeth, H.J.: The Supreme Court and the Attitudinal Model Revisited. Cambridge University Press, New York (2002)
13. Sirovich, L.: A pattern analysis of the second Rehnquist US Supreme Court. Proc. Natl. Acad. Sci. (USA) **100**, 7432–7437 (2003)
14. Jaynes, E.T.: Information theory and statistical mechanics. Phys. Rev. **106**, 620–630 (1957)
15. Shannon, C.E.: A mathematical theory of communication. Bell. Syst. Tech. J. **27**, 379–423 & 623–656 (1948)
16. Cover, T.M., Thomas, J.A.: Elements of Information Theory. Wiley, New York (1991)
17. Epstein, L., Segal, J.A., Spaeth, H.J.: The norm of consensus on the US Supreme Court. Am. J. Pol. Sci. **45**, 362–377 (2001)
18. Black, D.: On the rationale of group decision-making. J. Pol. Econ. **56**, 23–34 (1948)
19. Schneidman, E., Still, S., Berry II, M.J., Bialek, W.: Network information and connected correlations. Phys. Rev. Lett. **91**, 238701 (2003)
20. Mézard, M., Parisi, G., Virasoro, M.A.: Spin Glass Theory and Beyond. World Scientific, Singapore (1987)
21. Kemp, C., Tenenbaum, J.B.: The discovery of structural form. Proc. Natl. Acad. Sci. (USA) **105**, 10687–10692 (2008)

22. Lee, E.D.: Information in Justice and Conflict: Formulating a Quantitative Approach to Social Data,Senior Thesis, Princeton University (2012)
23. Zipf, G.K.: Human Behavior and the Principle of Least Effort. Addison-Wesley, Cambridge (1949)
24. Martin, A.D., Quinn, K.M.: Dynamic ideal point estimation via Markov chain Monte Carlo for the US Supreme Court, 1953–1999. Political Anal. **10**, 134–153 (2002)
25. Bialek, W. (ed.): Biophysics: Searching for Principles. Princeton University Press, Princeton (2012)
26. Apostol, T. (ed.) : Calculus. Volume II: Multi-Variable Calculus and Linear Algebra with Applications. 2nd edn Wiley, New York (1969)
27. Miller, G.A.: Note on the bias of information estimates. In: Quastler, H., (ed.) Information Theory in Psychology: Problems and Methods II-B, pp. 95–100 Free Press, Glencoe (1955)
28. Treves, A., Panzeri, S.: The upward bias in measures of information derived from limited data samples. Neural Comput. **7**, 399–407 (1995)
29. Strong, S.P., Koberle, R., de Ruyter van Steveninck, R.R., Bialek, W.: Entropy and information in neural spike trains. Phys. Rev. Lett. **80**, 197–200 (1998)
30. Paninski, L.: Estimation of entropy and mutual information. Neural Comput. **15**, 1191–1253 (2003)
31. Nemenman, I., Bialek, W., de Ruyter van Steveninck, R.R.: Entropy and information in neural spike trains: progress on the sampling problem. Phys. Rev. E **69**, 056111 (2004)
32. Feller, W.: Probability Theory and Its Applications, vol. I. Wiley, New York (1950)
33. Ma, S.K.: Calculation of entropy from data of motion. J. Stat. Phys. **26**, 221–240 (1981)