QUANTITATIVE MODELING OF COLLECTIVE BEHAVIOR

A Dissertation

Presented to the Faculty of the Graduate School

of Cornell University

in Partial Fulfillment of the Requirements for the Degree of

Doctor of Philosophy

by

Edward D. Lee

December 2019

QUANTITATIVE MODELING OF COLLECTIVE BEHAVIOR

Edward D. Lee, Ph.D.

Cornell University 2019

We show how ideas, models, and techniques from statistical physics prove useful for building quantitative theories of collective behavior with a focus on social systems.

In the first chapter, we develop a statistical mechanics of political voting on the US Supreme Court. By building minimal models of voting behavior, we find that signatures of strong consensus and partisanship are captured by a maximum entropy model only relying on the pairwise correlations between voters. We extend this maximum entropy approach to collective voting outcomes on a Super Supreme Court, inferring how the set of historically disjoint justices from 1946–2015 might have voted with one another. When we measure how correlations decay between justices across time, we find a long, institutional timescale approaching a century, a quantitative signature of historical precedent. Beyond consensus, we find that voting blocs fracture in many possible ways, belying a common assumption that partisan intuition generalizes to the history of the court; actually, Supreme Court voting over time is immensely more complex. Then, we use minimal models to measure how sensitive collective outcomes are to perturbations to "pivotal components," akin to how majority outcomes are sensitive to swing voters. We demonstrate how to extract these pivotal components from an information-geometric analysis of example social systems including Twitter, financial markets, legislatures, and judicial courts. Our approach

presents a principled, quantitative step towards characterizing the robustness of social institutions to changes in component-level behavior.

The second topic we address is about conflict dynamics. In a society of pigtailed macaques and in armed conflict in human society, we find remarkable, emergent regularities suggesting that conflict is dominated by a low-dimensional process that scales with physical dimensions in a surprisingly unified and predictable way. For macaque conflict, we discover a temporal scaling collapse for conflict duration distributions. This collapse indicates the presence of long-time correlations that connect early conflict events with later ones. We propose a model that explains this collapse and consider how we might predict conflict evolution. For armed conflict, we find that social and spatiotemporal properties of conflict can be unified by a reduced scaling framework, and we make initial steps towards a model that captures the dynamics of observed properties of conflict.

The final topic we address is that of interpersonal coordination of motion. We describe an experimental apparatus that combines a commercial virtual reality platform, a human motion-capture suit, and a mirroring game with an avatar. We use this apparatus to show how frequency-based auditory cues enhance the ability to mirror motion. Our work lays the groundwork for future experiments: better understanding of how information is encoded in visual or auditory cues could facilitate joint coordination when navigating visually occluded environments, improve reaction speed in human-computer interfaces or measure altered physiological states and disease.

BIOGRAPHICAL SKETCH

Edward Dongmyung Lee was born in Silver Spring, Maryland on October 14, 1990 to two immigrants from South Korea. After an unsuccessful start to a musical career as a teenager, Edward became interested in interdisciplinary physics at Princeton University, where he earned his A.B. in Physics and Certificate in Biophysics in 2012. Following an immensely enjoyable period of open-ended research at the Lewis-Sigler Institute of Princeton University and at the Wisconsin Institute for Discovery of the University of Wisconsin-Madison, he enrolled in the Department of Physics at Cornell University to earn his doctorate in Physics in 2014. In 2018, he earned his Masters of Science from Cornell. He received his Doctorate in Philosophy in December 2019 under Professor Paul H. Ginsparg. While earning his doctorate, he become an avid fan of bboy/bgirl culture and a dancer, and subsequently he became a Omega Miller Program post-doctoral research fellow at the Santa Fe Institute under Professors Jessica C. Flack and David C. Krakauer.

To my parents, who made my endeavors possible.

ACKNOWLEDGEMENTS

The thesis is the culmination of a long journey, and there have been many people who have guided me in ways both great and small. I thank all of those who have obligingly led me through the academic world. In chronological order, the mentors who have lighted the way: Bill Bialek who showed me how to savor the journey that a research paper entails. Chase Broedersz, then as a Lewis-Sigler Fellow and now friend, guided me through the grinding effort of writing my first paper. My long-time mentors and friends Jessica Flack, David Krakauer, and especially Bryan Daniels. They, above anyone else, have shaped my scientific personality and perspective. At Cornell, I have matured as a researcher under the tutelage of Itai Cohen, Jim Sethna, and my advisor Paul Ginsparg. I have also had the fortune of learning from a number of patient members of the Physics Department including Veit Elser, Guru Khalsa, and Chris Myers.

I have also benefited from many insightful discussions with my co-authors Dan Katz, Mike Bommarito, and Ted Esposito and colleagues Danilo Liarte, Colin Clement, Brian Leahy, Archishman Raju, Katherine Quinn, Matt Bierbaum, Jaron Kent-Dobias, Giles Hooker, and other members of the Sethna and Cohen groups. Without this vibrant research community, I would not have discovered many of the inspiring ideas that inform my research today.

I must thank my parents and brothers for supporting me through these long, sometimes arduous, years, and I must thank my friends who helped me preserve my sanity.

I thank the wonderful administration of the Physics Department who tried their best to smooth the trajectory for graduate students, especially Kacey Acquilano,

Rosemary Barber, Debra Hatfield, Sue Sullivan, Craig Wiggers, and tech guru Barry Robinson.

Finally, all research requires financial investment. I acknowledge generous support from the National Science Foundation, Dirksen Congressional Research Center, National Institutes of Health, Army Research Office, Cornell University, American Physical Society, Templeton Foundation, St. Andrew's Foundation, Santa Fe Institute, and Omega Miller Program.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

FOREWORD

...the spirit of physics, the idea of discovery, the idea of understanding.

- Hans Bethe

...'thinking like a physicist' is supposed to mean something, and it is this, above all else, that we try to convey to our students.

- William Bialek

What is Physics?

This question has dogged me in one form or another during graduate school, arising in conversation, lectures, and referee reports.

I was lucky that this question, cornering me into the box marked "physicist," only arose during my graduate career. As an undergraduate of the Integrated Sciences program at Princeton, I was introduced to physics as the mathematical framework tying together our understanding of Nature including the many wonders of Biology. What was for me the "normal" approach to physics turned out to be somewhat avante-garde. As a research experience undergraduate at the Santa Fe Institute, I was challenged to explore big questions that didn't hew to disciplinary boundaries, but invoked a heretical mix of transciplinary incantations. And finally, at the forward-looking Physics Department at Cornell, I found thinkers who weren't interested in defining the limitations of their disciplines, but rather in dismantling the walls to build cross-disciplinary bridges. Disciplinary walls keep thoughts in as much as they keep wanderers out.

Physics has made striding advances and correspondingly has evolved over the centuries. Through the Copernican revolution, classical mechanics, quantum

mechanics, statistical mechanics, nuclear physics, astrophysics, biophysics, physicists have incorporated entire new volumes into the disciplinary canon. What was once, sometimes unacceptably, new—be it statistical mechanics or astrophysics—is now standard. I hope that this dissertation represents a step into the wilderness guided by the question,

What could the next chapter in Physics become?

CHAPTER 1

POLITICAL VOTING

Political voting provides a data-rich opportunity for studying collective human behavior. Both legislative and judicial bodies in the US keep extensive historical records specifying the rules of the body, membership, transcripts, text of legislative bills, court opinions, and votes. Committees in the executive branch also make decisions by voting such as on the Federal Reserve Board and the Export-Import Bank. Beyond these official records, there are often records of interaction with the public on news media or social media. These sources of data provide a detailed timeline of political behavior and possibly insight into how politicians make decisions. Since political decisions can have major social consequences, it is useful to be able to anticipate them or to explain them quantitatively.

We focus on votes because they are a distilled representation of the dynamics that lead to a political outcome. Voting behavior presents a testing ground for honing our techniques, leading us towards an increasingly detailed understanding of political behavior. Of the reasons for which voting is particularly amenable are that votes are straightforward to map to physical models and are accessible in digital format—though seemingly trivial is really essential.[1] Here, we develop a statistical physics approach to political voting that emphasizes sparsely-parameterized models, use it to characterize collective aspects of vot-

[1] Until the advent of the Supreme Court Database Project, for example, report (final) votes on the US Supreme Court had to be determined from reading physical copies of the opinions. This is straightforward in principle but presents major practical difficulties. Even more inaccessible are the conference votes, which are votes made in private by the court previous to the report vote. Some of these are recorded in now public docket records kept by previous justices. The author at one point visited the Congressional Research Library to look at these docket records and can appreciate the labor it takes to decipher such records.

ing, and touch on the topic of intervention.

We discuss voting on the US Supreme Court in the context of a single *natural court*, a period during which the membership of the nine justices remains fixed, using a minimal model that agrees closely with the data and recapitulating work published in reference [36]. Then, we extend our analysis to consider simultaneously many natural courts using the fact that the members overlapped in tenure to infer how historically disjoint justices might have voted together [35]. As in physical systems where the most interesting properties arise when there are interactions between components, we find that voting on the court is characterized by strong competing interactions that generate collective structure beyond the binary partisan picture—a widespread simplification of the court as divided between liberal and conserative factions. This picture is violated in an irreconciliable way when we look at Supreme Court voting over time as is published in reference [35], where we discover that any low-dimensional description of voting falls short of describing the wide variety of potential outcomes. Results like these show how an approach informed by statistical physics can lead to new, even surprising, insights about political voting.

The statistical framework that we develop opens the door to many other interesting questions. One such question is how sensitive collective outcomes are to changes in individual behavior. For example, swing voters in majority-rule voting systems are often the focus of those wishing to change the outcome. Building on observations gleaned from a reduced toy model capturing the essential features of the swing voter, we use the information geometry of our minimal models to search for *pivotal components* on which collective outcomes are

the most sensitive. We survey systems including Twitter, financial indices, state and federal legislatures, and state and federal courts [37]. We find large variety in behavior, but also hints that institutional structure might strengthen pivotal components, ultimately rendering collective outcomes highly dependent on a few system components. This observation suggests that, by comparing many systems, we could determine the factors that make an institution more or less robust to changes in individual behavior and perhaps use that knowledge to better design our institutions.

1.1 Statistical mechanics of the US Supreme Court

The US Supreme Court is the highest court in the nation and constitutes the third complement to the legislative and executive branches as is articulated in the US Constitution. In its modern form, it consists of nine justices who are nominated by the president, confirmed by the Senate, and appointed for life. Though the court is often portrayed as impartial, it is political [61], and its decisions consequential. Perhaps because of its importance, patterns of internal division, and tractable size, its voting behavior has attracted much research interest across many fields [68].

Here, we focus on the second Rehnquist Court (lasting from 1994–2005 with 909 recorded nine-member votes). The court writes majority and minority opinions, sometimes supplemented with other opinions. Although these can be nuanced (e.g., dissenting in part, providing alternative legal justification), the justices must choose to vote in the majority or minority:[2] each justice casts a yes ($s_i = 1$) or no ($s_i = -1$) vote, and the majority decides the outcome. 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 $\{s_i\}$ and $\{-s_i\}$ are equally likely, and we return to this problem below. With this symmetry, the average vote is neutral, $\langle s_i \rangle = 0$ for all justices.

[2] We are not considering cases where any justice is recused or absent.

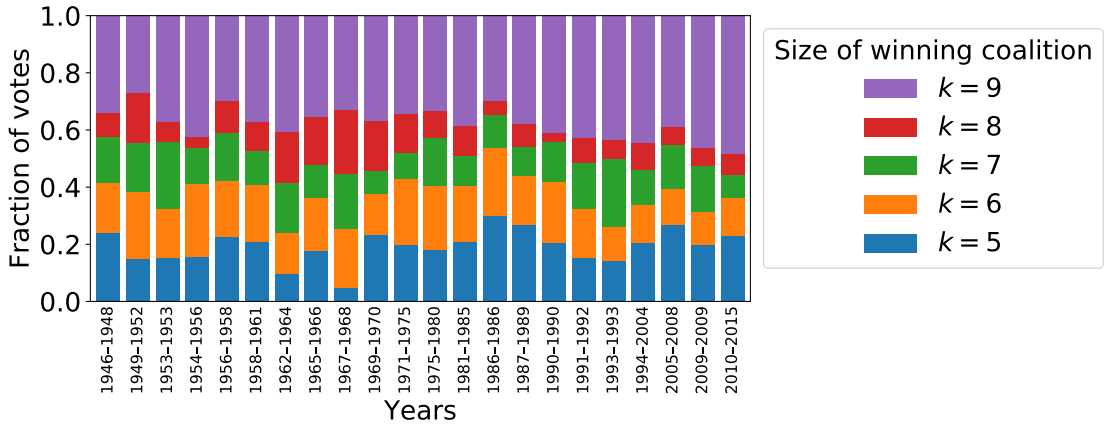


Figure 1.1: Majority-minority divisions on US Supreme Court (SCOTUS) over time. These include all nine-member votes from natural courts who voted at least 30 times together.

1.1.1 A pairwise maximum entropy approach

One of the major themes in the study of the court is the partisan divide [44, 61]. In the modern Supreme Court, which is considered to have begun in 1946 when dissenting rates rose substantially, the justices are often discussed as divided into ideological left and ideological right. In truth, such a divide only characterizes a fraction of the decisions. The plurality of decisions for a given natural court, a period of time during which membership is fixed, is usually unanimity as is given by the probability $q(k)$ of having k votes in the majority in Figure 1.1. It is only behind this strong tendency to unanimity that partisanship is apparent in the voting record.

It is immediately clear from inspecting the distribution dissenting votes $q(k)$ that each justice does not vote independently of the others. If they were independent, the distribution of votes would be described by a binomial distribution and the

probability of k votes in the majority would be

$$q(k) = 2^{-8}\binom{n}{k}. \tag{1.1}$$

A unanimous vote would be unlikely, $q(k = 9) = 2^{-8}$, which is evidently far from the case for any of the natural courts shown. These examples demonstrate that strong correlations between justices are essential for explaining how they vote.

To explain such structure, we begin by writing the average votes, or the "mean magnetizations," that were fixed by the up-down symmetry previously as

$$\langle s_i \rangle = \sum_s p(s) s_i = 0. \tag{1.2}$$

Eq 1.2 involves a sum over all possible votes $s \equiv \{s_i\}$. We will relax the assumption of up-down symmetry later when we consider the ideological direction of a vote in a case. Then, a good model should be able to capture the statistics of the data as described fully by the set of all correlations without imposing excessive complexity [43]. We turn to the principle of maximum entropy (maxent) to formalize these criteria.

The information entropy of a statistical model describes the amount of statistical structure encoded in it. As Shannon proved in his seminal work [64], the unique measure of uncertainty up to a scaling factor given three basic assumptions is

$$S = -\sum_s p(s) \ln p(s). \tag{1.3}$$

The Shannon entropy is maximized when the probabilities of all states s, here different votes, are equally likely. The statement that all configurations

are equally likely is a restatement of the equipartition principle, in which case the Shannon entropy reduces to the classical thermodynamic entropy $S \propto \ln[\#\,\text{of states}]$. At the other extreme, $S = 0$ when only a single configuration is possible with unit probability, and we have perfect information about the state of the system. If we imagine starting with no prior knowledge about a system, we will through observation constrain the probabilities in configuration space and reduce the entropy of model in a way that reflects additional incorporated structure. By maximizing the entropy as we constrain the model, we ensure that we infer models from a set of observations while being explicit about the information that we have used.

How do we choose which properties of the data to impose on the model? From the physics perspective, it is most natural to consider how structures emerge from lower-order interactions, so we are motivated to consider a maxent model that agrees with the pairwise correlations while ignoring higher-order constraints as a first step. So, we start by specifying the joint probability distributions of every pair of variables s_i and s_j. It is entirely reasonable to fit higher-order patterns first or any combination thereof, and we might do that by comparing how quickly the model converges to the data as in reference [36] or by identifying features that are essential for a close fit to the data [14, 16, 65].[3] For the Supreme Court as we consider here, the set of pairwise correlations efficiently captures much of the probability distribution $p(s)$ [36].

[3]When starting with higher order correlations, however, there arises an issue in the choice of representation. An entirely equivalent way to define the votes is as $s_i \in \{0, 1\}$ with a concomitant change in the fields and couplings such that any correlation in one basis immediately includes a combination of lower order correlations in the other basis. By building from the lowest order correlations up, we avoid this issue entirely.

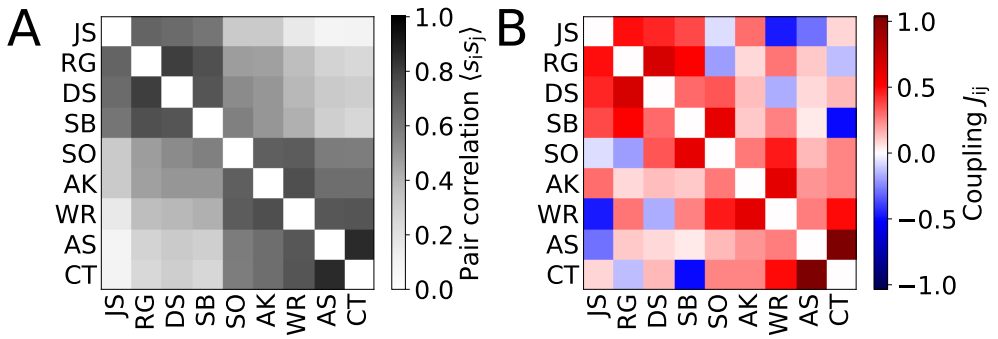


Figure 1.2: (A) Pairwise correlations $\langle s_i s_j \rangle$ for Second Rehnquist Court. The standard error in estimating C_{ij} is given by $\delta C_{ij} = [(1 - C_{ij}^2)/K]^{1/2}$ with $K = 909$ votes, and we find $\delta C_{ij} < 0.03$ for all ij. (B) Solved couplings J_{ij}. Justices are ordered from most liberal to most conservative according to a standard political science score of ideology [44]. Their full names are John P. Stephens, Ruth B. Ginsburg, David H. Souter, Stephen G. Breyer, Sandra D. O'Connor, Anthony M. Kennedy, William H. Rehnquist, Antonin G. Scalia, and Clarence Thomas.

To wring any additional statistical structure out of the model beyond the specified constraints, we use the standard method of Lagrangian multipliers to maximize the Shannon entropy.[4] We construct a Lagrangian functional of the form

$$\mathcal{L}[p] = -\sum_s p(s) \log p(s) - \frac{1}{2} \sum_{ij} J_{ij} \left\langle s_i s_j \right\rangle - \sum_i h_i \langle s_i \rangle . \tag{1.4}$$

The Langrangian multipliers are the set of "couplings" $\{J_{ij}\}$ and "fields" $\{h_i\}$. The unique optimum of Eq 1.4 is the Ising model with Boltzmann form,

$$p(s) = e^{-E(s)} \Big/ Z, \tag{1.5}$$

normalized by partition function

$$Z = \sum_s e^{-E(s)} \tag{1.6}$$

[4] Another interpretation of this procedure is to recognize that the Lagrangian in Eq 1.4 is the Helmholtz free energy where the units $k_B T$ have been set to unity. So it is the case that the maximization of entropy is really the minimization of free energy—after all they are Legendre transforms of one another—if free energy be a more familiar way of describing the situation.

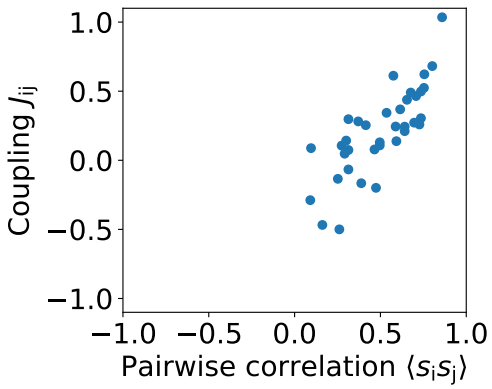


Figure 1.3: Comparison of pairwise correlations with the corresponding couplings for the Second Rehnquist Court.

and Hamiltonian

$$E(s) = -\frac{1}{2}\sum_{ij}^{N} J_{ij} s_i s_j - \sum_i h_i s_i. \tag{1.7}$$

Since we insisted that the average of every vote was $\langle s_i \rangle = 0$, all the fields $h_i = 0$, and this ensures that there is no bias in the vote. When two voters, or spins, are aligned and share a positive coupling $J_{ij} > 0$, this lowers the energy and enhances the probability of observing the vote s by a factor of $e^{J_{ij}}$. When the coupling is negative, the pair tends to align against each other, lowering the probability of the vote by a factor of $e^{-|J_{ij}|}$ instead.

The question of now solving for the couplings that match the pairwise correlations is straightforward in principle but difficult in practice. Since for each coupling we have a corresponding pairwise correlation, there is no parameter fitting because there is a single solution for the data. However, the equations are nonlinear and it becomes exponentially expensive to calculate the partition function with system size N. For the Supreme Court, we are fortunate that $N = 9$, and it is straightforward to enumerate the partition function and find the solution with standard optimization algorithms. For larger N, more sophisticated

approximation and sampling techniques must be used like those that are discussed in Appendix A.1.

In Figure 1.2A, we show the matrix of pairwise correlations between the justices ordered from ideological left to ideological right for the Second Rehnquist Court. Consistent with this ideological picture, the strongest pairwise correlation, $\langle s_{\mathrm{AS}} s_{\mathrm{CT}} \rangle = 0.86 \pm 0.02$, is unsurprisingly between the pair of highly conservative justices, A.S. and C.T. The "swing voters" S.O. and A.K. straddle the divide, though they tend to vote more with the conservative voters than the liberal ones. Yet, despite the frequent emphasis on partisan lines, the entire matrix of correlations is positive, demonstrating that the court is actually dominated by unanimity. Ideological divisions manifest as blocks of stronger correlations between the ideological wings within this consensus-dominated matrix.

The couplings, shown in Figure 1.2B, are clearly different. They span a range of both positive and negative values, reflecting antagonistic tendencies within the court that are buried within the entirely positive set of correlations. The strongest positive couplings tend to concentrate close to the diagonal which is consistent with the idea that there is an ideological spectrum rather than a simple bipartite structure to the interaction network. Finally, it is clear that Justices A.S. and C.T. share strong tendencies to agree, as we found in the correlations. Despite such satisfactory agreement with partisan intuition for the court, it is important to point out that the strong correlation between A.S. and C.T. does not alone imply a strong coupling. Couplings describe change to the log-probabilities, whereas pairwise correlations are a marginalization of this probability distribution, and as we show in Figure 1.3 these two sets share no simple

relationship. Thus, the nontrivial mapping between the couplings and the correlations mean that couplings provide an alternative way of extracting insight about the interactions in the system.

A common misconception is that the couplings represent physical or even causal interactions betweeen voters, confusion that perhaps arises from the nomenclature "interactions." The couplings, by definition, are given by the set of pairwise correlations as specified by the principle of maximum entropy in Eq 1.4. Thus, in the strictest sense the couplings represent statistical interactions. There is, however, evidence that couplings reflect physical interactions in protein folding, where Coulomb forces and thermal noise generate an ensemble of protein shapes [74].[5] For bird flocks, the topological nature of the coupling network (i.e., starlings tend to follow a fixed number of neighbors) suggests that the couplings capture social interactions mediated through perception [7]. In the case of a voting system like SCOTUS, the analogy is less clear, and the couplings reflect direct social interaction between voters and indirect ones like those induced by the selection of cases faced by the court.

We must check if the model with only pairwise correlations is sufficient to capture the higher-order statistics of the system without explicitly having fixed the higher-order correlations. A useful property of maxent models is that the entropy must decrease monotonically as we add further constraints and is minimized when the model matches the data exactly. With this in mind, we define a sequence of maxent models that incorporate all correlations up to order n with

[5] Maxent techniques have been used as an inference procedure to identify residue contacts in protein-protein interactions. Relying on statistical correlations in sequence, Weigt et al. were able to identify through the couplings pairs of proteins with physical interactions without relying on tuned fitting parameters [74].

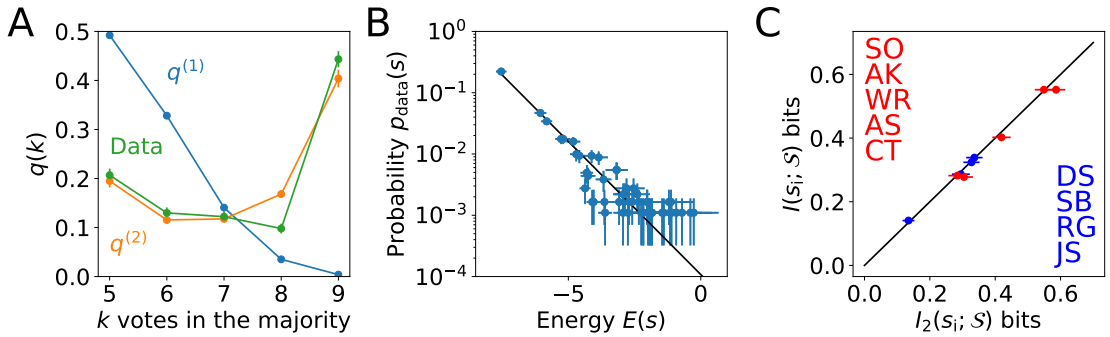


Figure 1.4: Testing the maximum entropy model for the Rehnquist court. (A) Probability of k votes in the majority. We compare the data (green) with the predictions of the pairwise maximum entropy model $p^{(2)}$ (orange), and with independent voters $p^{(1)}$ (blue). (B) Probability of each of the 106 observed voting patterns s versus the "energy" in Eq 1.5; line is Eq 1.7. Errors in probability arise from counting, the standard errors; 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 $p(s)$. (C) Mutual information $I(s_i; \mathcal{S})$ between individual votes s_i and the decision $\mathcal{S}$ of the majority, compared with $I_2(s_i; \mathcal{S})$ from the model. Conservatives are red and liberals blue, from highest $I(s_i; \mathcal{S})$ to lowest according to data.

entropy S_n to measure how much of the total correlation has been captured by the model compared to the first-order model. This is called the fraction of multi-information captured [59],

$$f_n = \frac{S_1 - S_n}{S_1 - S_{\text{data}}}, \tag{1.8}$$

and f_n ranges from $f_1 = 0$, the independent model, to $f_N = 1$, when the model matches the probability distribution in its entirety.[6]

As a more direct method of verification, we show that the pairwise maxent model presents a vast improvement over the independent model in matching the higher-order statistical structure encoded in the probability of k votes in the majority, $q(k)$ (Figure 1.4A). We also show by comparison of the probabilities of

[6]The estimation of entropy is biased for small data sets, an issue that we discuss in further detail in reference [36].

each vote $p(s)$ with the energies $E(s)$, that the model does well predicting almost the entire distribution of votes (Figure 1.4B). This close agreement is reflected in the fraction of multi-information captured of $f_2 = 0.95$. Thus, perhaps surprisingly, the model matching only the pairwise statistics matches features of the data that were not explicitly specified in its construction.

1.1.2 The energy landscape

The competition between positive and negative couplings lead to local minima in the energy landscape defined by Eq 1.7, where we might think of votes that belong to the same basin of an energy minimum as noisy versions of "prototypical" voting configurations; these minima are equivalently maxima in the probability landscape. We find these prototypical votes by taking a vote and flipping a single voter at a time that leads to the largest decrease in energy. When no more votes can be flipped, we have found an energy minimum. For the Second Rehnquist Court, the energy minima that contain over 99% of all possible votes are the unanimous 9–0 vote, the partisan 5–4 divide, and the conservative core 7–2 of everyone against A.S. and C.T. These basins constitute 50%, 32%, and 18% of the probability distribution, respectively.[7]

To visualize how strongly the voting configurations "order" into the prototypical voting states, we measure the overlap of a vote s with the nth most frequent minimum $\xi^{(n)}$

$$m^{(n)} = \sum_{i=1}^{N} \xi_i^{(n)} s_i. \tag{1.9}$$

[7] There is a fourth basin accounting for less than 1% of the distribution consisting of the three most liberal and two most conservative justices, but this basin occurs with frequency on the order of statistical sampling noise.

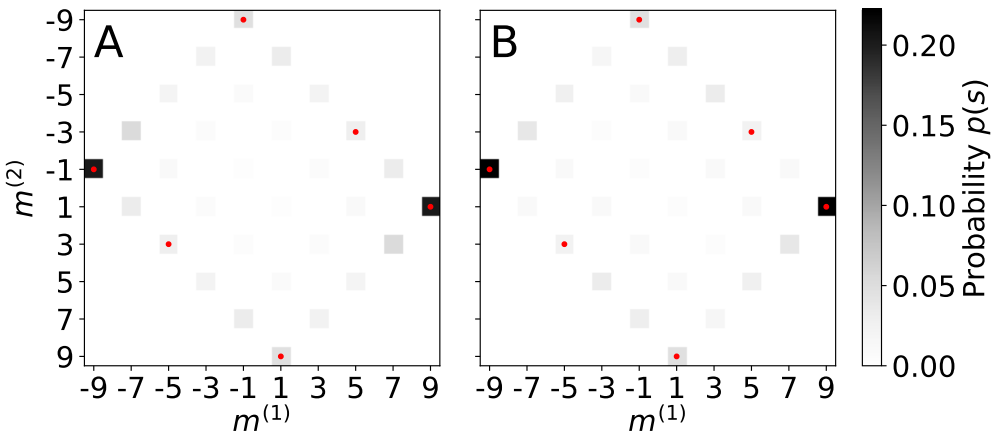


Figure 1.5: Projection of the energy landscape in the data (A) and the predictions of the pairwise maximum entropy model (B). The horizontal axis shows the projection $m^{(1)}$ onto the unanimous +1 basin, and the vertical axis shows the projection $m^{(2)}$ onto 5–4 basin oriented so the majority voters are +1. The 7–2 basin lies in between. Local energy minima are marked with red dots. Note that blocks 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.

In the density of states by overlap with the first two minima as in Figure 1.5, we find that the density is localized on the boundary of the two-dimensional 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 maxent model using only the observed pairwise correlations among votes as inputs.

1.1.3 Measuring influence

This Court is often described as being divided between liberal and conservative voters with the majority outcome being decided by the swing voters in the

middle. To test whether or not the statistics are consistent with this description, we measure the mutual information between each voter and the majority vote,

$$I(s_i; \mathcal{S}) = \sum_{s_i;S} p(s_i; \mathcal{S}) \log_2 \frac{p(s_i; \mathcal{S})}{p(s_i; \mathcal{S})} \tag{1.10}$$

$$\mathcal{S} \equiv \frac{\sum_{i=1}^{N} s_i}{\left|\sum_{i=1}^{N} s_i\right|} \tag{1.11}$$

As we show in Figure 1.4C, the two swing voters S.O. and A.K. have the highest mutual information with the majority outcome of $I(s_{SO}; \mathcal{S}) = 0.59 \pm 0.03$ bits and $I(s_{AK}; \mathcal{S}) = 0.55 \pm 0.03$ bits, which is consistent with the perception that they are important swing voters on the court.

For a system symmetric with respect to the two possible outcomes -1 and 1, the mutual information is a monotonic function of the positive correlation between any voter and the majority vote, so these two different measures of influence contain the same information. We might use our metaphor for a physical system, where we would measure the response of the majority outcome to an applied "field" on a single voter, the susceptibility $X_i = \partial \langle \mathcal{S} \rangle / \partial h_i = \langle s_i \mathcal{S} \rangle$. Yet again, the susceptibility returns the correlation, and it is the case that multiple measures of statistical influence on the outcome are redundant.

As an alternative measure, we imagine what would happen if a justice were to signal a change in a particular direction of voting without actually changing his or her own vote. Given that each Justice i "feels" an effective local field from others $h_i^{\text{eff}} = \sum_{j \neq i} J_{ij} s_j$, what would happen if Justice i exerted a tendency to change on Justice j? We would see a small change in j's local field $\Delta h_{j \neq i}^{\text{eff}}(i) = J_{ij} \epsilon$. Yet, through feedback Justice i's votes will become biased as well. In order to

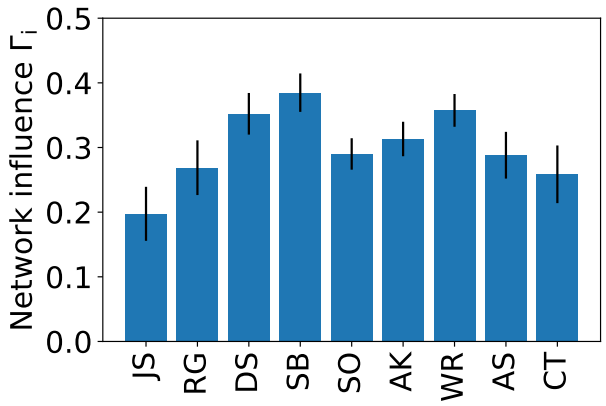


Figure 1.6: Network influence Γ_i of the majority to a signal from Justice i, as defined in Eq 1.12. Note that these values were calculated incorrectly in reference [36].

isolate the influence of Justice i on others, we add to its local field $\Delta h_i^{\text{eff}}(i) = -(\epsilon/\chi_{ii})\sum_j \chi_{ji} J_{ij}$, where $\chi_{ji} = \partial \left\langle s_j \right\rangle / \partial h_i = \left\langle s_i s_j \right\rangle$, and as a result fixing $\langle s_i \rangle = 0$. The resulting susceptibility of the majority vote to a signal from i is

$$\Gamma_i \equiv \lim_{\epsilon \to 0} \frac{1}{\epsilon} \sum_{j=1}^{N} X_j \Delta h_j^{\text{eff}}(i). \tag{1.12}$$

We summarize this measure of network influence in Figure 1.6. This measure of influence Γ_i tends to decrease towards the ideological extremes, in agreement with measures given by the correlation of justice votes with the majority, but it does not peak on the medians. Interestingly, Γ_{SB} is the largest, which suggests that S.B.'s role may be unnoticed but important on the court. On the other hand, W.R. has both the third largest mutual information I and measure of network influence Γ_i, highlighting a relatively important role in the court. This observation would be consistent with the chief justice's keeping of special privileges with respect to procedural rules for the agenda and the assignment of written opinions. The network influence thus captures a dimension of importance that differs from the correlation with the majority.

More generally, there are many such ways of considering how voters might "influence" each other. Though it remains unclear how causal influence, if it can be measured, would be mapped to perturbations in the pairwise maxent model, such ways of experimenting on the model present hypothetical opportunities for changing outcomes. We further discuss this concept of perturbing outcomes as a measure of influence using the information geometry of minimal models in Chapter 1.3.

1.1.4 Considering ideology

At the beginning of this section, we assumed that there was a symmetry between the orientation of the votes -1 and 1 such that the model only captured majority-minority dynamics. Yet, the fact that we were able to capture accurately the entire probability distribution of votes may have been surprising because it is possible that the patterns of agreement and disagreement reflect the political biases of each of the actors. In other words, the correlations between voters might be a result of sharing similar biases when presented with a case. Here, we consider this possibility by including the ideological orientation of the votes.

Measuring ideological orientation, however, is not a straightforward task. For one, the typical measure of ideology is unidimensional, mapping voters to a liberal-conservative axis, whereas cases can be complex and contain multiple issues that are not straightforward to map to a scalar. Second, ideology is, to an extent, defined by the relative differences between voters. For example, standard political science voting models learn ideological tendencies from the voting record by relying on the distribution of disagreements [52], which assumes

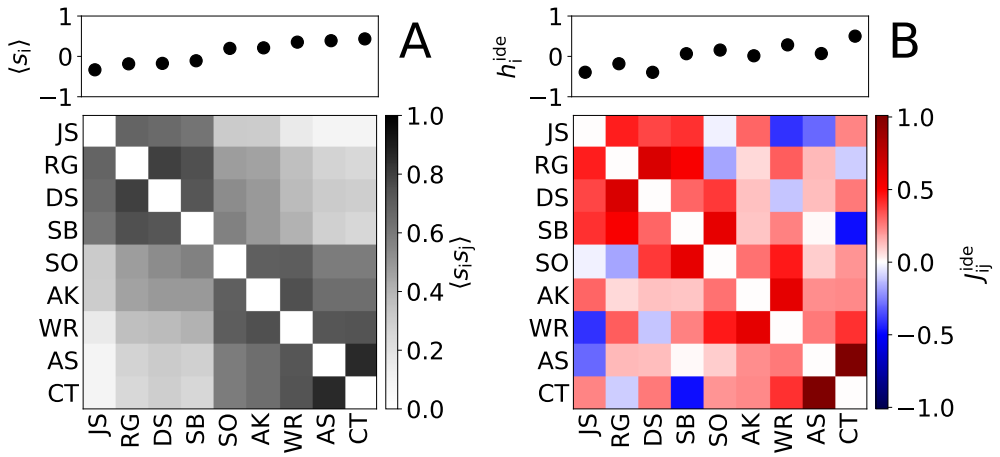


Figure 1.7: (A) Average vote $\langle s_i \rangle$ and pairwise correlations $\langle s_i s_j \rangle$ for Second Rehnquist Court with ideologically labeled data. The value -1 corresponds to liberal. (B) Solved fields h_i and couplings J_{ij}^{ide}. Compare with Figure 1.2.

the same majority-minority symmetry we considered earlier. Third, interpretation of both qualitative and quantitative differences can be subjective, and there is debate amongst Supreme Court scholars about how to best label the ideological leanings of votes in cases [68]. Finally, some cases do not have a natural interpretation as left or right because they fall outside the realm of partisan policy lines such as interstate disagreements. Given such qualifications, we do not attempt to identify ideology ourselves, but rely on the Supreme Court Database, which contains a binary labeling of the votes as either liberal $s_i = -1$ or conservative $s_i = 1$. Now accounting for ideology, the voters have clear biases: the libertarian J.S. with bias $\langle s_{JS} \rangle = -0.33$ and the conservative ideologue with bias $\langle s_{CT} \rangle = 0.43$. We now include these ideological biases into the pairwise maxent model.

As a result of considering bias in the "magnetizations" of each voter, the fields h_i^{ide} in the energy function from Eq 1.7 are no longer fixed at 0. As we show in Figure 1.7, negative fields indicate a tendency to vote liberally and positive

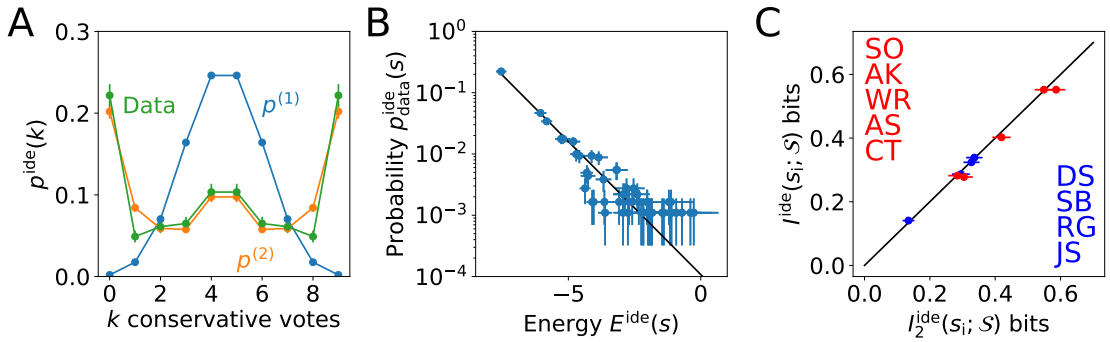


Figure 1.8: Testing the maximum entropy model for the Rehnquist court with ideologically labeled votes. (A) Probability of k conservative votes $p^{\text{ide}}(k)$. We compare the data (green) with the predictions of the pairwise maximum entropy model $p^{(2)}$ (orange), and with independent ideologues $p^{(1)}$ (blue). (B) Probability of voting patterns s from the data versus the "energy" in Eq 1.5; line is Eq 1.7. Errors in probability arise from counting, the standard errors; errors in the energy are propagated from errors in estimating the parameters J^{ide}_{ij}. Only states that appear more than once are shown, setting a floor for $p^{\text{ide}}(s)$. (C) Mutual information $I^{\text{ide}}(s_i; \mathcal{S})$ between individual votes s_i and the decision $\mathcal{S}$ of the majority, compared with $I_2(s_i; \mathcal{S})$ from the model. Conservatives are red and liberals blue, from highest $I(s_i; \mathcal{S})$ to lowest according to data.

fields indicate a tendency to vote conservatively. Though the average votes grow steadily from the most negative value to the most positive along the liberal-conservative ordering of justices, the fields h^{ide}_i do not in Figure 1.7A. This suggests, if ideology can be measured by average vote, that the tendency to vote either in the liberal or conservative blocs is mediated not just through individual biases but also through interactions with others. This hint that the interactions are important is evident in the failure of the independent model in Figure 1.8A, where we show the distribution of k votes in the conservative majority $p^{\text{ide}}(k)$. As was the case with the majority-minority model, there is an overwhelming tendency to form unanimous blocs at the expense of 5–4 divisions in contrast with the distribution predicted by the model of independent ideologues.

It could have been that by accounting for individual bias, the couplings would have been very different. Instead, we find that the ideological couplings J_{ij}^{ide} are very similar to J_{ij} with a correlation coefficient of $\rho = 0.90$. We show in Figure 1.7B the ideological couplings. The similarity between the couplings suggests that the couplings are responsible for generating the pattern of strong collective blocs. Once these tendencies to form political blocs is accounted for, it takes little for the system to become ideologically polarized. In analogy to strongly interacting systems in statistical physics, the energy required to break a symmetry encoded in the Hamiltonian is small.

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 sources of intuition 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.

1.2 Partisan intuition belies strong, institutional consensus and wide Zipf's law for voting blocs in US Supreme Court

In the previous section, we considered voting on a single natural court at a time. While the membership of the Supreme Court may change every few years, the institution has existed for centuries. Here, we model the Supreme Court over time, integrating over many natural courts during the era of the modern court to explore how collective patterns manifest over the ensemble of justices over time. As a result, we identify characteristics of the institution rather than of any single natural court.

The modern US Supreme Court is often described as being divided between liberal and conservative wings balanced on the fulcrum of one or two swing voters [23, 34, 45]. In the 1930s, a similar dynamic was at work where a conservative bloc known as the "Four Horsemen" relied on a fifth swing vote to countermand policy designed to ameliorate the economic impact of the Great Depression [72]. Over time, political issues and the composition of the court have changed, but the idea that a left and right divide characterizes voting on the court is a widespread observation in the literature [32, 36, 45, 62, 67]. The implicit hypothesis is that ideologically similar justices on different courts would have voted similarly on the same cases, but in reality they never faced the same cases nor voted with the same set of colleagues. Although we could take similar cases and then compare two justices' votes [62, 69], a more direct approach would be to compare how the two vote with a shared cohort to infer how the two might have voted together.

Long overlapping stints between pairs of justices are informative about how they vote relative to each other. If two pairs out of three voters are highly correlated, then transitivity implies that the third pair would likewise vote together often. If, however, the two pairs are anticorrelated, then the proverbial logic "the enemy of my enemy is my friend" implies that the third pair would be again correlated. In general, there are an infinite number of models that would match the correlations between observed pairs while filling in the missing ones. If we insist that the model match the observed pairwise averages of agreement but otherwise make no further assumptions, we have specified pairwise maximum entropy (maxent) models. This minimal set of assumptions means that maxent models are not guaranteed to capture the statistics of political voting with few parameters. We show, however, that they do while remaining consistent with highly parameterized spatial voting models from political science.[8]

We take all 36 justices who formed part of the modern Supreme Court from 1946–2016 ($K = 8,737$ recorded votes) to build a pairwise maxent model of a "Super Court," capturing the probability distribution over the joint distribution of all justices. Although one might anticipate that, in the much larger group of voters, strong divisive forces tear apart the Super Court into small factions, we find that the Super Court is dominated by unanimity, showing the Supreme Court to be an extremely stable institution. Besides the strong consensus captured from only pairwise statistics, the model generates a rich structure of dissenting blocs that goes far beyond a low-dimensional, ideological picture. This comparison of Supreme Court justices across time is just one example of the comparative study of political institutions. We show how minimal models from

[8]For the approach we discuss here, we integrate over the entire history of each justice including some that show clear changes in voting patterns over their tenure.

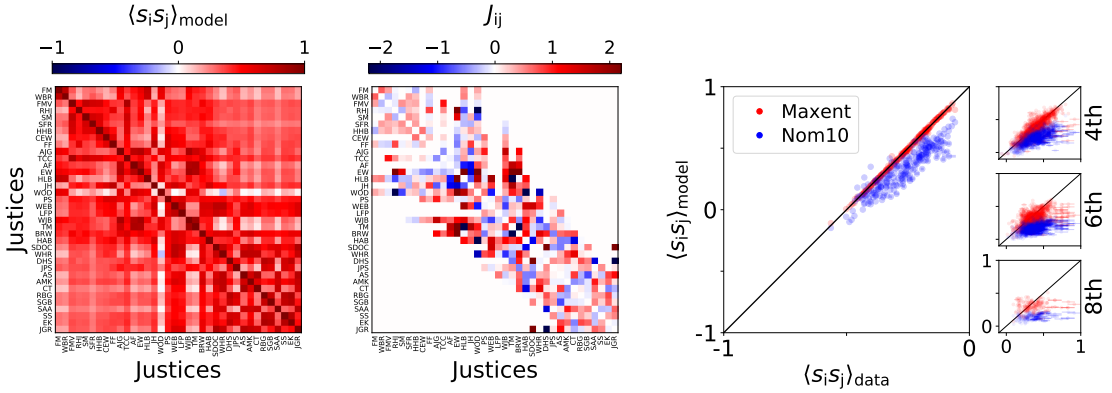


Figure 1.9: (left) Predicted pairwise correlations from pairwise maxent model are almost all positive. (middle) Inferred couplings J_{ij} as defined in Eq 1.7. Couplings between pairs of justices that did not vote together are excluded from the model and are effectively 0. (right) The pairwise maxent model fits the pairwise, 4th, 6th, and 8th order correlations better than the 10-dimensional W-Nominate model even after the spatial model is adjusted for unanimous votes. For the maxent model, over 97% of the 197 fit correlations are within two standard errors of the mean given K_{ij} observations for justices i and j, $\delta_{ij} = \sqrt{\langle s_i s_j\rangle + 1)(1 - \langle s_i s_j\rangle}/4K_{ij}$, except for a few major deviations accounted for by conflicting constraints (Appendix B.1). Strong deviations in 4th, 6th, and 8th order correlations can be explained by short temporal fluctuations in the distribution of case types (Appendix B.2). For clarity, only points with absolute error $\delta_{ij\ldots k} = \left|2\sqrt{p(s_i = s_j = \ldots = s_k)[1 - p(s_i = s_j = \ldots = s_k)]}/K_{ij\ldots k}\right| > 0.25$ have error bars.

statistical physics make quantitative predictions that could be rigorously tested against alternative models or new data.

1.2.1 Pairwise maxent approach

As in the previous section with the Second Rehnquist Court, we represent the vote of Justice i to be s_i where the binary majority vote is represented by $s_i \in \{-1, 1\}$. For nearly all votes, the sign of the vote corresponds to whether to affirm or to reverse the previous court's ruling, but the orientation is determined by the history of the case. Since we do not expect the pattern of internal disagreement to depend on whether the question is formulated as "Is

A constitutional?" or as "Is A unconstitutional?", we remove any such bias by symmetrizing the data set such that [26, 36]

$$p(s) = p(-s) \tag{1.13}$$

Thus, every justice votes Yea as frequently as votes Nay, but the patterns of internal pairwise disagreement are left unchanged. Along with observed pairwise correlations, Eq 1.13 and the principle of maximum entropy specify the values of the remaining unobserved pairwise correlations.

For each $\langle s_i s_j \rangle$ we fix, we have a corresponding coupling J_{ij} that, in the strictest sense, only has probabilistic implications. A positive J_{ij} lowers the energy E when the two justices i and j vote together, increasing the probability that we observe such a vote by a factor $e^{J_{ij}}$. A negative J_{ij} would decrease the probability of observing i and j voting together. Since justices who never voted together have no correlations to constrain, the corresponding couplings are not specified, effectively setting $J_{ij} = 0$. Although one might interpret this to indicate the absence of real interaction, the precise statement is that the predicted correlations are mediated through the structure implied by the constrained correlations. The couplings do not imply causal interaction between the justices. If there were causal interactions, then the consequences would be hidden in the pairwise correlations. In other words, the only information in the model is that available in the pairwise correlations and no more as is explicitly imposed by the maxent approach.

Finding the couplings that match the observed pairwise correlations is a hard computational problem, but a number of efficient numerical techniques and open source packages have been developed recently [35, 48]. The problem is

more complicated here because we have no votes of the full Super Court. With a finite set of data and votes from different subsets of the Super Court, it is possible to measure a set of marginals that are incompatible with a joint probability distribution. Instead of exactly fitting the marginals, we minimize the distance to the observed pairwise correlations and are able to find a good fit using the Monte Carlo Histogram technique as we show in Figure 1.9 [13]. To check that the model predicts other features of the data, we compare it with fourth, sixth, and eight order correlations and find that it does well. The few measured correlations that have strong deviations correspond to subsets that voted together for short periods of time and are not well-characterized by the long-time averaged statistics that are well fit by the model (Appendix B.2).

In the maxent model for the Super Court, the pairwise correlations are almost exclusively positive as shown in Figure 1.9, confirming that the Supreme Court has always been dominated by consensus. Out of the 197 observed pairwise correlations, only 2 are negative—between W.O. Douglas and W.E. Burger and between W.O. Douglas and W.H. Rehnquist—hinting that W.O. Douglas is an unusual justice in the history of the court [70]. After solving the model, we can predict the hypothetical correlations between justices who never served together, and even in this enlarged set of possibilities the only further negative correlations are between W.O. Douglas and conservatives S.D. O'Connor, W.H. Rehnquist, A Scalia, and C. Thomas. Additionally, the ideologically opposed pairs W.J. Brennan and C. Thomas, T. Marshall and W.H. Rehnquist, and T. Marshall and C. Thomas are negatively correlated. While they are all small negative values, the correlations are nevertheless prominent because they are unusual in a political body that prizes public agreement [20, 72, 76]. Although the re-

maining correlations are positive, the left-right divide is indicated by stronger correlations within and weaker correlations between ideological sides when the justices are ordered by ideology (Figure B1). This tension between competing blocs is reflected in the interaction matrix J_{ij}, where couplings of both signs are recovered from the model of majority voting. Thus, both consensus and ideological tendencies manifest in the pattern of pairwise correlations and are reflected in the structure of the model.

1.2.2 W-Nominate model

We compare the maxent model with a standard spatial voting model from political science called W-Nominate (Appendix B.3) [52]. Canonical spatial voting models assume that each justice lives in some low-dimensional, policy space along with the Yea and Nay votes for every case considered; a justice is more likely to vote for the position to which he or she is closer.

In this model, each justice votes independently such that the probability of observing any given vote can be factorized into the product of the probabilities of each justice, correspondingly the sum of the energies, $\log p(s) = \sum_i \log p(s_i) \propto -\sum_i \beta E(s_i)$ and correlations are induced by similar positions in policy space. In physical terms, the parameter β would be called an inverse temperature and controls how randomly justices are voting such that when $\beta \to 0$ features of the cases do not matter and justices vote randomly.

Given justice i's preferred policy position θ_{id} corresponding to vote s_i with distance from case k, $f_{kd}(\theta_i, s_i)$, along dimension d of a D-dimensional policy space,

every justice contributes to the total energy

$$E_i(s_i) = -\exp\left(-\sum_{d=1}^{D} w_d^2 f_{kd}(\theta_{id}, s_i)^2/2\right) \tag{1.14}$$

The parameters w_d weight each dimension, hence the "w" in the name. Eq 1.14 stipulates that when cases are located close to a justice's preferred policy position, justices are more likely to vote for it than against it.

In the language of statistical learning, spatial voting is a kernel technique [9]. The kernel describes how the positions of points in the space relate to one another. Eq 1.14 describes a radial basis kernel for a Gaussian process for Justice i where the position of the justice and cases are parameters and the weights w_d and inverse temperature β are hyperparameters [55]. The rotation, translation, and scaling symmetries of this kernel reflect the fact that only relative positions of the justices matter for the spatial voting model. All parameters are found simultaneously by maximizing the Bayesian posterior probability across all justices and all cases. This is implemented in an R-package described in Ref [53].

Although unanimous votes are a particularly notable feature in the pairwise maxent model, unanimous votes in the data are excluded from the training set for the W-Nominate model. They are degenerate in the sense that every voter has the same apparent policy preference. In this sense, they give no information about the relative positions of the voters in policy space (and can negatively impact numerical accuracy), and so it is standard procedure to exclude them from the analysis [53, 54]. Still, unanimous votes can be generated by sampling from the W-Nominate model even if they occur rarely. To compare the models directly, we explicitly include the probability of a unanimous vote as given by

the pairwise maxent model, p_{unan}. We reweight the W-Nominate model such that a unanimous vote s_{u}

$$p(s_{\text{u}}) \leftarrow p(s_{\text{u}})(1 - p_{\text{unan}}) + p_{\text{unan}}. \tag{1.15}$$

While for other votes

$$p(s) \leftarrow p(s)(1 - p_{\text{unan}}) \tag{1.16}$$

to conserve probability. Eqs 1.15 and 1.16 do not fix the two models to have the same $p(s_{\text{u}})$—even if comparable—but they do ensure that similarities beyond the unanimous mode are readily apparent.

The D = 10 spatial voting model referred to as Nom10 has $> 10^5$ parameters yet is minimal compared to other variations of spatial voting models. In comparison, the pairwise maxent model has $< 10^3$ parameters. Despite this large disparity in parameterization, the maxent model fits the observed statistics of the data better as shown in Figure 1.9.

1.2.3 Strong consensus spans a century

As times change, the justices on the court are replaced one by one, eventually yielding an entirely new cohort. By tradition and precedent, the new members are bound to the court's past [72]. On the other hand, it is not unusual for the Supreme Court to overturn previous rulings or to pass rulings that align with the political preferences of the court.[9] How does the tension between keeping and breaking with the past manifest in the voting record? If each new set of justices were radically different, we would expect correlations to decay very

[9] *Bush v. Gore* is cited as an example of how political preferences are no more unusual in the court than in the other branches of the US government [61].

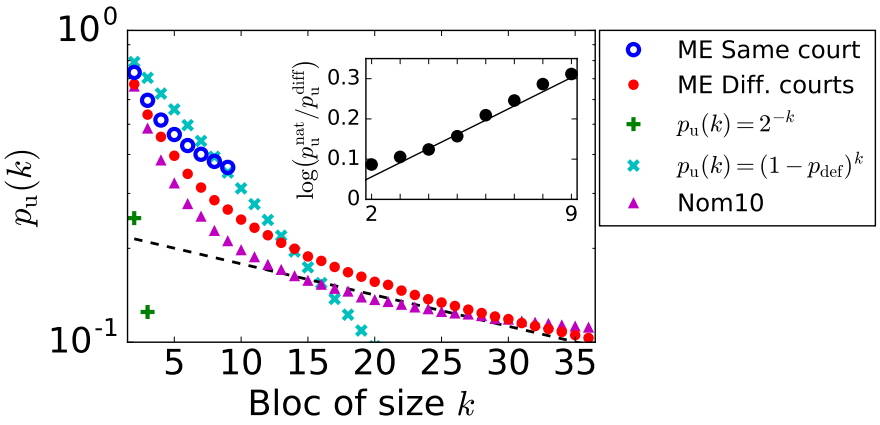


Figure 1.10: Decay length of consensus spans a century for both pairwise maximum entropy (ME) and the 10-dimensional W-Nominate (Nom10) spatial voting models. The maxent model probability that a bloc of size k vote together $p_u(k)$ decays exponentially when $k \gtrsim 20$ with decay length of $l = 43$ justices (dashed black line) corresponding to a century of justices. We show justices that sat on the same court (blue), justices who did not (red), independent voters (green), independent defectors from unanimity (cyan), and the 10-dimensional W-Nominate model (magenta). (inset) "Norm to consensus" force increases the probability that justices in a natural court all vote together p_u^{nat} relative to justices who did not vote together p_u^{diff}. The ratio of probabilities grows by a constant factor with each additional justice. We measure this as an energy difference that grows linearly for justices in the same natural court compared to justices did not work together. The rate of increase, 0.035 per justice, is faster than the inverse decorrelation length in the limit of large courts $l^{-1} = -0.023$ per justice showing that consensus dominates over random disagreement.

quickly, evidence of independence from historical precedent. If new justices were maintaining strong coherence with the past, the time scale for correlations would be long. In order to ask this question, we consider how the pairwise correlations change as we consider increasingly larger subsets of the Super Court spanning increasingly distant spans of time.

Despite consisting of members from 24 nine-member natural courts, the probability of unanimity on the Super Court is an astounding 10% of the time. In comparison, the maxent model constrained to fit only the individual voting av-

erages $\langle s_i \rangle$ describes independent justices who vote unanimously with a minuscule probability of $p_u(k = 36) = 2^{-36}$. The wide gulf between the independent model and pairwise models shown in Figure 1.10 implies that the interactions impose strong collective structure. As another test of the strength of cohesion, we might consider that justices already have an idea of how the court is leaning before they cast their vote. If we instead imagine that each justice independently defects from unanimity with probability p_{def}, we can measure for a natural court of $k = 9$ that typically votes unanimously about 36% of the time, $p_{def} = 0.89$. Here, p_u accounts for factors like ideology that induce consensus for a typical nine-member court. Extrapolating this defection model to 36 individuals, we find that the Super Court would vote unanimously with probability $p_u(k = 36) = 0.015$.[10] This is an order of magnitude below what we find on the Super Court and shows that independent defection is suppressed by consensus. Thus, the Super Court shows a strong tendency to consensus that spans the many natural courts it represents.

Even in a chain of tightly correlated voters, we might expect that the probability that justices agree with others far away in time becomes dominated by drift [28]. Looking at Figure 1.9, we see that pairwise correlations far from the diagonal tend to be smaller than those closer, indicative of weaker correlations spanning longer periods of time. To test this more systematically, we consider the probability of unanimity in blocs of size k, $p_u(k)$ and compare justices that sat together with random subsets in Figure 1.10. In the limit of large blocs of 20 or more justices, or roughly two disjoint Supreme Courts, the unanimous mode

[10] We use the probability of unanimity of a group of size $k' = 9$ to estimate the probability that a each justice would defect if there were independently doing so, $p_{def} = 1 - p_u^{1/k'}(k')$. The probability of no defections with k independently defecting justices is then $p_u(k) = (1 - p_{def})^k$, shown as teal x's in Figure 1.10.

begins to decay like an exponential, indicating the regime of finite correlation length. Here, adding a new justice is like adding an independent voice to the system. The decay length l, however, is 43 justices, roughly equivalent to a century of justices. When explicitly accounting for unanimity in the W-Nominate model, it likewise predicts a similar decay length of 77 justices with a similar decay profile as shown in Figure 1.10. In this quantitative sense, the Supreme Court is an extremely stable and conservative institution.

The force quelling dissent, a "norm of consensus" [20], is even stronger for justices that sit on the same court. We again calculate the probability of unanimity for blocs of size k, but only for justices who sat on the same natural court. As k increases from 2 to 9, the log-ratio of the probabilities, or energy difference, of finding a unanimous vote on a natural court $p_{\mathrm{u}}^{\mathrm{nat}}$ compared to justices from different courts $p_{\mathrm{u}}^{\mathrm{diff}}$ increases linearly with the bloc size $\log\left(p_{\mathrm{u}}^{\mathrm{nat}}(k)/p_{\mathrm{u}}^{\mathrm{diff}}(k)\right) = E^{\mathrm{diff}}(k) - E^{\mathrm{nat}}(k) \propto 0.035k$ as we show in Figure 1.10. This linear relation reveals a force encouraging the formation of blocs and especially for consensus specific to natural courts. Notably, the energy term for consensus grows faster than the inverse decorrelation length in the limit of large courts $l^{-1} = -0.023$ per justice. This suggests that consensus would dominate even when random justices who had never voted together are made to vote together which is akin to the situation when new justices are nominated into the Supreme Court.

These results together provide evidence for collective effects that impose strong consensus on the court—in fact strong enough that unanimity occurs ten times more frequently than the most frequent vote with dissenters (Figure 1.11). We show that collective effects are necessary to describe not only how justices band

together to form a united decision but also in how they dissent, indicated by the slower than exponential decay in Figure 1.10 in comparison to null models we consider. Furthermore, the long correlation length for unanimity even at the regime of uncorrelated justices shows that consensus effects decay very slowly over time. Finally, comparison of the decay length of the exponential tail with the measured force to consensus shows that the norm of consensus dominates over entropic effects. In this sense, the Super Court shows strong consensus and this institutional consensus is a signature of Supreme Court voting.

1.2.4 Heavy-tailed Zipf's for voting configurations

Besides consensus, the distribution of Super Court votes that emerges from pairwise interactions shows much richer structure including and going far beyond binary ideological division. After ordering the votes from the pairwise maxent model by frequency rank r, the probability $p(r)$ clearly follows a power law over four orders of magnitude in Figure 1.11. This form is described by Zipf's law [47, 60]

$$p(r) = (1 - \alpha) r^{-\alpha} / (r_{\text{min}} - r_{\text{max}}) \tag{1.17}$$

where $\alpha = 0.85$ when excluding the strong consensus mode that appears prominently for $r = 1$ and $r = 2$. The small value of $\alpha < 1$ indicates a distribution overwhelmingly dominated by the tail. In the limit of a large number of unique dissenting coalitions as shown here, votes are almost exclusively found in the tail rather than in any fixed number of votes with high rank. This finding is consistent with data on the Second Rehnquist Court, but over a limited range of $\sim 10^2$ unique votes. There, a power law fit yields $\alpha \approx 1$ [36]. Such heavy-tailed distributions show that the intuitive notion that we could compress the voting

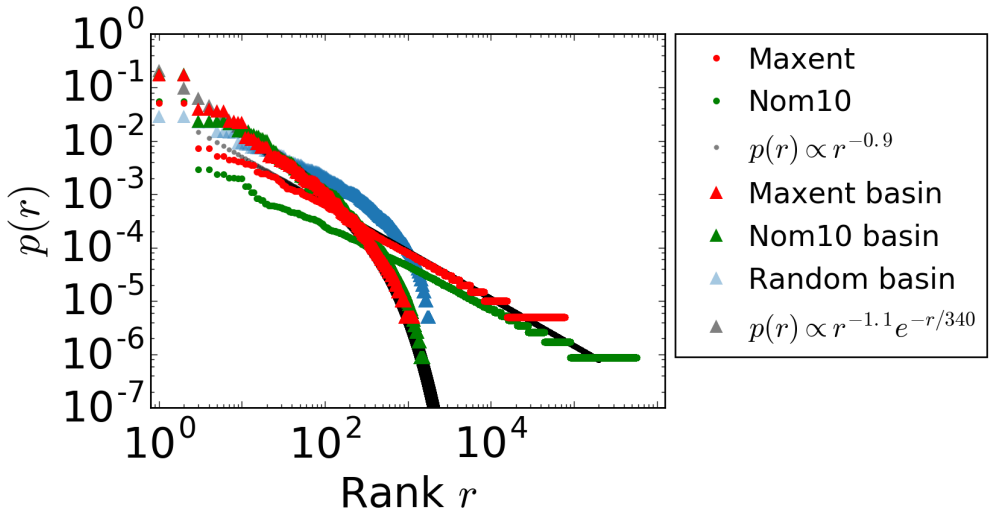


Figure 1.11: (circles) Zipf's plot for the pairwise maxent model (red circles) and the W-Nominate model (green circles) as described by Eq 1.17 (black circles). Symmetry about the sign of the vote means that there are always two votes with the same probabilities. The maxent distribution is very heavy-tailed with exponent $\alpha = 0.85$. Note that the unanimous vote (the first pair of points) occurs about 10 times more frequently than the most frequent dissenting coalition. (triangles) Maxent distribution over energy basins (red triangles) follows a truncated power law described by Eq 1.19 (black triangles) with large cutoff $n = 340$. Nom10 (green triangles) does not have a natural energy landscape relating votes, but the distribution over basins found assuming the maxent energy equation in Eq 1.7 shows remarkable similarity. In comparison, random, independent voters show strong deviations from the maxent distribution over basins when assuming the maxent landscape (blue triangles).

of the Super Court into a few ideological modes is misleading because even votes that defy this intuition are probable.

The scale-free nature of the distribution of votes $p(r)$ furthermore indicates the lack of a cutoff for a simple description of Super Court voting. Like previous approaches for constructing a sparse description of Supreme Court voting based on spectral techniques [34, 67], we consider the eigenvalue spectrum of the covariance matrix $\langle s_i s_j \rangle$ estimated from the pairwise maxent model. If it were possible to compress the covariance matrix into a few principle components with

high fidelity, then there would be a natural scale at which including further components would do little to improve the accuracy of the model. Only a few components would matter, for example, if a binary ideological spectrum were sufficient to describe Supreme Court voting across time such that the cutoff would always include a fixed number of dimensions: a dimension for consensus, left vs. right, liberal and conservative cores, etc. [36]. However, we find that the eigenvalues of the principle components $\lambda(r)$ ordered by rank r and averaged over random subsets of size N are well-described by the scale-free form

$$\lambda(r) = r^{b/a}\Lambda(rN^{-a}). \tag{1.18}$$

with $a = 0.8$, $b = -1.1$, and universal scaling function Λ shown in the scaling collapse in Figure 1.12. Since the spectrum is close to a power law with cutoff that grows with N, there is no fixed length scale for the spectrum. In contrast, previous work has shown that a small number of principle voting modes can indeed summarize sitting nine-member natural courts [34, 67]. we show that by considering how voting behavior scales with the number of justices in the larger set that the addition of every justice reveals yet another level of complexity. In this sense, Supreme Court voting cannot be reduced to a few principle dissent coalitions, but is inherently complex.

By grouping votes together such that similar votes belong to dominant clusters of "noisy" prototypical votes, however, we might find that the clusters serve as a compressed description of the Super Court [27, 36, 59]. A natural way of clustering votes is given by the energy landscape from Eq 1.7. By taking a vote and flipping the single voter most likely to change his or her vote at a time, we will eventually reach an energy minimum, where no vote flip will lower the energy of the vote. This defines an energy landscape where states can be

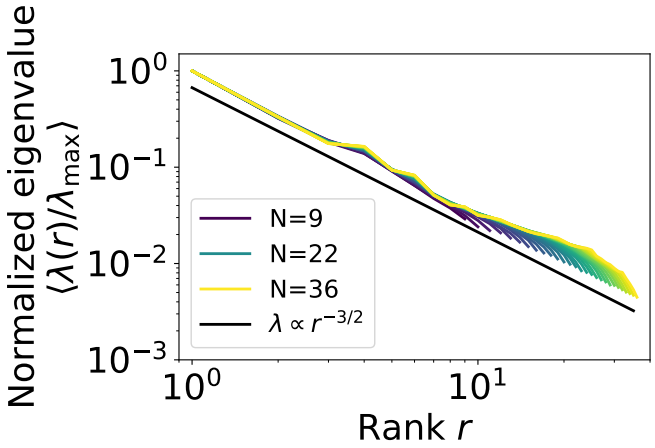


Figure 1.12: Rank-ordered spectrum of eigenvalues of the pairwise maxent covariance matrix $\langle s_i s_j \rangle$ when averaged over random subsets of size N of justices from the Super Court. Eigenvalues are normalized by the largest eigenvalue $\lambda_{\max}$ before averaging. Spectrum decays like a power law with $\lambda(r) \sim r^{b/a}$ where $b/a \approx -3/2$ as proposed in Eq 2.12 and with exponents estimated by scaling collapse shown in inset. The cutoff of the distribution scales with the number of justices considered N, revealing dependence on the system size and not a fixed number of dominant dimensions. (inset) Universal scaling function $\Lambda(rN^{-a})$ obtained from a scaling collapse.

grouped by local energy minima. This procedure is equivalent to searching for local maxima in probability where single voter flips climb hills in a probability landscape. When we measure the probability of being located on a given hill, we find again Zipf's law but now with an exponential cutoff

$$p_{\text{basin}}(r) \propto r^{-\alpha'} e^{-r/n} \tag{1.19}$$

with $\alpha' = 1.1$ and $n = 340$. The exponential cutoff means that clusters of noisy prototypical votes could serve as dominant modes, but the slow decay of the power law and the size of the cutoff means the important modes are many.

The large number of energy minima is in stark contrast with the Second Rehnquist Court where there are only six meaningful clusters despite being a quarter of the size of the Super Court [36]. This comparison, however, is misleading. For

spin glasses, systems with many frustrated interactions as with the Super Court, the number of maxima grows exponentially with system size [12]. Like with the Second Rehnquist Court, we find that many of the largest energy minima correspond to votes that fracture along ideological lines. The "noisy" perturbations away from the energy minima describe how a few individuals tend to buck the line. Thus, ideological tendencies manifest in the structure of the energy landscape, but they are not the full story. Dominant modes of voting are indeed useful and accurate descriptions of voting on natural courts because the systems are small [34, 67], yet a low-dimensional representation of collective voting over time misses the heavy tail generating a rich range of behavior.

1.2.5 Relating votes to legal issues

At the beginning of this section, we started with two possible ways of comparing voters across time, either by comparing similar cases or by using overlapping voting records, and we decided to use the latter to construct the joint probability distribution of the Super Court. Yet, given that justices prefer to vote in certain ways for different types of cases, we would expect that the way that justices align contain information about the type of cases they are considering. We might expect, for example, that especially partisan issues are reflected in partisan divides (e.g., *Bush v. Gore*), whereas cases dealing with uncontested application of precedent show unanimous outcomes.

The statement that the pattern of internal divisions contains information about the type of case being considered is the assertion that the joint probability distribution $p(s, \tau) \neq p(s)p(\tau)$ of the vote s and type of case τ. Then, the distribution over votes $p(s)$ represents a marginalization over all the types of cases τ consid-

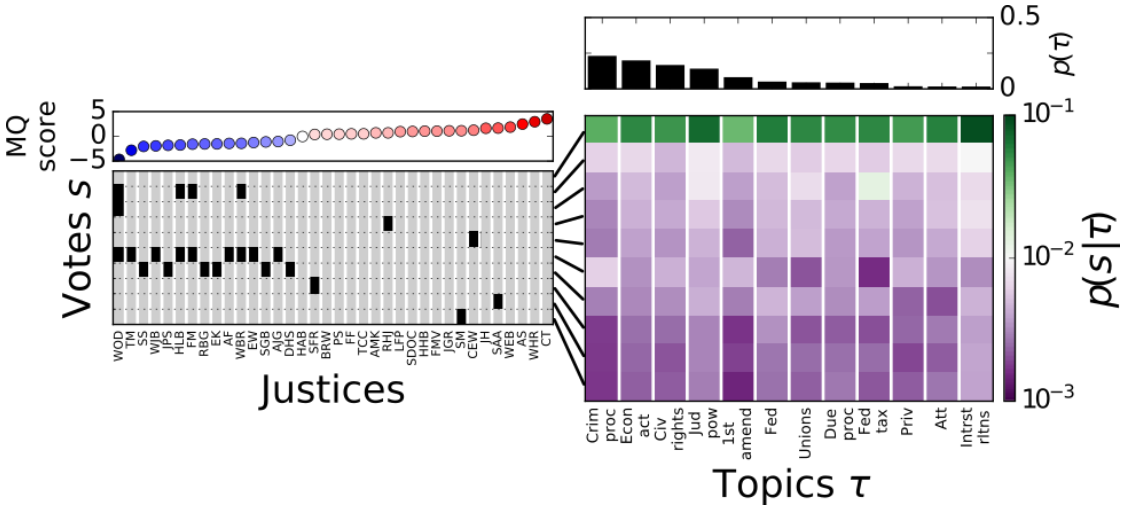


Figure 1.13: Joint distribution of legal topics and Super Court voting. (right) Each column is the conditional probability of seeing a vote given the topic $p(s|\tau)$. The topics occur with frequencies $p(\tau)$ as shown on top. Topic labels are abbreviations for Criminal Procedure, Economic Activity, Civil Rights, Judicial Power, First Amendment, Federalism, Unions, Due Process, Federal Taxation, Privacy, Attorneys, and Interstate Relations. (left) Each column is a different vote: justices have been ordered left-to-right by ideological scores (term-average Martin-Quinn scores), where blue corresponds to liberal and red conservative.

ered by the justices across the overlapping history of these 36 justices from the Super Court,

$$p(s) = \sum_{\tau} p(s, \tau). \tag{1.20}$$

We never witness a full vote of the entire Super Court, but if we were to take a particular type of case where we know the votes of a subset of the Super Court s_k of k voters, we could used the joint correlations in the pairwise maxent model to guess how the remaining justices would have voted on a case on the same topic, $p(s|\tau, s_k)$. In other words, the distribution of voting patterns of the Super Court when conditioned on case type τ encodes information about the case. By Bayes' Theorem, the converse is true: the case encodes information about the voting patterns of the Super Court.

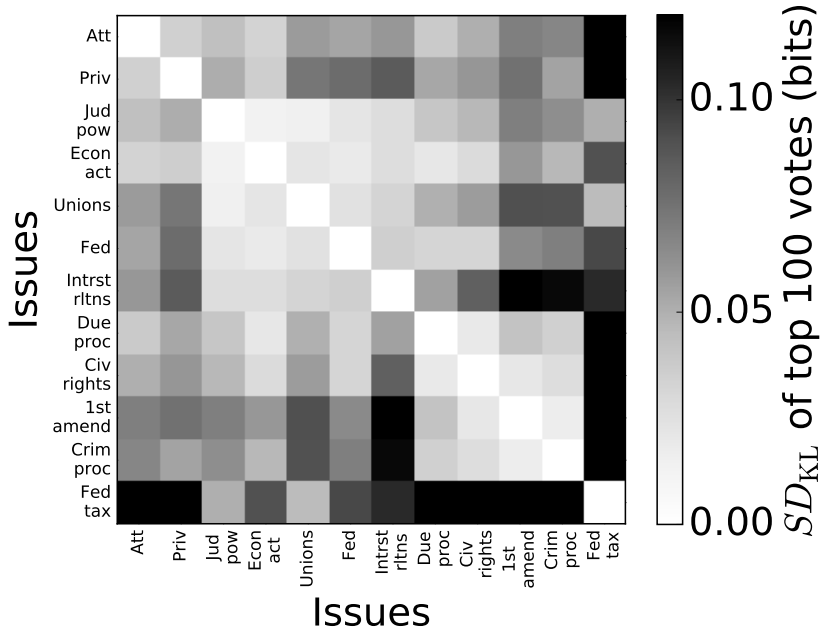


Figure 1.14: Symmetrized Kullback-Leibler divergence between legal topics on Super Court voting as in Eq 1.21. See caption of Figure 1.13 for full topic names.

The Supreme Court Database labels cases by their primary legal topic and provides 14 such topics [68].[11] We focus on the top 12 topics that have at least 50 data points with distribution $p(\tau)$. We compare the probabilities of the ten most frequent votes for each legal topic in Figure 1.13. Even with just these ten votes, it is clear that the way that the Super Court divides is strongly variable, suggesting that the voting distributions serve as fingerprints for legal topic. Such a picture aligns with the intuition that we could use divisions between justices on the different topics to distinguish cases from one another.

To measure the differences between the distributions of voting patterns conditioned on two different legal topics $p(s|\tau)$ and $p(s|\tau')$, we measure the sym-

[11] An alternative approach would be to learn a useful labeling of the cases like with natural language processing techniques or by the origins of the cases.

metrized Kullback-Leibler divergence,

$$SD_{\mathrm{KL}} = \frac{1}{2} \sum_{s} p(s|\tau) \log_2 \frac{p(s|\tau)}{p(s|\tau')} + p(s|\tau') \log_2 \frac{p(s|\tau')}{p(s|\tau)}, \tag{1.21}$$

limiting ourselves to the top 100 states that appear across all conditional distributions because of numerical sampling limits. We show the pairwise distances in Figure 1.14, having ordered the legal topics such that each topic is more similar to neighboring topics than ones further away. It is apparent that some topics yield very similar distributions of voting outcomes, whereas others like "Federal taxation" are substantially different from almost every other topic.

Taking this matrix of distances, we hierarchically cluster the legal topics. Starting with the most similar pair of legal topics, we join the two into a single cluster, continuing to join either pairs of individual topics or adding new topics to an existing cluster if the average pairwise distance to the members of the cluster is smaller. By keeping track of the value of the distances at which topics are clustered together, we build a tree from the roots as we show in Figure 1.15, a method called agglomerative clustering based on average linkage [50].

We find that the quantitative similarities between the topics seem to align with the conceptual similarity suggested by the names. Dividing topics into clusters by $SD_{\mathrm{KL}} = 0.04$ around which a large gap appears across several clusters, we find that topics dealing with legal practice—Attorneys and Privacy—go together. Economic questions that also bring in questions of federalism—Judicial power, Economic activity, Unions, Federalism, and Interstate relations—belong together. Topics dealing with civil liberties go together: Due process, Civil rights, First amendment, Criminal procedure. These intuitive groupings disagree with suggestions made by the legal scholars such as to group Federal

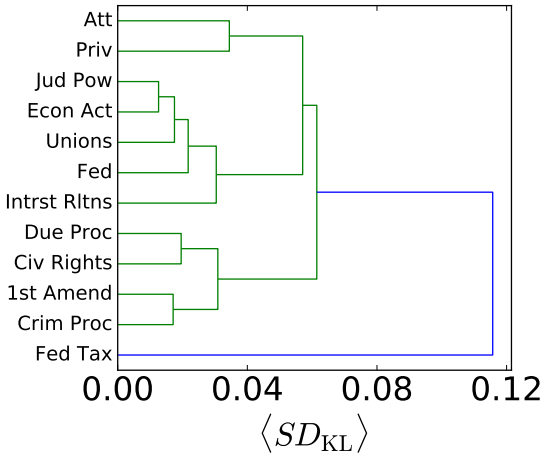


Figure 1.15: Agglomerative clustering by symmetrized Kullback-Leibler divergence (Eq 1.21) between Super Court vote distributions conditioned on case type. See caption of Figure 1.13 for full topic names.

Taxation with Economic Activity, to group Attorneys with Economic Activity, and to group Privacy with Civil Rights [68]. According to our model, Federal Taxation clearly belongs in its own bracket, Attorneys is quite distinct from Economic Activity, and Privacy cases are about as different from Civil Rights as they are from the cluster including Economic Activity. This disagreement suggests that a more quantitative approach to grouping these cases informed from the pattern disagreement between voting justices and historical overlap in voting records might better highlight the similarities between the topics.

1.2.6 Discussion

Although voting in a political context is complex, such collective social phenomena are natural to approach from the perspective of statistical physics [8, 16, 24, 59]. Voting on the Supreme Court shows strong collective effects—often obscured by the focus on individual ideology [19, 44, 61]—that we exploit by using observations of justices that voted together to infer how justices who never did might have (Figure 1.9). We find that the hypothetical Super Court is domi-

nated by strong consensus that nearly spans a century (Figure 1.10) even in the modern era where dissent is much more likely [73]. Hiding beneath this dominant mode is a rich range of dissenting coalitions that emerge from the pattern of pairwise correlations (Figure 1.11), illustrating the hypothetical behavior of historically disjoint groups of justices. The wide distribution of dissenting coalitions is scale-free, and the absence of a cutoff for a sparse representation of Super Court voting is reflected in the scaling behavior of the eigenvalue spectrum of the covariance matrix (Figure 1.12).

The remarkable similarity in the distribution of votes beyond the unanimous mode of the maxent and W-Nominate models suggests that they are largely capturing the same collective structure encoded in the maxent energy landscape. From the perspective of model selection, however, the principle of Occam's Razor favors the pairwise maxent model for predicting voting across Supreme Courts. The focus of voting models in political science are typically for extracting case, justice, and year specific parameters [44], and so the goal is to fit many presumably informative parameters. Yet, we show here that the number of parameters dwarfs the number of available data points and a much simpler model captures similar structure. This shows that the spatial voting model is overparameterized for inferring the distribution of voting across Courts. More generally, accounting for model complexity is useful for good prediction because simpler models tend to generalize better [43].

Consideration of such hypothetical situations has long played a significant role in legal research. Legal scholars study the influence of past precedents by using written opinions as sources of legal reasoning for future justices [72]. Political

scientists compare justices by considering similarities and differences between the issues brought up in cases [5]. These approaches suggest alternative ways of using or testing the predictions of this model. A direct quantitative test and application of the maxent model would be for forecasting Supreme Court votes [31, 57]. For example, the strong historical correlations that we find here suggest how one might leverage historical votes on similar cases across time to enhance predictions. Knowing how correlations decay over time are essential for weighting historical information as a function of temporal separation.

The slow decay of consensus in time highlights the role of the institution in the Supreme Court. Given that the typical length of tenure is on the order of a few decades (although a few cases span the range of a year to nearly 40 years), the fact that consensus spans nearly a century implies the existence of an institutional timescale that surpasses any single justice's role in the Supreme Court. Such a phenomenon perhaps arises from the use of precedent or other institutional norms [72]. Indeed, we find evidence supporting the presence of such an example, a "norm for consensus," when comparing rates of unanimous voting between random subsets of justices and subsets of natural courts (Figure 1.10) suggesting that strong consensus is mediated by real interaction between justices [19].

More generally, the emergence of collective features explains why directly extrapolating from natural courts can fail to characterize the set of historically disjoint justices. We find, for example, that when we move from the Second Rehnquist Court to the Super Court, the number of metastable states in the energy landscape increases from only 6 to the order of 340, respectively. This immense

growth in complexity only becomes apparent when considering how voting behavior scales with system size. Yet previous, quantitative studies of Supreme Court voting focus almost exclusively on natural courts where the number of justices is fixed [23, 24, 34, 36, 67]. When justices are compared across natural courts, dependence on N is not considered and much analysis hews closely to the ideological narrative based on the study of natural courts [32, 45, 62]. It is natural in statistical physics to look for signatures of collective behavior as a function of system size, and we show that this approach provides quantitative evidence for inherently high-dimensional voting on the Supreme Court based on only binary voting data.

In contrast to partisan ideological intuition obtained from news on the Supreme Court, the collective behavior of the court over time reveals an institution focused on consensus and insulated from the seemingly rapid pace of political change. When we look beyond consensus, ideology captures only a part of the story if we account for the changing composition of the court: ideological modes are lost amongst the cacophony of dissenting voices. As we know from statistical physics, even simple models can generate an incredibly rich range of behavior when there are many competing, frustrated interactions [49, 66]. we show that this intuition is expressed in the complex voting patterns of the Supreme Court across time.

1.3 Sensitivity of collective outcomes identifies pivotal components

In previous sections, we developed a statistical physics of political voting, relying on the principle of maximum entropy to construct quantitative models grounded in data to characterize voting on the US Supreme Court. Here, we will go beyond this picture in two ways.

We will consider the influence of individuals on collective outcomes. We will think of influence as the response of the aggregate, e.g., majority outcome, to fluctuations in individual voting behavior. In statistical physics, this is analogous to susceptibility coupling fluctuations along one degree of freedom with those along another. The susceptibility, by the fluctuation-dissipation theorem, is related to the linear response to a perturbation in a system at equilibrium [56]. Leaning on the exact mapping between the maxent model and a system at thermodynamic equilibrium—though we note that such an assumption was not necessary for the maxent derivation—we might think of such a measurement of influence as an experiment probing the behavior of a voter and watching the perturbation cascade throughout the network of couplings connecting individuals.

As our second extension, we will expand our analysis to a wider range of examples of collective behavior that can likewise be mapped to binary states: voting in legislatures and other courts, the rise and fall of stock indices, and the agreement or disagreement in choice of vocabulary by Twitter users. This set of examples demonstrates the generality of our approach building on the maxent

framework and yet represents a small sample of a great variety of systems that remain to be studied with our systematic approach.

1.3.1 Introduction

When collective outcomes are highly sensitive to the behavior of few individual components, these components are pivotal. Collective outcomes could be the partition of voters into blocs, the pattern of co-moving financial indices, or the coalescence of shared vocabulary in a social community. A classic example is the swing, or median, voter, prominent in political science and economics: if voters can be deterministically ranked according to preference, the median will always vote in the majority and thus is predictive of the outcome [3, 10, 18]. In real systems, this simple picture becomes much more complicated because the median might change depending on the contested issue [33], multiple issues may be at stake simultaneously [17], voters might exchange votes strategically [19, 61], etc. In other words, competing interactions between voters imply that changes in individual voting behavior may cascade into alignment or antagonistic changes in others resulting from direct physical interactions or indirect ones, i.e., mediated through a new compromise on the contents of a legislative bill. In contrast with an idealized notion of a median, we consider a "pivotal" voter one that could change collective outcomes even when accounting for such complexity. Here, we develop this generalized notion and use it to identify components that are especially indicative of collective changes in political voting, financial indices, and social media on Twitter.

Information geometry provides a natural framework for measuring how sensitive collective properties are to change in component behavior. The Fisher in-

formation, the fundamental quantity of information geometry [1], establishes a metric over probability distributions $p(s;\{J_{ij}\})$ of states s (e.g., a voting configuration, stock and sector price movement, behavior on social networks) determined by the set of parameters $\{J_{ij}\}$ indexed by ij. When the parameters are infinitesimally changed to $\{\tilde{J}_{ij}\}$, the distribution becomes $p(s;\{\tilde{J}_{ij}\})$, and the distance between the two distributions is given by the Fisher information (FI) [15]. By measuring the FI for perturbations to each pair of variables J_{ij} and $J_{i'j'}$ in turn, we construct the Fisher information matrix (FIM) $F_{iji'j'}$, whose eigenvectors describe how changes to the parameters lead either to sharp change in the model (large eigenvalues) or slow change (small eigenvalues) [42, 71]. Often, collective outcomes are a coarse-graining of the high-dimensional state s to a lower-dimensional variable $f(s)$. An example is when the full vector of individual votes is compressed into a binary variable indicating the direction of the majority outcome. By coarse-graining, we map the probability distribution to one over the coarse-grained state $p(s) \rightarrow q[f(s)]$. By calculating the FIM for q instead of p, we consider the sensitivity of collective outcomes. When the FIM indicates that aggregate properties are highly sensitive to a few components, we determine that they are key contributors to the collective properties of the system.

We outline our approach in Figure 1.16, where we fit a minimal statistical model to a data set, measure the FIM, and extract properties of the local information geometry. We first discuss this approach on a toy Median Voter Model to build intuition, and we extend detailed analysis to voting on an example from the US Supreme Court (SCOTUS) and State Street Global Advisors SPDR exchange-traded funds [11, 68]. Then, we perform a survey across multiple systems in

society including examples of judicial voting across US state high courts [40], California (CA) state legislatures [38], US federal legislatures [39], and communities on Twitter [25]. Across these examples, we find large diversity: ranging from examples of median-like systems, with pivotal voters or components, to other examples in which no special component emerges.

1.3.2 Median Voter Model (MVM)

The role of the median derives from the fact that in a majority-rule voting system, the voting outcome depends on a coarse-graining instead of the detailed nature of every individual's vote. The margin by which the majority wins, as is captured in the probability $q(k)$ that k voters of the system are in the majority, can reflect the appeal of the voting outcome or even its legitimacy. These perceptions feed back into the decision process [72]. Thus, $q(k)$ serves as an aggregate measure of underlying decision dynamics that we will use to identify pivotal blocs.

To outline our approach, we study the sensitivity of $q(k)$ in the context of a reduced toy model that captures the essence of a median voter. The ideal median voter exists in a majority-rule system where voters' preferences are unidimensional. By virtue of a unique ranking of preference, the median is always in the majority [10]. We propose a statistical generalization, the Median Voter Model (MVM), with an odd number of N voters. The MVM consists of $N-1$ Random voters and one Median voter who always joins the majority. The binary vote of voter i, s_i, is equally likely to be -1 and 1 such that only majority-minority divisions are relevant. Thus, the average votes are all the same, but the set of pairwise correlations as shown in Figure 1.16A display nonzero correlations be-

tween M and R, $\langle s_M s_R \rangle = 0.3125$, and no correlations between R's, $\left\langle s_{R_i} s_{R_{j\neq i}} \right\rangle = 0$. Thus, this model consists of a special voter, the Median (M), who after a voting sample has been taken, is perfectly correlated with the majority, whereas Random (R) voters all behave in a statistically uniform and random way.

To capture the network of interactions between individuals from which majority-minority coalitions emerge, we take a pairwise maximum entropy (maxent) approach [63]. The maxent principle describes a way of building minimal models based on data. We maximize the information entropy $S = -\sum_s p(s) \ln p(s)$ while fixing the model to match the pairwise correlations from the data, $\left\langle s_i s_j \right\rangle_{data} = \left\langle s_i s_j \right\rangle$ as defined in Figure 1.16A. The result is a minimal model parameterized by *statistical* interactions between voters, or "couplings" J_{ij} in Figure 1.16B [29]. For each pair of voters with pairwise correlations in Figure 1.16A, there is a corresponding coupling such that the set of couplings is specified exactly by the pairwise correlation matrix. For the MVM, the $5^2 - 5 = 20$ couplings only take two possible values, one for each of the two unique correlations. The couplings for the MVM indicate that all R's tend to vote with M (agreement between M and R leads to an increase in the log-probability $\ln r(s_M = s_R) \propto J_{MR}$ as in Figure 1.16B) with a slight tendency for R's to disagree with each other more than would be expected given their shared correlation with M (disagreement between R_i and R_j decreases the log-probability of the vote by $\ln r(s_{R_i} = s_{R_j}) \propto J_{RR}$). In principle, any probabilistic graph model is a viable alternative for the approach we outline, but the pairwise maxent model has been shown to capture voting statistics better than other models of voting with surprisingly few parameters [35, 36], fits the data well (Appendix C.2), and presents a particularly tractable model for calculating

information quantities.

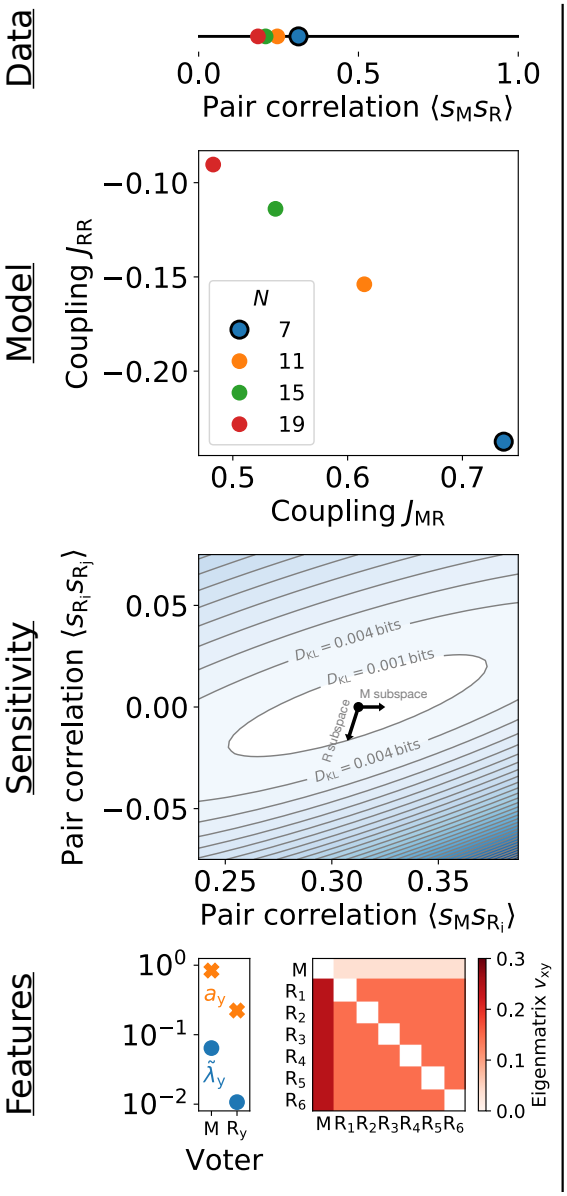
To probe how the collective properties captured by the distribution $q(k)$ depend on the voters, we ask how the distribution changes if the voters were slightly different. In this example of majority-minority voting, any change in voting behavior is reflected in the pairwise correlations and preserves the symmetry between the two possible outcomes -1 and 1. A natural endpoint for the set of possible $q(k)$ as we increase the pairwise correlations is when all voters are perfectly correlated, so we consider perturbations that take us towards this endpoint: with some small probability, voter y's votes are replaced by x's,

$$\tilde{r}(s_{\mathrm{y}} = s_{\mathrm{x}}) = (1 - \epsilon) r(s_{\mathrm{y}} = s_{\mathrm{x}}) + \epsilon. \tag{1.22}$$

Eq 1.22 is a weighted average that interpolates from the current probability of agreement between x and y, $r(s_{\mathrm{y}} = s_{\mathrm{x}})$, when $\epsilon = 0$ to perfect agreement when $\epsilon = 1$. We then account for the changes to y's correlations with the remaining voters:

$$\tilde{r}(s_{\mathrm{y}} = s_{\mathrm{x}' \neq \mathrm{x}}) = (1 - \epsilon) r(s_{\mathrm{y}} = s_{\mathrm{x}'}) + \epsilon r(s_{\mathrm{x}} = s_{\mathrm{x}'}). \tag{1.23}$$

Eq 1.23 interpolates from the current probability of agreement between y and x′ when $\epsilon = 0$ to that between x and x′ when $\epsilon = 1$. If replacing M with any R voter such that y = M and x = R, the operation defined in Eq 1.22 increases the pairwise correlation $\langle s_{\mathrm{M}} s_{\mathrm{R}} \rangle$ while simultaneously changing M's correlations with the others to be more like those with R, pushing them to zero. When the statistical model exactly matches the entire distribution of votes $p(s) = p_{\mathrm{data}}(s)$, the perturbation described in Eqs 1.22 and 1.23 is equivalent to shifting the probability from any voting configuration where i and j disagree to the voting configuration where i and j agree, holding all others probabilities constant. With the pairwise



A

$$\langle s_i s_j \rangle = \sum_s p(s) s_i s_j$$

(Pairwise correlation)

B

$$p(s;\{J_{ij}\}) = e^{-E(s)} \Big/ Z$$

$$E(s) = -\frac{1}{2}\sum_{i,j=1}^{N} J_{ij} s_i s_j$$

$$Z = \sum_s e^{-E(s)}$$

(Pairwise maxent model)

C

$$F_{x'y'xy} = \lim_{\epsilon\to 0} \frac{2}{\epsilon^2} D_{KL}\left[q_J \| \tilde{q}_{\tilde{J}}\right]$$

(Fisher information matrix)

D

$$\tilde{\lambda}_y = \lambda_y \Big/ |\lambda|$$

(Component eigenvalue)

$$a_y = \frac{1}{4}\sum_{x=1}^{N} \left(v_{xy} - v_{yx}\right)^2$$

(Component asymmetry)

Figure 1.16: Overview of method for identifying pivotal voters for the Median Voter Model. (A) Taking the pairwise correlations ($\langle s_{R_i} s_{R_j} \rangle = 0$ is not shown), (B) we solve a pairwise maxent model to learn the probability distribution $p(s; \{J_{ij}\})$ parameterized by the couplings J_{ij}. MVMs of different sizes N correspond to different coordinates in this two-dimensional space, but we focus on $N = 7$ as an example. (C) We calculate the FIM for $q(k)$, the probability of k votes in the majority, measuring the sensitivity of $q(k)$ to changes in voter behavior as described by Eqs 1.22 and 1.23. As we describe in Appendix C.3, the sensitivity corresponds to the curvature of the Kullback-Leibler divergence D_{KL}. We show a two-dimensional cut of D_{KL} when M copies R_i and R_i copies R_j. The principal directions in the full space, v_{xy}, determine the combinations of perturbations to which $q(k)$ is most sensitive. We show projections of eigenvectors obtained from limiting perturbations to the Median or an Random voter as black arrows. (D) The principal eigenvector of the FIM, reshaped into an "eigenmatrix" on the right, specifies the relative change in the rate that x's votes are replaced by y's (i.e., a positive value is the rate at which x's voting record becomes y's and a negative the rate at which its disagreements increase). The vector of principal subspace eigenvalues per voter $\tilde{\lambda}_i$, corresponding to the outlined diagonal blocks in Appendix Figure C7C, are divided by norm $|\lambda|$ to give our pivotal measure. The asymmetry a_y measures the difference in perturbations localized to a specific voter vs. all its neighbors in turn. If a voter and all its neighbors are similar, the asymmetry is close to zero. Otherwise, it is bounded by a maximum value of one (see Appendix C.6).

maxent model, however, the perturbation is only reflected in the pairwise correlations, moving us from one model to another within the class of pairwise maxent models. In this case, the perturbations can be mapped to changes in the couplings J_{ij} in the limit of $\epsilon \to 0$ that we use to determine the entries of the FIM shown in Appendix Figure C7C (Appendix C.4).

The variation in the entries of the FIM indicates the unique role of the median. The FIM describes the curvature of the Kullback-Leibler divergence D_{KL} as the probabilities of pairwise agreement are modified. Under small perturbations, the contours of D_{KL} form an ellipse, whose major and minor axes represent components of the FIM's eigenvectors, as in Figure 1.16C. We show the principal

eigenvector in Figure 1.16D. Its entries represent the relative amount by which pairs should be simultaneously varied for maximal local change to $q(k)$ — as if one could change all the pairwise voting "knobs" at once. To be clear about the pairwise grouping of index, we reshape the principal eigenvector into an "eigenmatrix" in Figure 1.16D. Each column corresponds to a directed change where voter y is made more similar to the corresponding row voter x. Since R's are all the same, the first column connecting M to each R is uniformly valued. In the first row, the entries all correspond to making the neighbors of M more like M, so these are also all uniformly valued given that the R's are interchangeable. Thus, each column of the eigenmatrix describes perturbations localized to the column voter and each row corresponds to changes across all the neighbors of a particular voter such that the symmetry between R's and the unusual role of M manifests in the comparison of local neighborhood with the local neighborhood of neighbors.

This local versus neighborhood asymmetry presents one way of pinpointing an unusual voter by using the difference between the eigenmatrix v_{xy} and its transpose v_{yx}. We define this per voter asymmetry a_y in Figure 1.16E. Given a normalized eigenmatrix, the total asymmetry over all voters $A \equiv \sum_y a_y$ is 0 when the eigenmatrix is perfectly symmetric and is 1 when perfectly antisymmetric $v_{xy} = -v_{yx}$. The point $A = 1/2$ marks the maximum asymmetry possible when all the nonzero elements are of the same sign — such as when for each $v_{xy} > 0$, $v_{yx} = 0$ (Appendix C.6). For the MVM with $N = 7$, we find that M's asymmetry $a_M = 0.06$, whereas $a_R = 0.01$, clearly distinguishing M from R. The total $A = 0.13$, a point of reference for systems that are more complex than the MVM. For larger N, the MVM asymmetry a_M grows as the role of M more visibly skews the

distribution. Thus, both the asymmetry in the roles of voters and the growing importance of a median with system size is reflected in the symmetry of the eigenmatrices.

To measure the sensitivity of $q(k)$ to each voter, we inspect the subspace eigenvalues λ_i, specifying the sensitivity of $q(k)$ to change in a single voter's behavior. These values are calculated from the subspace of the FIM describing localized perturbations — the diagonal blocks of the FIM as outlined in Appendix Figure C7C and whose eigenvectors are projected into Figure 1.16C. The upper leftmost block corresponds to M and the remaining blocks correspond to each R in turn. For each subspace, we retrieve the principal eigenvalue. To compare the eigenvalues across voters, we calculate the normalized eigenvalue as defined in Figure 1.16D, $\tilde{\lambda}_i$, defining our measure of how "pivotal" a component is relative to others. For the $N = 7$ MVM, the principal eigenvalues are $\tilde{\lambda}_M = 0.70$ and $\tilde{\lambda}_R = 0.05$. This large difference indicates that $q(k)$ is over 10x more sensitive to variation in M than R, again reaffirming the special role of the median. It is important to note that voters with strong asymmetry are not necessarily the most pivotal — clearly because eigenvalues and eigenvectors present different information. Still, asymmetry in the eigenmatrix indicates heterogeneity amongst the voters; thus, large asymmetry is necessary, if insufficient, for the pivotal measure to vary across a wide range. Overall, the information geometry of this minimal class of models provides a way of quantifying the role of individual components on collective outcomes, identifying key components with pivotal roles that can emerge given strong heterogeneity in the population.

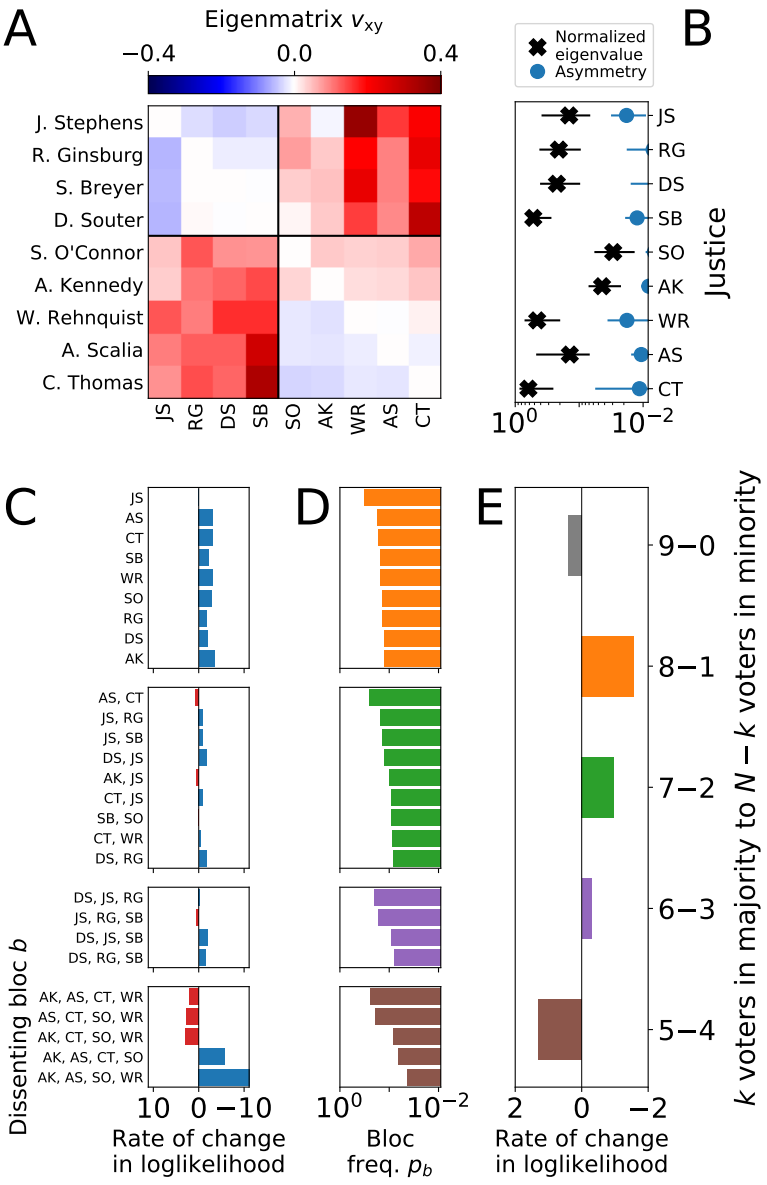


Figure 1.17: SCOTUS example. (A) Principal eigenmatrix of the FIM. Justices are ordered from most liberal to most conservative according to a standard measure of ideology [44]. We indicate the typical divisions between the liberal and conservative blocs with black lines. (B) Normalized voter-subspace eigenvalues and asymmetry per justice (Figure 1.16D). (C) Rate of change in log-probability of dissent for dissenting blocs $\ln p_b$. (D) Each bloc b's probability of dissenting together according to the pairwise maxent model, p_b. (E) Rate of change in log-probability of k dissenters $\ln q(k)$. Error bars represent 95% confidence intervals from repeating the full procedure outlined in Figure 1.16 for 10^2 bootstrapped samples of the data.

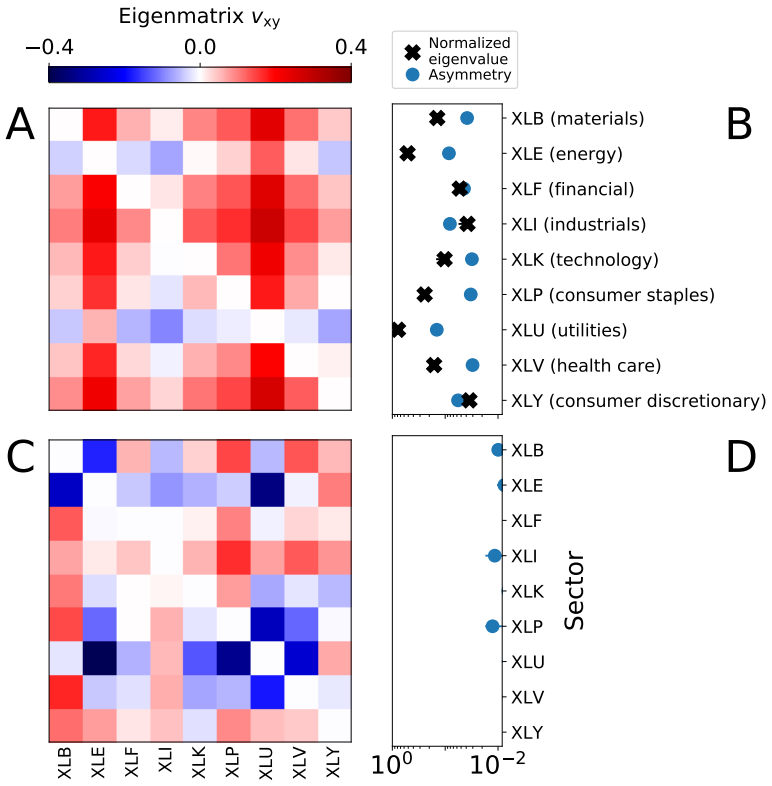


Figure 1.18: S&P SPDR example. (A) Principle and (C) secondary eigenmatrices of the FIM with eigenvalues $\Lambda_1 = 0.41$ and $\Lambda_2 = 0.13$, respectively. (B, D) Relative sector subspace eigenvalues and asymmetry by sector (Figure 1.16D). Error bars represent 95% bootstrapped confidence intervals.

1.3.3 US Supreme Court (SCOTUS) and S&P 500

We perform the same analysis on an example from SCOTUS of $N = 9$ voters, $K = 909$ votes, and between the years 1994–2005 (see Appendix E.1 for details about data sets). We show the principal eigenmatrix in Figure 1.17 that consists of perturbations primarily increasing similarity across ideological wings given by the positive values connecting liberals and conservatives.[12] The principal mode has a total asymmetry of $A = 0.10$ compared to $A = 0.25$ for the $N = 9$ MVM, indicating the absence of a median-like, pivotal voter. This absence is surprising

[12]Since the recovered eigenvector is arbitrary with respect to sign, we could just as well consider the negative eigenvector that would reverse the sign but preserve the magnitude of the elements.

because discussion of medians A. Kennedy and S. O'Connor is prominent in the context of this court. When we consider voter-subspace eigenvalues shown in Figure 1.17, we find the justices in ranked order: C. Thomas, S. Breyer, and Chief Justice W. Rehnquist. A change in C.T., given his strongly conservative voting record, would naturally constitute consequential change, but the roles of W.R. and S.B. are more subtle [34, 36, 67]. Despite A.K. and S.O.'s prominent role in the narrative of Supreme Court voting, we find that other justices come to the foreground when we consider the sensitivity of the Court to behavioral change.

The principal mode can be projected into the more intuitive space of dissenting coalitions in terms of the rate of change of the probabilities for dissenting blocs (Appendix C.5). Though the eigenmatrix in Figure 1.17 shows increasing similarity between ideological wings, suggesting suppression of partisan 5–4 divides, the frequency of any 5–4 divide actually increases strongly along with a decrease in lone and pair dissents as in the bottom of Figure 1.17. Seven of the nine most common pair dissents found in the data decrease in likelihood. Thus, this shift reflects an increasing tendency for justices to join larger blocs, reflected in the suppression of every Justices' lone dissents, in a way that breaks the typical partisan divide. To visualize changes in the existing 5–4 conservative-liberal dynamic, we inspect defections from the liberal bloc, or 6–3 votes where a single liberal vote is missing, and likewise defections from the five-member conservative bloc. On the whole, defections from the liberal bloc are less surprising than those for the conservative bloc, consistent with the balance of power favoring conservatives. For the liberal bloc, the most prominent change entails R.G. defecting, leaving D.S., J.S., S.B., which reflects the central role of R.G. in the liberal

coalition. On the other side, increasing the probability of S.O. or A.K. defecting is important though not as much as the defection of W.R., which reflects his often-understated, unusual statistical role in the Court [36]. Consistent with pundits' understanding is the large surprise associated with C.T.'s defection from the conservative majority, a change that would represent a fundamental shift in the established partisan dynamics. Overall, this individual variation in the context of the partisan 5–4 dynamic reveals a portrait of much deeper subtlety than that suggested by unidimensional partisan intuition [22, 33, 36]. Thus, the information geometry of statistical models of social systems can provide detailed insight into specific components or blocs in direct connection to their role in collective modes of the system.

In Figure 1.18, we analyze the founding set of State Street Global Advisors SPDR exchange-traded funds ($N = 9$; $K = 4,779$; 2000–2018), which replicate the indices and provide daily price data (binarized to positive $s_i = 1$ or negative daily changes including no change $s_i = -1$ in analogy to votes). In contrast with SCOTUS, the collective behavior of each index reflects the aggregation of many individual investors: no stock index is monolithic in the sense of an individual voter. Given this aggregate nature, it is natural to consider the eigenvectors as the most surprising set of unanticipated global changes — although entire sectors might be "perturbed" by government policy like sector-specific regulation or tariffs. From this point of view, fluctuations in the pivotal blocs might reveal notable shifts in economic conditions or collective perceptions thereof (Appendix C.7). Taking a look at the model, we find that the principal mode displays large asymmetry across every index, reflecting the diversity of roles played by the various sectors of the economy as captured in price movements.

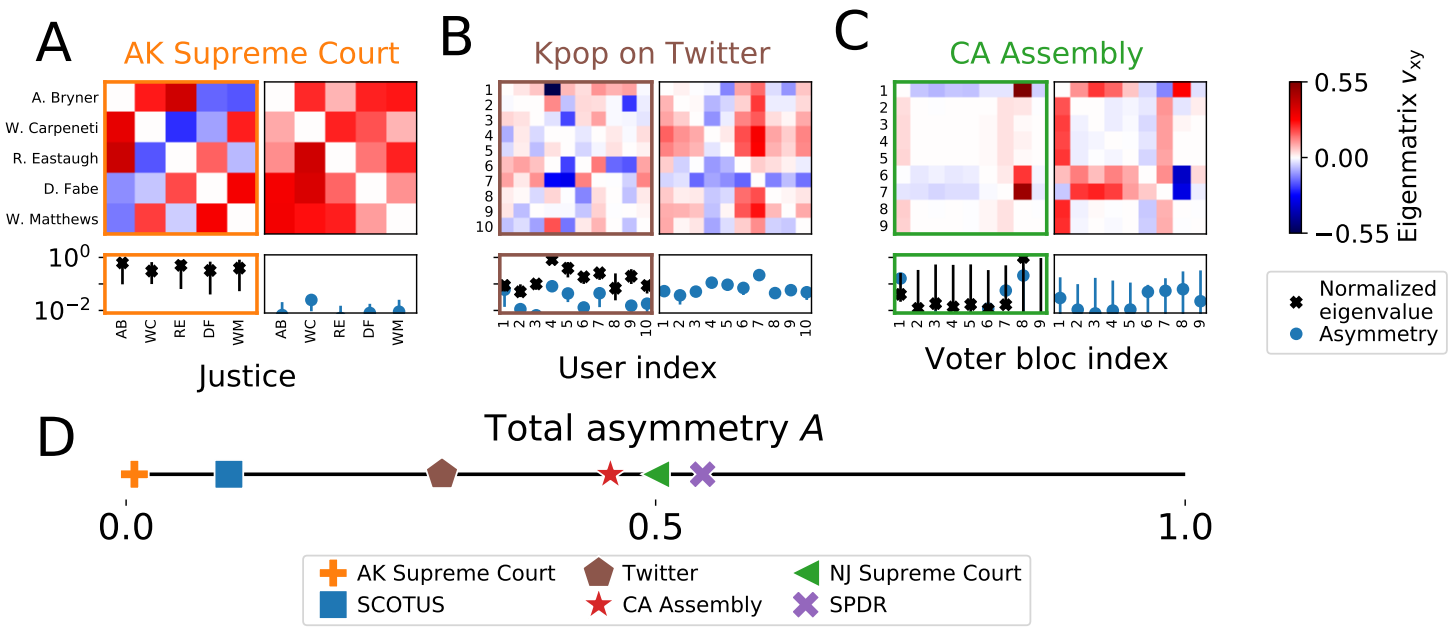


Figure 1.19: Example systems. (A, B, C) On top, we show the first two eigenmatrices for AK Supreme Court, K-pop on Twitter, and CA Assembly. Below each eigenmatrix, we show the pivotal measure and the asymmetry per component (see Figure 1.16). For the CA Assembly, voter are grouped into nine blocs after rank ordering by the first W-Nominate dimension (see Appendix Figure C21). Large error bars in the pivotal measure indicate that the unique pivotal role of Bloc 8 depends on a few crucial votes in this session. (D) Asymmetry of the principal eigenmatrix of the FIM. Error bars represent 95% bootstrapped confidence intervals.

Relatively large subspace eigenvalues highlight XLE (energy) and XLU (utilities), in agreement with their role as drivers of the economy on whose outputs many of the other sectors depend [2, 30]. Perhaps unsurprisingly, we also find a "bellweather" XLP (consumer staples) and XLV (healthcare) as notably pivotal whereas XLF (financial) and XLI (industrials) seem to be relatively not. Going beyond the principal mode, we inspect the secondary mode and find that it is remarkably symmetric, with an asymmetry score of $A = 0.07$, in contrast with the second mode of the SCOTUS example where $A = 0.44$. This secondary symmetry is reminiscent of the MVM where a prominent asymmetric mode hides a nearly symmetric mode arising from the uniformity of Random voters. Such a symmetry is not found for the SCOTUS example, where at lower modes, asymmetry actually increases, signaling notable individual roles in determining collective outcomes. Thus, these examples are a comparison of opposites, where the apparent asymmetry in components obscures shared structure for SPDR, whereas for SCOTUS the overarching tendency to consensus overshadows individual roles on the Court.

1.3.4 Pivotal components in society

We explore other examples of social systems, including votes from US state high courts [40], the California State Assembly and Senate [38], the US federal legislature [39], and communities on Twitter [25]. As with the previous two examples, we map behavior in these systems to binary form. To reduce the larger legislative bodies to a comparable number of blocs, we first separate voters into 9 blocs nearly-equally-sized blocs by ranked similarity according to a standard political science measure of ideology, the first W-Nominate dimension [53]. The bloc vote is given by the majority vote of the members and is randomly chosen if

equally divided (see Appendix Figure C21). For Twitter communities, we identify individuals as high-dimensional binary vectors where an element is positive if they used a corresponding keyword, or else negative, such that the pairwise correlations reflects overlap in their use of keywords. Thus, our analysis of the information geometry involves the same procedure outlined above, but for a wider variety of social systems.

Considering the principal eigenmatrix of the Alaskan (AK) Supreme Court ($N = 5$; $K = 1,021$; 1998–2007),[13] we find a remarkable degree of symmetry between justices and a small value for the total asymmetry $A = 0.01$. Such symmetry implies that the justices on this court all dissent in a statistically uniform way as described by the set of their pairwise correlations. Though this could be trivially true if all pairwise couplings were the same, this is not the case, a fact that is mirrored in the spread of positive and negative values in the eigenmatrix in Figure 1.19. Checking the local interaction networks described by the set of couplings to every neighbor j for justice i (Appendix Figure C9), we find that the sets are all similar for every justice i. This symmetry is mirrored in the similarity of the individual subspace eigenvalues shown in Figure 1.19. Consistent with this symmetry extracted from the voting record, four out of the five justices served as Chief Justice during this period,[14] a regular rotation of roles imposed by the state constitution stipulating that the Chief Justice only serve for three consecutive years at a time. In contrast, we show that the New Jersey (NJ) Supreme Court ($N = 7$, $K = 185$, 2007–2010) has strong asymmetry of $A = 0.5$ (Appendix Figure C14). Appointments to the NJ Court follow a tradition of maintaining

[13] With one case Millette v. Millette published in 2008 though Justice Bryner officially retired in 2007.

[14] W. Matthews, D. Fabe, and A. Bryner rotated as Chief Justice during the period 1997–2009 and W. Carpeneti from 2009–2012 (following the period of analysis).

partisan balance, apparently codifying a median role into the institution, and we find two nearly equal pivotal voters. Despite the seeming alignment between each of these two examples and the institutional norms, AK Supreme Courts are not always less symmetric than their NJ counterparts. The asymmetry is highly variable for previous years, suggesting that codified institutional rules only partially determine the role of pivotal voters (Appendix E.1).

We also show the eigenmatrices of the 1999 session of the CA State Assembly ($N = 77$; $K = 5,424$; 1999–2000) and a K-pop Twitter community ($N = 10$; $K = 7,940$; 2009–2017). The CA Assembly is an example of strong asymmetry ($A = 0.46$). For the sessions starting between the years 1993–2017, we find that the Assembly displays stronger signatures of asymmetry (average total asymmetry $\langle A \rangle = 0.4 \pm 0.1$) compared to the Senate ($\langle A \rangle = 0.3 \pm 0.1$), showing how the rules of the institution might be reflected in the distribution of pivotal blocs. We then compare the distribution of the single largest pivotal measure $\tilde{\lambda}_{\max}$ with that of the similarly coarse-grained US House of Representatives and Senate. Though the CA distributions for $\tilde{\lambda}_{\max}$ are statistically indistinguishable from each other (Kolmogorov-Smirnov test statistic $k = 0.31$ and significance level $p = 0.5$) and the federal bodies' distributions are similar ($k = 0.31$, $p = 0.05$), the state vs. federal levels show larger differences ($k > 0.46$, $p < 0.03$). This separation between the behavior at state and federal legislatures reflects institutional differences that are captured in the sensitivity of majority-minority coalitions (see Appendix C.8).

As for the Twitter Kpop community, we find much heterogeneity amongst users with total asymmetry $A = 0.30$, exceeding that of the $N = 9$ MVM. In contrast

with the MVM, this community contains multiple pivotal members but wide variation in the strength of their subspace eigenvalues. Twitter communities may be on average sensitive to the behavior a few individuals regardless of identity [25], but this individual-level variation suggests that collective behavior may be much more sensitive to a select Twitter users even within smaller communities [4]. Going beyond the detailed few examples in Figure 1.19, we find large diversity within political institutions that highlights the important role of heterogeneity in social institutions, heterogeneity that is captured in the information geometry of minimal, maxent models.

1.3.5 Discussion

An important question in the study of social institutions is whether or not collective decisions are robust to perturbation targeting individual components. Robustness is reciprocal to sensitivity: when a system is highly sensitive to small changes to components, its collective properties are not robust. In neural networks with avalanches of firing activity [21, 51, 59], in bird flocks with propagating velocity fluctuations [6], or in macaque societies with conflict cascades [16], such sensitivity might have an adaptive functional role. In the context of human society, questions of robustness are relevant to the stability of voting coalitions or the susceptibility of a population to disease or disinformation. For example, we might be interested in comparing the impact of different judicial nominees on the dynamics of voting on a judicial bench or the impact of modified user behavior on the spread of disinformation in social networks. By relying on the formal framework of information geometry to investigate statistical signatures of sensitivity, we present a data-driven and general approach to characterizing robustness. As a result, our approach is not model-specific, only relying on the

calculation of how sensitive a model is to changes in observable individual behavior (Figure 1.16D).

In voting systems, median voters are conventionally considered to be power brokers who have outsize influence [10, 44, 61]. Building on this idea, we propose a reduced toy model to extract features of the Fisher information matrix that correspond to signatures of a median voter. We show how to identify and interpret signatures of strong sensitivity on individual components in multiple social contexts, generalizing the intuition behind the median to pivotal components on which aggregate properties, measured by majority-minority divisions, depend strongly. Intriguingly, we find hints that institutional differences may contribute to structuring individual roles in collective outcomes both in courts and legislatures. Though it is unsurprising that the particular rules of a voting body may structure bloc dynamics, pivotal components provide a principled way of comparing social systems with differing composition, from different eras, and across different institutions in a unified, quantitative framework.

We might think of pivotal components as "knobs" that could drive a system out of its current configuration described by the ensemble $p(s)$. If the subspace eigenvectors are knobs, the pivotal measure is inversely proportional to the spacing of the dials such that for large eigenvalues the smallest turn results in the strongest effect. Since each pivotal component only considers the effects of perturbations localized to a single component, these knobs are independent. If n components were accessible simultaneously, however, we would consider the joint space of multiple pivotal components, and the principal subspace eigenvalue must increase beyond (or stay at) the maximum eigenvalue over the set

of component subspaces: this reflects the fact that enhancing the breadth of control only increases the range of possible outcomes [41, 75]. By considering which knobs are accessible experimentally, our analysis could be extended to measuring signs of statistical control in real systems. For judicial voting, the realizable knobs that change judicial voting behavior may be the submission of *amicus curiae* briefs, choice of litigating cases, or lobbying.[15] Those trying to craft a legislative coalition might "perturb" aspects of proposed policy to affect its acceptability to potential supporters [46]. In controlled biological systems, localized perturbations to single components could include manipulation of single neurons or the upregulation of specific genes.[16] Our work presents the possibility of informing the direction of such external perturbations in the broader context of control.

The understanding of the interplay between components and social structures across social and biological examples remains nascent at mesoscopic and macroscopic scales. With this principled, quantitative approach for measuring pivotal components, we might, by comparing systems, better understand how institutional and envirnomental factors shape the emergence of social structure.

[15] We are careful to point out that the ensemble of votes for political systems already include such effects so it is important to distinguish between endogenous and exogenous factors.

[16] For example, manipulation of single neurons is possible by electrical stimulation, optogenetic techniques, or chemical stimulation, all ways of enacting the localized perturbations of neural "votes" [59]. Analogously, gene expression might be perturbed by switching genes on and off or by adding protein directly to simulate changed expression levels [58].

Chapter 1 references

[1] Shun-ichi Amari. *Information Geometry and Its Applications*. en. Vol. 194. Applied Mathematical Sciences. Japan: Springer, 2016. ISBN: 978-4-431-55977-1.

[2] Vipin Arora and Jozef Lieskovsky. *Electricity Use as an Indicator of U.S. Economic Activity*. en. Tech. rep. Washington, D.C.: U.S. Energy Information Administration, 2014, p. 17.

[3] K.J. Arrow. *Social Choice and Individual Values*. 3rd. Yale University Press, 2012. ISBN: 978-0-300-18698-7.

[4] Eytan Bakshy et al. "Everyone's an Influencer: Quantifying Influence on Twitter". en. In: *Proceedings of the Fourth ACM International Conference on Web Search and Data Mining - WSDM '11*. Hong Kong, China: ACM Press, 2011, p. 65. ISBN: 978-1-4503-0493-1. DOI: 10.1145/1935826.1935845.

[5] Lawrence Baum. "Comparing the Policy Positions of Supreme Court Justices from Different Periods". en. In: *Western Political Quaterly* (1988), pp. 509–521.

[6] W. Bialek et al. "Social Interactions Dominate Speed Control in Poising Natural Flocks near Criticality". en. In: *Proceedings of the National Academy of Sciences* 111.20 (May 2014), pp. 7212–7217. ISSN: 0027-8424, 1091-6490. DOI: 10.1073/pnas.1324045111.

[7] W. Bialek et al. "Statistical Mechanics for Natural Flocks of Birds". en. In: *Proceedings of the National Academy of Sciences* 109.13 (Mar. 2012), pp. 4786–4791. ISSN: 0027-8424, 1091-6490. DOI: 10.1073/pnas.1118633109.

[8] William S. Bialek. *Biophysics: Searching for Principles*. en. Princeton, NJ: Princeton University Press, 2012. ISBN: 978-0-691-13891-6.

[9] Christopher M Bishop. *Pattern Recognition and Machine Learning*. Singapore: Springer Verlag, Aug. 2006.

[10] Duncan Black. "On the Rationale of Group Decision-Making". en. In: *Journal of Political Economy* 56.1 (1948), pp. 23–34.

[11] Michael J. Bommarito and Ahmet Duran. "Spectral Analysis of Time-Dependent Market-Adjusted Return Correlation Matrix". en. In: *Physica A: Statistical Mechanics and its Applications* 503 (Aug. 2018), pp. 273–282. ISSN: 03784371. DOI: 10.1016/j.physa.2018.02.091.

[12] A J Bray and M A Moore. "Metastable States, Internal Field Distributions and Magnetic Excitations in Spin Glasses". en. In: *Journal of Physics C: Solid State Physics* 14.19 (July 1981), pp. 2629–2664. ISSN: 0022-3719. DOI: 10.1088/0022-3719/14/19/013.

[13] Tamara Broderick et al. "Faster Solutions of the Inverse Pairwise Ising Problem". en. In: *arXiv:0712.2437 [cond-mat, q-bio]* (Dec. 2007). arXiv: 0712.2437 [cond-mat, q-bio].

[14] Xiaowen Chen et al. "Searching for Collective Behavior in a Small Brain". en. In: *arXiv:1810.07623 [cond-mat, physics:physics, q-bio]* (Oct. 2018). arXiv: 1810.07623 [cond-mat, physics:physics, q-bio].

[15] Thomas M. Cover and Joy A. Thomas. *Elements of Information Theory*. 2nd. Hoboken: John Wiley & Sons, 2006.

[16] Bryan C. Daniels, David C. Krakauer, and Jessica C. Flack. "Control of Finite Critical Behaviour in a Small-Scale Social System". en. In: *Nature Communications* 8 (Feb. 2017), p. 14301. ISSN: 2041-1723. DOI: 10.1038/ncomms14301.

[17] Philippe De Donder, Michel Le Breton, and Eugenio Peluso. "Majority Voting in Multidimensional Policy Spaces: Kramer-Shepsle versus Stack-

elberg". en. In: *Journal of Public Economic Theory* 14.6 (Dec. 2012), pp. 879–909. ISSN: 10973923. DOI: 10.1111/jpet.12001.

[18] Anthony Downs. *An Economic Theory of Democracy*. New York: Harper, 1957.

[19] Lee Epstein and Tonja Jacobi. "The Strategic Analysis of Judicial Decisions". en. In: *Annual Review of Law and Social Science* 6.1 (Dec. 2010), pp. 341–358. ISSN: 1550-3585, 1550-3631. DOI: 10.1146/annurev-lawsocsci-102209-152921.

[20] Lee Epstein, Jeffrey A. Segal, and Harold J. Spaeth. "The Norm of Consensus on the U.S. Supreme Court". en. In: *American Journal of Political Science* 45.2 (Apr. 2001), p. 362. ISSN: 00925853. DOI: 10.2307/2669346.

[21] Nir Friedman et al. "Universal Critical Dynamics in High Resolution Neuronal Avalanche Data". en. In: *Physical Review Letters* 108.20 (May 2012), p. 208102. ISSN: 0031-9007, 1079-7114. DOI: 10.1103/PhysRevLett.108.208102.

[22] Noah Giansiracusa and Cameron Ricciardi. "Computational Geometry and the U.S. Supreme Court". en. In: *Mathematical Social Sciences* 98 (Mar. 2019), pp. 1–9. ISSN: 01654896. DOI: 10.1016/j.mathsocsci.2018.12.001.

[23] Bernard Grofman and Timothy J. Brazill. "Identifying the Median Justice on the Supreme Court through Multidimensional Scaling: Analysis of "Natural Courts" 1953–1991". en. In: *Public Choice* 112.1/2 (July 2002), pp. 55–79. ISSN: 0048-5829. DOI: 10.1023/A:1015601614637.

[24] Roger Guimerà and Marta Sales-Pardo. "Justice Blocks and Predictability of U.S. Supreme Court Votes". en. In: *PLoS ONE* 6.11 (Nov. 2011). Ed. by

Yamir Moreno, e27188. ISSN: 1932-6203. DOI: 10.1371/journal.pone.0027188.

[25] Gavin Hall and William Bialek. "The Statistical Mechanics of Twitter". en. In: *arXiv:1812.07029 [physics]* (Dec. 2018). arXiv: 1812.07029 [physics].

[26] Daniel E Ho and Kevin M Quinn. "How Not to Lie with Judicial Votes: Misconceptions, Measurement, and Models". en. In: *California Law Review* 98 (2010), pp. 813–876.

[27] John J. Hopfield. "Neural Networks and Physical Systems with Emergent Collective Computational Abilities". In: *Proceedings of the National Academy of Sciences* 79 (Apr. 1982), pp. 2554–2558.

[28] Ernst Ising. "Beitrag zur Theorie des Ferromagnetismus". de. In: *Zeitschrift für Physik* 31.1 (Feb. 1925), pp. 253–258. ISSN: 0044-3328. DOI: 10.1007/BF02980577.

[29] E. T. Jaynes. "Information Theory and Statistical Mechanics". en. In: *Physical Review* 106.4 (May 1957), pp. 620–630. ISSN: 0031-899X. DOI: 10.1103/PhysRev.106.620.

[30] Wensheng Kang, Ronald A. Ratti, and Kyung Hwan Yoon. "The Impact of Oil Price Shocks on the Stock Market Return and Volatility Relationship". In: *Journal of International Financial Markets, Institutions and Money* 34 (Jan. 2015), pp. 41–54. ISSN: 1042-4431. DOI: 10.1016/j.intfin.2014.11.002.

[31] Daniel Martin Katz, Michael J. Bommarito, and Josh Blackman. "A General Approach for Predicting the Behavior of the Supreme Court of the United States". en. In: *PLoS ONE* 12.4 (Apr. 2017). Ed. by Luís A. Nunes

Amaral, e0174698. ISSN: 1932-6203. DOI: 10.1371/journal.pone.0174698.

[32] C. Kemp and J. B. Tenenbaum. "The Discovery of Structural Form". en. In: *Proceedings of the National Academy of Sciences* 105.31 (Aug. 2008), pp. 10687–10692. ISSN: 0027-8424, 1091-6490. DOI: 10.1073/pnas.0802631105.

[33] Benjamin E. Lauderdale and Tom S. Clark. "The Supreme Court's Many Median Justices". en. In: *American Political Science Review* 106.4 (Nov. 2012), pp. 847–866. ISSN: 0003-0554, 1537-5943. DOI: 10.1017/S0003055412000469.

[34] Brian L Lawson, Michael E Orrison, and David T Uminsky. "Spectral Analysis of the Supreme Court". In: *Mathematics Magazine* 79.5 (Dec. 2006), p. 340.

[35] Edward D. Lee. "Partisan Intuition Belies Strong, Institutional Consensus and Wide Zipf's Law for Voting Blocs in US Supreme Court". en. In: *Journal of Statistical Physics* 173.6 (Dec. 2018), pp. 1722–1733. ISSN: 0022-4715, 1572-9613. DOI: 10.1007/s10955-018-2156-0.

[36] Edward D. Lee, Chase P. Broedersz, and William Bialek. "Statistical Mechanics of the US Supreme Court". en. In: *Journal of Statistical Physics* 160.2 (July 2015), pp. 275–301. ISSN: 0022-4715, 1572-9613. DOI: 10.1007/s10955-015-1253-6.

[37] Edward D Lee et al. "Sensitivity of Collective Outcomes Identifies Pivotal Components". en. In: (2019), p. 14.

[38] Jeffrey Lewis. *California Assembly and Senate Roll Call Votes, 1993 to the Present*. http://amypond.sscnet.ucla.edu/california/. 2019.

[39] Jeffrey B. Lewis et al. *Voteview: Congressional Roll-Call Votes Database*. https://voteview.com/. 2019.

[40] Dwayne Liburd and Sonia Barbosa. *State Supreme Court Data Project*. en. Jan. 2009. DOI: 10.7910/DVN/Z80F7P.

[41] Yang-Yu Liu, Jean-Jacques Slotine, and Albert-László Barabási. "Controllability of Complex Networks". en. In: *Nature* 473.7346 (May 2011), pp. 167–173. ISSN: 1476-4687. DOI: 10.1038/nature10011.

[42] B. B. Machta et al. "Parameter Space Compression Underlies Emergent Theories and Predictive Models". en. In: *Science* 342.6158 (Nov. 2013), pp. 604–607. ISSN: 0036-8075, 1095-9203. DOI: 10.1126/science.1238723.

[43] David J C MacKay. *Information Theory, Inference, and Learning Algorithms*. en. Cambridge University Press, 2003.

[44] Andrew D. Martin and Kevin M. Quinn. "Dynamic Ideal Point Estimation via Markov Chain Monte Carlo for the U.S. Supreme Court, 1953–1999". en. In: *Political Analysis* 10.02 (2002), pp. 134–153. ISSN: 1047-1987, 1476-4989. DOI: 10.1093/pan/10.2.134.

[45] Andrew D. Martin et al. "Competing Approaches to Predicting Supreme Court Decision Making". en. In: *Perspectives on Politics* 2.04 (Dec. 2004), pp. 761–767. ISSN: 1537-5927, 1541-0986. DOI: 10.1017/S1537592704040502.

[46] M. A. Moore and Helmut G. Katzgraber. "Dealing with Correlated Choices: How a Spin-Glass Model Can Help Political Parties Select Their Policies". en. In: *Physical Review E* 90.4 (Oct. 2014), p. 042117. ISSN: 1539-3755, 1550-2376. DOI: 10.1103/PhysRevE.90.042117.

[47] Mark E. J. Newman. “Power Laws, Pareto Distributions and Zipf’s Law”. en. In: *Contemporary Physics* 46.5 (Sept. 2005), pp. 323–351. ISSN: 0010-7514, 1366-5812. DOI: 10.1080/00107510500052444.

[48] H. Chau Nguyen, Riccardo Zecchina, and Johannes Berg. “Inverse Statistical Problems: From the Inverse Ising Problem to Data Science”. en. In: *Advances in Physics* 66.3 (July 2017), pp. 197–261. ISSN: 0001-8732, 1460-6976. DOI: 10.1080/00018732.2017.1341604.

[49] Hidetoshi Nishimori. *Statistical Physics of Spin Glasses and Information Processing*. en. Oxford University Press, July 2001. ISBN: 978-0-19-850941-7. DOI: 10.1093/acprof:oso/9780198509417.001.0001.

[50] F. Pedregosa et al. “Scikit-Learn: Machine Learning in Python”. In: *Journal of Machine Learning Research* 12 (2011), pp. 2825–2830.

[51] Adrián Ponce-Alvarez et al. “Whole-Brain Neuronal Activity Displays Crackling Noise Dynamics”. en. In: *Neuron* 100.6 (Dec. 2018), 1446–1459.e6. ISSN: 08966273. DOI: 10.1016/j.neuron.2018.10.045.

[52] Keith T. Poole and Howard Rosenthal. “A Spatial Model for Legislative Roll Call Analysis”. en. In: *American Journal of Political Science* 29.2 (May 1985), p. 357. ISSN: 00925853. DOI: 10.2307/2111172.

[53] Keith Poole et al. “Scaling Roll Call Votes with Wnominate in R”. en. In: *Journal of Statistical Software* 42.14 (June 2011), p. 21.

[54] Katherine N. Quinn. “Patterns of Structural Hierarchies in Complex Systems”. PhD thesis. Cornell University, 2019.

[55] C E Rasmussen and C K I Williams. *Gaussian Processes for Machine Learning*. Cambridge: MIT Press, 2006.

[56] L.E. Reichl. *A Modern Course in Statistical Physics*. Physics Textbook. Wiley, 2009. ISBN: 978-3-527-40782-8.

[57] Theodore W Ruger et al. “Competing Approaches to Predicting Supreme Court Decision Making”. In: *Colum L Rev* (2002).

[58] Marc Santolini and Albert-László Barabási. “Predicting Perturbation Patterns from the Topology of Biological Networks”. en. In: *Proceedings of the National Academy of Sciences* 115.27 (July 2018), E6375–E6383. ISSN: 0027-8424, 1091-6490. DOI: 10.1073/pnas.1720589115.

[59] Elad Schneidman et al. “Weak Pairwise Correlations Imply Strongly Correlated Network States in a Neural Population”. en. In: *Nature* 440.7087 (Apr. 2006), pp. 1007–1012. ISSN: 0028-0836, 1476-4687. DOI: 10.1038/nature04701.

[60] David J. Schwab, Ilya Nemenman, and Pankaj Mehta. “Zipf’s Law and Criticality in Multivariate Data without Fine-Tuning”. en. In: *Physical Review Letters* 113.6 (Aug. 2014), p. 068102. ISSN: 0031-9007, 1079-7114. DOI: 10.1103/PhysRevLett.113.068102.

[61] Jeffrey A. Segal and Harold J. Spaeth. *The Supreme Court and the Attitudinal Model Revisited*. New York: Cambridge University Press, 2002.

[62] Jeffrey A. Segal et al. “Ideological Values and the Votes of U.S. Supreme Court Justices Revisited”. en. In: *The Journal of Politics* 57.3 (Aug. 1995), pp. 812–823. ISSN: 0022-3816, 1468-2508. DOI: 10.2307/2960194.

[63] C E Shannon. “A Mathematical Theory of Communication”. en. In: *The Bell System Technical Journal* 27 (1948), pp. 379–423, 623–656.

[64] Claude Elwood Shannon. “A Mathematical Theory of Communication”. In: *Bell Syst Tech J* 27 (July 1948), pp. 379–423.

[65] Yair Shemesh et al. “High-Order Social Interactions in Groups of Mice”. en. In: *eLife* 2 (Sept. 2013), e00759. ISSN: 2050-084X. DOI: 10.7554/eLife.00759.

[66] David Sherrington and Scott Kirkpatrick. “Solvable Model of a Spin-Glass”. en. In: *Physical Review Letters* 35.26 (Dec. 1975), pp. 1792–1796. ISSN: 0031-9007. DOI: 10.1103/PhysRevLett.35.1792.

[67] L. Sirovich. “A Pattern Analysis of the Second Rehnquist U.S. Supreme Court”. en. In: *Proceedings of the National Academy of Sciences* 100.13 (June 2003), pp. 7432–7437. ISSN: 0027-8424, 1091-6490. DOI: 10.1073/pnas.1132164100.

[68] Harold J. Spaeth et al. *Supreme Court Database*. http://Supremecourt-database.org. 2016.

[69] C. Neal Tate. “Personal Attribute Models of the Voting Behavior of U.S. Supreme Court Justices: Liberalism in Civil Liberties and Economics Decisions, 1946–1978”. en. In: *American Political Science Review* 75.2 (June 1981), pp. 355–367. ISSN: 0003-0554, 1537-5943. DOI: 10.2307/1961370.

[70] “The Court’s Uncompromising Libertarian”. In: *Time* 106.21 (Nov. 1975), p. 77.

[71] Mark K. Transtrum, Benjamin B. Machta, and James P. Sethna. “Geometry of Nonlinear Least Squares with Applications to Sloppy Models and Optimization”. en. In: *Physical Review E* 83.3 (Mar. 2011), p. 036701. ISSN: 1539-3755, 1550-2376. DOI: 10.1103/PhysRevE.83.036701.

[72] M.I. Urofsky. *Dissent and the Supreme Court: Its Role in the Court’s History and the Nation’s Constitutional Dialogue*. Knopf Doubleday Publishing Group, 2017. ISBN: 978-0-307-74132-5.

[73] Thomas G. Walker, Lee Epstein, and William J. Dixon. “On the Mysterious Demise of Consensual Norms in the United States Supreme Court”. en. In: *The Journal of Politics* 50.2 (May 1988), pp. 361–389. ISSN: 0022-3816, 1468-2508. DOI: 10.2307/2131799.

[74] M. Weigt et al. “Identification of Direct Residue Contacts in Protein-Protein Interaction by Message Passing”. en. In: *Proceedings of the National Academy of Sciences* 106.1 (Jan. 2009), pp. 67–72. ISSN: 0027-8424, 1091-6490. DOI: 10.1073/pnas.0805923106.

[75] Jorge Gomez Tejeda Zañudo, Gang Yang, and Réka Albert. “Structure-Based Control of Complex Networks with Nonlinear Dynamics”. en. In: *Proceedings of the National Academy of Sciences* 114.28 (July 2017), pp. 7234–7239. ISSN: 0027-8424, 1091-6490. DOI: 10.1073/pnas.1617387114.

[76] James R. Zink, James F. Spriggs, and John T. Scott. “Courting the Public: The Influence of Decision Attributes on Individuals’ Views of Court Opinions”. en. In: *The Journal of Politics* 71.3 (July 2009), pp. 909–925. ISSN: 0022-3816, 1468-2508. DOI: 10.1017/S0022381609090793.

CHAPTER 2

CONFLICT DYNAMICS

Conflict is omnipresent in nature. When components have misaligned interests, competition may lead to antagonistic interactions such as when the immune system fights a pathogen, voting blocs are competing to form a winning coalition, and when armed groups are battling for control of territory. Despite the ubiquity of conflict, the causes of variation in conflict size and duration are poorly understood. In this chapter, we study the dynamics of conflict in a pigtailed macaque society, published in reference [36], and the spread of armed conflict, developing quantitative models that allow us to peer into the social interactions driving the spread of conflict [37].

In this chapter, we discover surprising regularities in the statistics of conflict in social systems. One might expect that the vicissitudes of each particular conflict dominate such that the ensemble of many show no simple pattern, as is emphasized in historical descriptions of wars that focus on particular circumstances [35]. Instead, we find that for both pigtailed macaque and human societies that the dynamics of conflict show universal structure in their growth in size and temporal dynamics. These regularities suggest that the dynamics of conflict are largely shaped by physical constraints in each respective system.

Indeed, the idea that conflicts are shaped by physical constraints is not new. In 1948, the polymath L.F. Richardson discovered that the distribution of fatalities in interstate wars was distributed as a power law. Grasping at the enticing possibility of a general theory of armed conflict, he proposed that wars were

a function of the length of the boundary between countries [47]. Although his theory never became widely accepted, his initial foray led to much speculation on the mechanisms of conflict. Today, we have a much better theoretical understanding of how such statistical regularities arise from complex systems from our understanding of renormalization group theory [51]. The theory provides a reason for why we might be able to construct reduced theories of conflict that fail to capture all the microscopic details of the system yet enough of the essentials to be mathematically valid.

As a first step, we focus on a small, closed macaque model system that has been well-studied and for which a great of deal of detailed information has been collected. We study a society formerly housed in a social compound at the Yerkes National Primate Research Facility. These are a social species of monkey that form complex societies of many dozens of individuals with a range of social roles mediated by a web of interactions. Unlike human conflict where data can be difficult to collect, this society presents a high-resolution, controlled starting point for building intuition about how conflicts nucleate and grow [21, 22, 38]. We discover in this system that the distributions of the duration of conflict episodes and interleaving peaceful periods furnish hints for the dynamics driving them [36]. The statistics of conflict reveal that peace is interrupted by rogue actors, whereas conflict ends because pairs of interacting actors resolve their disputes sequentially.

Rescaling conflict distributions reveals a universal curve, showing that the typical time-scale of correlated interactions exceeds nearly all individual fights. This temporal correlation implies collective memory across pairwise interactions be-

yond those assumed in standard models of contagion growth or iterated evolutionary games. By accounting for memory, we make quantitative predictions for interventions that mitigate or enhance the spread of conflict. Managing conflict involves balancing the efficient use of limited resources with an intervention strategy that allows for conflict while keeping it contained and controlled.

Propelled by our success with finding emergent scaling regularities in macaque conflict, we move on to consider a much larger and more difficult-to-measure system of armed conflict spanning the African subcontinent during the two decades 1997–2016 [30, 37]. Despite the vastly more complex nature of human conflict, we likewise find striking emergent scaling regularities. In contrast with macaque conflict, armed conflict is richer, showing dependence on large spatial scales and evidence of disorder that drives conflict differently in varying locations.

By systematically clustering spatiotemporally proximate events into conflict avalanches, we show that the number of conflict reports, fatalities, duration, and geographic extent satisfy consistent power law scaling relations. The temporal evolution of conflicts measured by these scaling variables display emergent symmetry, collapsing onto a universal dynamical profile over a range of scales. The measured exponents and dynamical profiles describe a system distinct from prevailing explanations of conflict growth such as forest fire models. Our findings suggest a "thermodynamics" of armed conflict in which armed conflicts are dominated by a low-dimensional process that scales with physical dimensions in a surprisingly unified and predictable way.

2.1 Conflict dynamics in pigtailed macaques

In biology, conflict plays a central role in structuring interactions among components whether genes, cells, or individuals. Conflict occurs when components have only partially aligned interests [7, 24]. Examples of conflict include fights among group members in animal and human societies, infection, immune responses, and even autoimmunity, in which conflict arises when an immune response targets self instead of a pathogen. Conflict growth—the spread of conflict from a small number of antagonistic or infected components to many—has been modeled in a diversity of systems as a contagion process [34], where the resulting conflict duration (e.g., length of fights or duration of infection [31]) can vary over several orders of magnitude.

A growing body of work suggests that a benefit of small, contained conflicts is that they allow components to test and refine strategies at relatively low cost [33, 54]. This can facilitate adaptation and innovation [20]. Large and long conflicts have been associated with system instability and increased component mortality [5, 21, 34, 40]. The key factors influencing conflict size and duration across a range of biological systems are not yet understood, although in the specific case of primate conflict some factors contributing to conflict size have been identified [6, 56].

We investigate the dynamics of conflict duration using an animal society model system—a group of captive, socially-housed pigtailed macaques (*Macaca nemestrina*, $N = 64$) at the Yerkes National Primate Research Center (Appendix D.1). Here, conflicts can manifest as fights. A fight starts when one individual threatens or attacks a second individual. The total number

of individuals participating in a given fight ranges from 2 to 35, with third-parties becoming involved through intervention and redirected aggression (Appendix D.1).

Fights have clear, operationally-defined starting times and endpoints and generally only one fight is active at a given time (Appendix D.1). This produces a time series of fights separated by peaceful periods. The time series was measured at the resolution of seconds, collected over a period of roughly four months, and contains ~1000 cycles of peace and conflict. We observe a wide range of conflict durations (1–840 s) and peace durations (2–5,570 s). We use these data to infer and characterize the dynamics underlying the durations of conflict and peace.

We find that peaceful periods are characterized by a "first to fail" process (as in reliability theory [3]) in which the duration of peace depends only on the waiting time for the first pair of individuals to begin fighting. The distribution of peaceful periods is exponential, consistent with pairs independently choosing to begin fights, and the likelihood of remaining in the peaceful state does not depend strongly on the recent past.

We find conflict, on the other hand, displays increased variance in duration, consistent with strong correlations between the aggressive interactions that occur within a fight. This suggests a "first to fight" mechanism: the beginning of the conflict influences the duration of consecutive pairwise interactions, and the end of fighting retains a *collective memory* of the start. Collective memory is encoded in the aggregate interactions in a conflict (we return later to how this

collective memory arises and whether it implies individual memory). As a result, distributions of fight duration are consistent with a universal lognormal distribution under a simple scaling transformation that accounts for fight size. In contrast with common models of conflict growth [19], fight duration cannot be explained using a simple memoryless contagion process.

2.1.1 Peace dynamics

We first investigate the distribution of peaceful durations. As we show in Figure 2.1, this distribution has a nearly exponential tail with small deviations, and a maximum likelihood fit to the exponential distribution

$$p(t) = \Lambda e^{-\Lambda t} \tag{2.1}$$

returns a conflict outbreak rate parameter of $1/\Lambda = 414\,\mathrm{s}$. We plot this distribution as a complementary cumulative distribution function (CDF),

$$1 - \int_0^t \Lambda e^{-\Lambda t'}\, dt' = e^{-\Lambda t}, \tag{2.2}$$

which obviates the need for bins to construct a probability density and makes clear the exponential decay in the tail. An exponential distribution is a signature of a process where at each moment in time there is a constant probability that the peaceful period ends. When there are many processes running concurrently that could lead to the outbreak of conflict with timescales λ_i, the combination of those still return an exponential with rate $\Lambda = \sum_i \lambda_i$. Thus, peaceful durations in macaque conflict are consistent with a "rogue actor" model, where troublemakers interrupt the peace in an independent manner.

Despite the close agreement with the data, we might wonder why there is a systematic deviation at the tail of the distribution for the longest peace durations. If

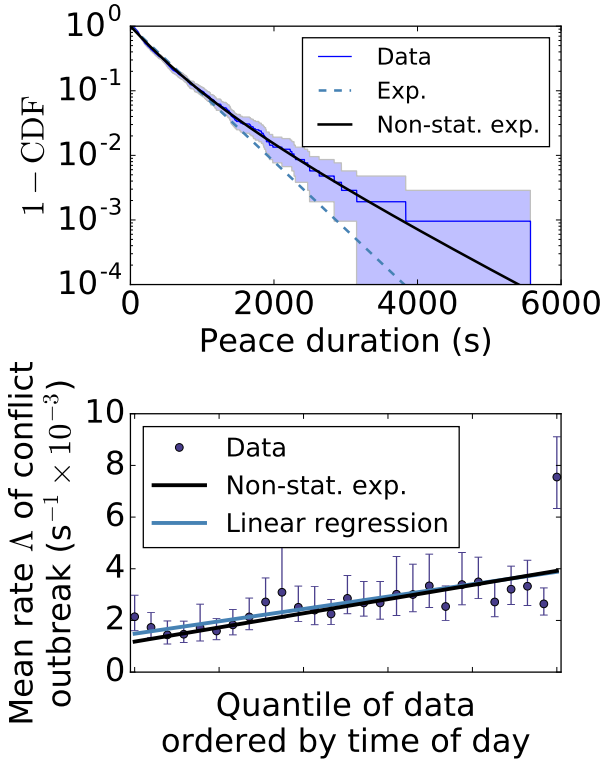


Figure 2.1: Peace durations in pigtailed macaque society. (top) Distribution of peace durations decay with an exponential tail with typical duration $1/\Lambda \approx 400\,\mathrm{s}$. (bottom) Rate of conflict outbreak increases close to linearly over the course of the day, proportional to the number of hours macaques spent in the social compound before being moved to individual cages. Linear regression to the shown points was weighted by the number of points in each data point. Rightmost point contains only 31 points compared to 39 for every other point.

the model were perfectly accurate, we would expect that the data would wander above and below the tail of the distribution, but this is not the case. We know from qualitative observation that the macaques are more agitated at certain parts of the day compared to others. These special times are around meal times and near the end of the day in the compound (and before being caged). Such nonstationary effects manifest quantitatively in the instantaneous rate of conflict outbreak over the course of the day, which indicates rise in rate around meal time and a generally increasing trend towards the end of the day as shown in Figure 2.1.

To capture this variation in the rate of conflict outbreak, we imagine that instead of a single outbreak rate Λ, the distribution is a mixture of many different average rates Λ—imagine that the set of subprocesses λ_i fluctuate. A simple version of this hypothesis is to assume that Λ comes from a uniform distribution with mean $\bar{\Lambda}$ and width $\Delta\Lambda$ such that

$$p(\Lambda) = \frac{1}{2\Delta\Lambda}. \tag{2.3}$$

The resulting distribution for peaceful duration is

$$p(t, \Lambda) = \frac{\Lambda}{2\Delta\bar{\Lambda}} e^{-\Lambda t} \tag{2.4}$$

and $\bar{\Lambda} - \Delta\Lambda \leq \Lambda \leq \bar{\Lambda} + \Delta\Lambda$. Maximizing the likelihood of the mixture model to find the parameters $\bar{\Lambda}$ and $\Delta\Lambda$, we recover reasonable agreement between the average timescale in this new model $\bar{\Lambda}^{-1} \approx 540\,\mathrm{s}$ and the simpler model's $\Lambda^{-1} = 414\,\mathrm{s}$. Furthermore, we find that we capture the rising tail of the distribution, the most obvious systematic deviation from the exponential, closely.

Typically, we expect that models with more parameters do better, so we go through another check of our more complicated hypothesis. We might imagine that the variability in Λ might be ordered by time such that the slowest rate of conflict outbreak occurs at the beginning of the day and the fastest rate near the end. If this were the case, the change in Λ would serve as a linear interpolation of the rate at which conflict outbreak changes. We overlay the data with the maximum likelihood fit and find close agreement that almost exactly aligns with a simple linear regression on the shown change in rate. Importantly, when we fit the mixture distribution defined in Eq 2.4, we imposed no chronological structure on the data. Yet, we find that increasing time spent in the compound explains nonstationarity in the rate of conflict outbreak.

As a final check, we consider another potential hypothesis that might serve as a reasonable, but alternative explanation for the unusual behavior in the exponential distribution. We might wonder if longer peaceful periods are simply less stable or become more stable over time instead of a changing mean over the course of the day. This possibility is captured by the Weibull distribution, with probability density function $p(t) = k/\Lambda(t/\Lambda)^{k-1}e^{(t/\Lambda)^k}$, such that the rate of decay changes as a power law. Unlike our nonstationary hypothesis, a maximum likelihood fit of the Weibull distribution yields no improvement, and we find $k = 1$, reaffirming our finding that peace durations are indeed exponentially distributed though the average rate evolves with a timescale exceeding the disintegration of any single peaceful period. Such a separation of timescales suggests there are two separate processes driving the dynamics of peace in pigtailed macaque society: one driving the outbreak of the next conflict and another determining the average rate of conflict outbreak over the course of many hours.

The durations of conflict follow a very different distribution from that of peace. Whereas peace is characterized by a quickly decaying, nearly exponential tail, conflict has a much heavier tail indicating that the longest conflicts occur with non-neglible probability. Fitting various distributions (exponential, Weibull, power law, and lognormal) with maximum likelihood to the distribution of conflict duration in Figure 2.2, we show that the distribution is most closely described by a lognormal distribution,

$$p(t) = \frac{1}{\sqrt{2\pi}\tilde{\sigma}x}\exp\left(-[\log(t/\tilde{\mu})]^2/2\tilde{\sigma}^2\right). \tag{2.5}$$

The heavy tail indicates that conflict is driven by processes that enhance variability—such as strong social interactions—providing a hint as to why

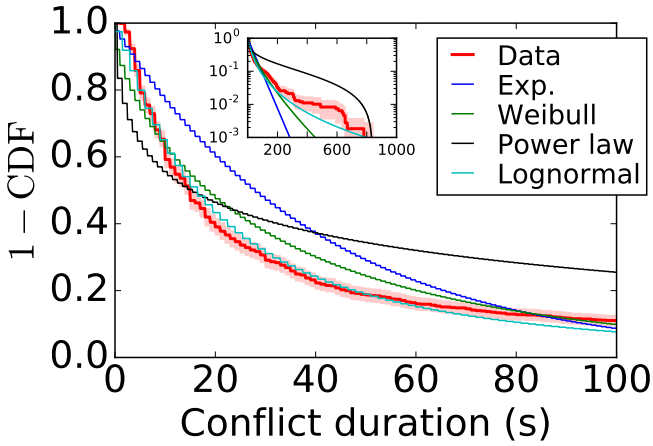


Figure 2.2: Distribution of fight duration. Maximum likelihood fits for the exponential, Weibull, power law, and lognormal distributions are shown after accounting for the temporal discretization of the data.

conflict differs from peace.

2.1.2 Conflict dynamics

If social interaction could explain why the duration distribution of fights were different from peace, then we might expect that this variability scale with the number of participants, the "fight size." We inspect the duration distribution when conditioned on the number of participants n, or $p(t|n)$. Since the means of these distributions grow with size, we first center them for a fair comparison in logarithmic scale, equivalent to rescaling in linear space. After rescaling by the geometric means in Figure 2.3, we find to our surprise that the distributions collapse onto a universal profile with the same width! The collapse shows that that the distributions follow a universal, scale-free form

$$p(t) = \phi(t/\tilde{\mu}_n)/\tilde{\mu}_n. \tag{2.6}$$

The form of Eq 2.6 depends on the presence of a dominant scaling variable that determines how larger fights last longer.

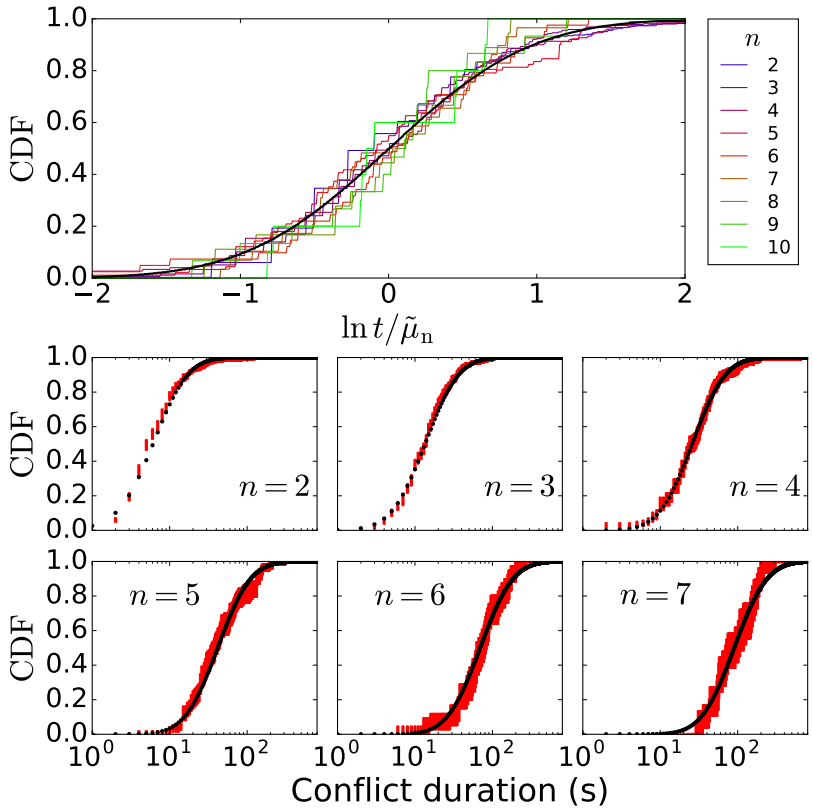


Figure 2.3: Scaling collapse of fight duration distributions to universal lognormal form. (top) All fight duration distributions for $2 \leq n \leq 10$ rescaled by geometric mean. These account for 97% of the data. (bottom) Lognormal model discretized to second resolution as is the data for fights of different sizes n.

The profile in Figure 2.3 is well-described by a lognormal distribution as given in Eq 2.5 and obeys the scale-free form given in Eq 2.6. The lognormal is characterized by the geometric standard deviation $\exp\tilde{\sigma} = 0.76 \pm 0.07$. This corresponds to a coefficient of variation of $\sigma_n/\mu_n = \sqrt{\exp\tilde{\sigma}^2 - 1} = 0.88$. That is, fights of a given size have fluctuations in duration with magnitude approximately 88% of the mean across all observed fight sizes. The collapse implies multiplicative scaling similar to Weber's Law, though the size of fluctuations is too large to be explained simply as errors in temporal perception.[1] Furthermore, we check that

[1] In perceptual tasks, this is related to the Weber ratio: if fights were to end simply by individuals stopping after some fixed perceived duration, then we expect a corresponding temporal Weber ratio roughly the size of the observed coefficient of variation. Temporal Weber ratios are typically 0.5 or less in both animals and humans, depending on the particular task and the dura-

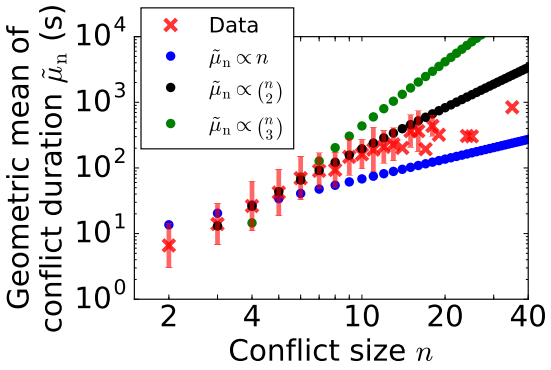


Figure 2.4: Scaling of the geometric mean of fight durations with fight size in pigtailed macaque society. Given the heavy tail in duration, we show the geometric mean as opposed to the arithmetic mean. Points $n = 2$ and $n = 3$ for the binomial and trinomial models, respectively, are given by the data. For the trinomial, no prediction is made for the $n = 2$ point.

the discretization of the data into seconds agrees with the distributions for small fights in the bottom of Figure 2.3. We plot the distribution after assuming that real-valued durations from the distributions are rounded to the nearest second. The model fits the data well across the entire range of 1 s to 1000 s for fights of sizes $2 \leq n \leq 7$ with error bars determined by bootstrap sampling of the data. Thus, the lognormal distribution provides a compelling model for the variation of conflict duration about the geometric mean irrespective of conflict size.

This temporal collapse indicates that the way that the average fight duration $\tilde{\mu}_n$ scales with the number of involved actors n is an important factor. Since conflict is defined by the presence of aggressive interactions between actors, we compare how the data scales with the number of subgroups of size k in Figure 2.4,

$$\tilde{\mu}_n = \alpha \tilde{\mu}_k \binom{n}{k}. \tag{2.7}$$

tion of the interval being estimated [1, 26, 39]. Our measured value of 0.9 is significantly larger, suggesting variation beyond that implied by individual temporal estimation.

Just from visual inspection, it is clear that the data scales most closely to $k = 2$. We check with the weighted least-squares error on log-scaled axes to ensure that this is indeed the closest scaling. Though none of the models fit the durations of the longest conflicts $n > 20$, the small number of data points at that range makes it difficult to distinguish the viability of the linear and binomial models from each other based only on those data points (Figure 2.5). The overwhelming majority of data points are concentrated for $n \leq 10$, so it is the case that the binomial model captures the scaling of conflict duration closely for the vast majority of conflicts that we observe in the data.

It is remarkable that the binomial scaling form given by Eq 2.7 provides such a close fit to the growth of the mean duration. By fixing a relationship between the linear and quadratic coefficients, we impose a much stronger hypothesis about the nature of the scaling, yet one that hews closely to the data. The resulting scaling signals to us that fights are dominated by the number of potential pairwise interactions. By measuring the coefficient in units of the average duration of pairwise interactions $\tilde{\mu}_2$, we find that $\alpha = 0.66$ from Eq 2.7. For triadic fights, this coefficient indicates that typically two of the three pairwise interactions contribute to the duration of the conflict, a prediction consistent with the number of aggressive interactions typically seen in triadic fights. From measuring the scaling of conflict duration, we thus gain insight into the structure of macaque conflict, suggesting an intuitive picture where on average 2/3 of all possible pairwise interactions are realized in a sequential manner.

This observation presents a simple model for the dynamics of macaque conflict. If we imagine that each pairwise interaction occurs roughly in sequence, per-

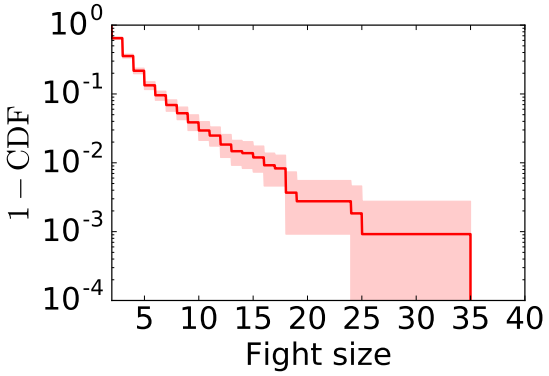


Figure 2.5: Distribution of fights with n participants in pigtailed macaque society.

mitting some random overlap, then the total duration of a conflict would be the summation of $\alpha\binom{n}{2}$ random variables with average duration $\tilde{\mu}_2$. If these random durations were independent of one another, however, the Central Limit Theorem shows that the distribution would have a well-defined mean with exponentially decaying tails. Instead, the way that the variance scales with the mean suggests that sequential conflict interactions have correlated durations. If these pairwise interactions had durations that were perfectly correlated, the variance would scale with the means as $\mu_n \propto \sigma_n^2$. In the data, we find that $\mu_n \propto \sigma_n^{1.8}$, close but not quite the perfect scaling that would lead to a perfect scaling collapse.

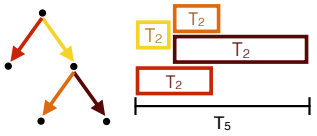
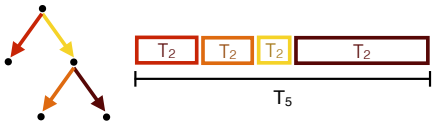
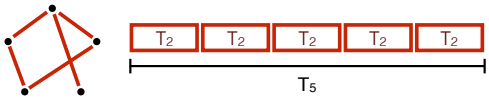
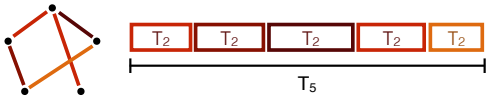
Yet another conundrum arises when we consider how the fights are spreading to new participants. As shown in previous work, the number of participants grows as a contagion process that recruits new participants into an ongoing fight such that conflict spreads in the social network in a tree-like fashion [13]. If it were the case that each new interaction recruiting an additional member contributed to the duration of the conflict, we would expect that the mean $\tilde{\mu}_n \sim n$, in disagreement with the observed binomial scaling. If we assume that each of

those "recruitment" interactions do not have duration that grows quadratically with conflict size, it must be the case that tree-like growth that characterizes conflict recruitment does not describe how the duration increases.

We organize several such possible explanations for the dynamics of conflict in Table 2.1. First, we consider the tree-like model where the temporal dynamics live on the same network as the recruitment process. If each recruitment interaction were to contribute to the final duration, but overlap as multiple branches of the tree grow simultaneously, then we would find sublinear scaling of duration with size n. If these recruitment interactions were to happen sequentially, we would still recover the incorrect scaling for mean duration. In contrast, when interactions occur between older and newer participants, the number of interactions scales with $\binom{n}{2}$ gives both the correct scaling for the mean. When the durations of sequential pairwise interactions are likewise correlated, the variance scales close to what is observed in the data.

In light of our observations, we propose a model of sequential pairwise interactions whose correlations in duration decay with a timescale τ. When $\tau \to 0$, each sequential pairwise interaction has a different random value selected from the lognormal model of the distribution of dyadic conflicts as given in Eq 2.5, where as $\tau \to \infty$ every sequential pairwise interaction has the same duration and we recover a perfect scaling collapse of conflict duration distributions conditioned on fight size. As we change τ to vary the amount of correlation between subsequent interactions, we cross over between these two limits. From the scaling collapse, we expect that $\tau \gg 1$. Thus, we picture a fight as a sequence of interactions whose durations randomly wander in duration space pictured in

Table 2.1: Hypotheses for conflict duration. Fight durations obey a simple scaling law in which the durations of fights of size n are rescaled versions of smaller fights, by a factor that scales as $\binom{n}{2}$. This growth with the number of pairs suggests that the durations of fights of size greater than 2 may arise from a process consisting of multiple pairwise interactions, each with duration T_2 sampled from the duration distribution for fights of size 2. In this table, we compare the outcomes of some possible processes. First, branching models are inconsistent with the data because the number of interactions, and thus the mean duration, grows too slowly with n. Second, sequentially adding a fraction of possible pairwise interactions ("pairwise sequential") produces the correct scaling of the mean, but insufficiently large standard deviation. Finally, sequentially adding correlated versions of pairwise durations can produce the correct scaling of the first two moments of the duration distribution. Only the perfect correlation case obeys an exact rescaling, but our data cannot distinguish between the "correlated" and "perfectly correlated" cases.

Model	Schematic for fight of size 5	Mean	Variance	Self-similar	Fits data
Branching simultaneous		μ sub-linear in n			✗
Branching sequential		$\mu \propto n$	$\sigma^2 \propto \mu$		✗
Pairwise sequential		$\mu \propto \binom{n}{2}$	$\sigma^2 \propto \mu$	No	✗
Pairwise sequential, perfectly correlated		$\mu \propto \binom{n}{2}$	$\sigma^2 \propto \mu^2$	Yes	✓
Pairwise sequential, correlated		$\mu \propto \binom{n}{2}$	$\sigma^2 \propto \mu^c$, $1 < c < 2$	Nearly	✓

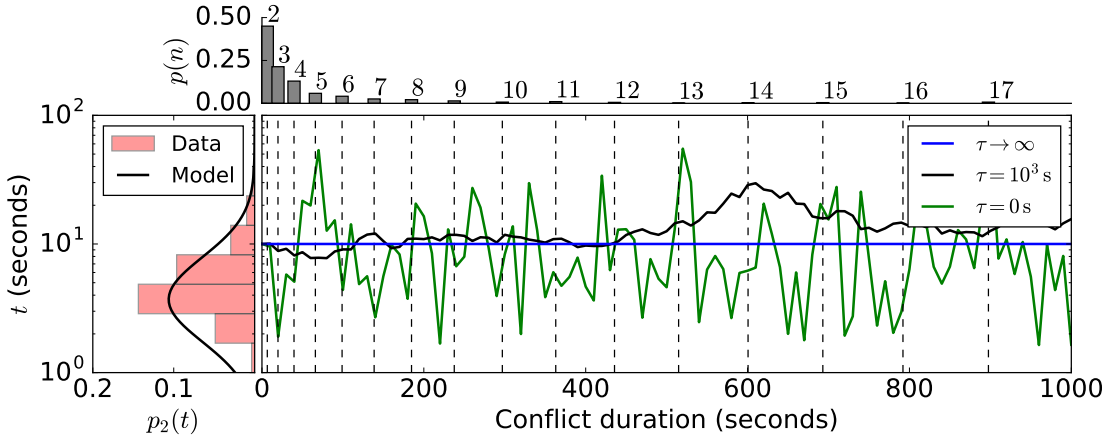


Figure 2.6: Diffusion model for fight durations in pigtailed macaque society. (left) The observed histogram of pairwise fight durations compared with the lognormal model in black. (right) We simulate three different fight trajectories for different correlation times τ. When $\tau \to \infty$, every sequential pairwise interaction has the same duration as the previous one. When $\tau \to 0$, every sequential pairwise interaction has a random duration sampled from the lognormal model. (top) The probability that a conflict with n participants lasts that long.

Figure 2.6, acting as a proxy for aggression, with a diffusion constant that determines the rate at which the system forgets how the fight started.

By maximizing the likelihood of the data, we find that sequential fight models with decorrelation times of $\tau > 270\,\mathrm{s}$ align closely with the data as we show in Figure 2.7. Although short correlation times $\tau = D^{-1}$ are sufficient to explain the distribution of triplet fight size distributions, the decorrelation between the durations of sequential interactions means that the distributions for larger fight sizes are poorly fit. As we increase τ, we improve our fit to the distribution of large conflict durations the log-likelihood of the model saturates. This flattening occurs for $\tau > 270\,\mathrm{s}$, a lower bound that corresponds to strong correlations between sequential pairwise episodes in a conflict. We combine the distributions per conflict size to construct the distribution of conflict duration beyond dyadic

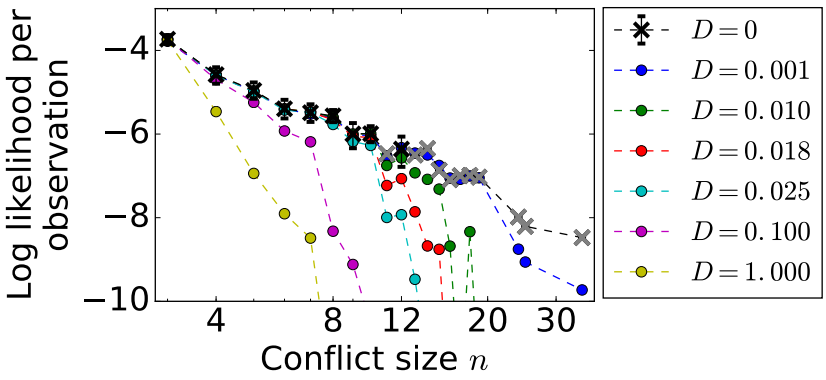


Figure 2.7: Maximum likelihood fit to decorrelation time to decorrelation time τ in diffusion model for pigtailed macaque conflict. The data is well fit when $\tau > 270$ s, the limit of strong correlations between sequential dyadic interactions.

fights, $p(t) = \sum_{n>2} p(t|n)p(n)$, in Figure 2.8. The model with no correlations $\tau = 0$ s fails to capture the distribution at short conflict duration. At long times, it converges to the strongly correlated model ($\tau = 380$ s) when the timescale for decay is comparable to the duration of the typical pairwise interaction and so correlations become less important for determining the distribution. The model with long-time correlations agrees closely with the data at both short and long conflicts, capturing the heavy-tailed nature of the distribution with relatively small deviations [2].

The lower bound on the correlation time at $\tau > 270$ s corresponds to strong correlations between sequential agonistic interactions. The typical conflict with three or more participants is 60 s long, and over 96% of those conflicts are shorter than 270 s, showing that nearly all observed fights retain correlated interactions over their entire duration. This observation suggests that temporal correlations are strong over the course of a single fight and that the length of the first pairwise incident significantly influences the evolution of the conflict over time.

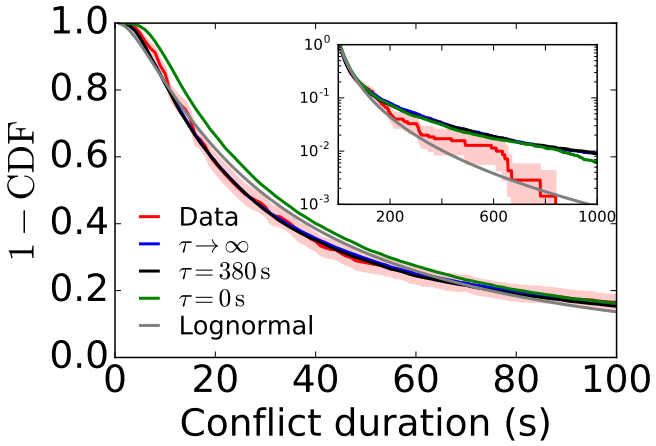


Figure 2.8: Diffusion model fit to the distribution of fight duration distribution $p(t)$.

Correlated interactions are likely to persist over the course of conflict due to cognitively-mediated or emotionally-mediated memory for past interactions. This collective memory could be reducible to individual memory if it is caused by individuals participating in multiple dyads within the same conflict and behaving similarly throughout. Alternatively, irreducible collective memory occurs when individual decisions to join and remain in a fight are a function of these decisions by others.[2] Different mechanisms for producing irreducible collective memory make different demands on individual-level memory: individuals may remember the duration or severity of (1) all previous dyads within the fight, (2) only the initial bout, or (3) only the immediately previous dyad within the fight.

2.1.3 Conflict prediction

To understand how conflict can be controlled, a crucial question to answer is how long and how large an ongoing conflict might become. For example, po-

[2]Such irreducibility is reflected in measures of synergy applied to conflict participation data [15].

licers, high-power individuals, in some macaque societies regulate conflict by intervening in fights [21, 22]. Monitoring conflicts, however, consumes time and attention and interventions carry the risk of injury, and this risk increases with fight size [21]. Knowing how long a conflict is likely to last given the number of individuals involved, or how big it will become given how long it has already lasted, would help intervening individuals decide how to distribute their interventions.

Using our model, we can estimate the probability that more individuals will join an ongoing fight and the probability that the fight will have a given duration. With the joint distribution of conflict size and duration $p(t, n) = p(t|n)p(n)$, the probability that a fight might be extended by time Δt with Δn additional members given that we have observed n_0 participants after t_0 elapsed time is an application of Bayes' theorem, which yields (Appendix D.4)

$$p(\Delta t, \Delta n | t_0, n_0) = \frac{p(t_0 + \Delta t | n_0 + \Delta n) p(n_0 + \Delta n)}{\sum_{\Delta n} \int p(t_0 + \Delta t | n_0 + \Delta n) p(n_0 + \Delta n)\, d\Delta t}. \tag{2.8}$$

In Figure 2.9, we show how the expected total fight size changes the longer a conflict with 2 individuals lasts. Since the probability that the dyadic conflict does not grow beyond the initial pair decays exponentially, one strategy for minimizing conflict size favors earlier intervention: it is more likely that the fight will grow the longer it has lasted. If the probability of a successful intervention terminating a fight becomes very difficult at, for example, a size of 6, our model suggests that an effective intervention time is ~15 s. After this point, the probability of the fight reaching size 6 is high. Similarly, if duration has important functional consequences, our model suggests the policer or conflict manager can estimate duration by monitoring conflict size (Appendix D.4).

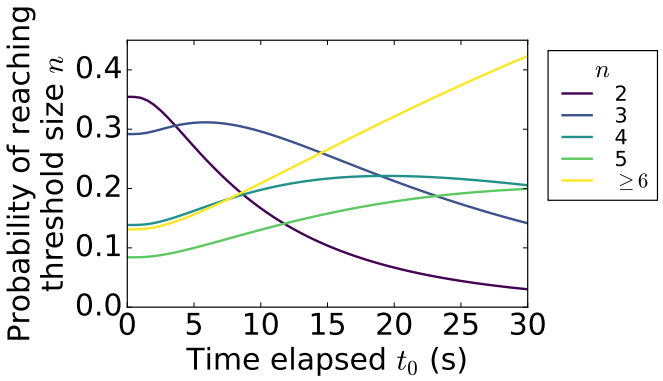


Figure 2.9: Predicted fight trajectories. Probabilities of conflict size growth for a dyadic fight as a function of time elapsed t_0 as in Eq 2.8. It is most likely that the conflict will not grow in first few seconds, but it quickly decays as the probability that it grows by at least one more member dominates between 4–15 s. At 15 s, the most likely outcome is a fight of size 6 or greater. Monitoring and intervention strategies can be developed depending on, for example, knowledge that interventions are ineffective after a critical fight size.

We cannot, however, use this analysis to identify the intervention strategies used by individuals in this system. The fights analyzed here already contain policing and other types of interventions. Some interventions terminated the fight or reduced its intensity, others had no apparent effect (except to increase fight size), and some exacerbated fight severity. This is a typical feature of fights in this system. Hence this analysis only reveals how an intervener could strategically apportion additional interventions given these events.

2.1.4 Discussion

Simple branching processes have been proposed to underlie many growth and contagion processes in biology and social science [14, 19]. We observe that for primate conflict a standard branching process does not naturally capture the temporal dynamics since (1) conflict durations are proportional to the number of pairs of individuals in a fight and not the number of individuals, and (2) bouts

within a conflict are influenced by how long previous dyadic events last [13]. Superlinear scaling suggests that conflict resolution requires time not merely for each agitated individual to become inactive but for a large fraction of pairwise relationships between involved individuals to be separately resolved. The duration correlation suggests that for fighting pairs the duration of fight bouts in the population is a salient conflict feature and conflict growth is a function of the collective memory of a group.

Collective memory, when it is not entirely reducible to individual memory, implies a long time scale associated with repeated interactions between participants in a conflict [17]. In other words the history of the conflict affects the game under consideration and the set of payoffs. The idea that conflict is shaped by a persistent social memory is of course not new. It was captured, for example, by Von Clausewitz when he wrote that "war is the continuation of politics by other means" [57]—the initial disagreement is present in every conflict. Our results show that there is a quantitative basis for this idea and hence that many of the frameworks typically applied to conflict, such as Markovian games with iterated interactions, may be inappropriate because they assume a short timescale.

The universal scaling of conflict durations in this society demonstrates the key role of memory in conflict growth. This scaling has implications for conflict management strategies, which are likely to reflect trade-offs between the costs and benefits of intervention and monitoring. In systems with collective memory, a relatively simple strategy to control conflict growth is to target initial interactions, which could be effective and efficient if sustained conflict monitoring is

costly or interventions are ineffective in larger conflicts.

In systems in which conflict managers can reliably estimate the relationship between conflict duration and the number of participants, we predict that interventions will be targeted toward specific optimal sizes or durations. This type of *targeted* conflict intervention tailored to conflict features stands in contrast to conflict management mechanisms that control conflict systemically at regular intervals [16]. We might expect such regular temporal control when managers cannot respond reliably to specific conflicts.

2.2 Emergent regularities and scaling in human conflict

In the previous section, we studied the temporal dynamics of conflict in a pig-tailed macaque society. Macaque conflict involved acts of physical aggression between a small number of individuals that did not result in death. Human violence, on the other hand, can spread to massive scales involving millions of people and thousands of kilometers in a coordinated fashion. The use of arms makes it much more likely to be fatal. Moreover, human conflict is seemingly more complex because it evolves on a background of varying geographic terrain, is influenced by local and global social institutions, and depends on the state of technological development.

Yet, human conflict displays signs of regularity that belie the additional layers of complexity. In 1941, Lewis Fry Richardson announced that the distribution of estimated casualties from interstate wars followed a power law distribution closely [48, 49]. More recently, Clauset, with another half century of conflict and more sophisticated statistical tests, confirmed that Richardson's Law held up well today as it did then [11]. Worryingly, it seems that the era of modern and catastrophic conflict has not abated, at least statistically speaking.

The scale-free distribution of fatalities in interstate wars presents a clue that the statistics of armed conflict, a more general category of conflict including interstate wars, may show strong regularities. Here, we search for those signatures across a range of behavior including organized, armed resistance by non-state actors and more spontaneous acts of violence like riots and protests. Here, we focus on political violence grouped into sometimes overlapping categories called Battles, Violence Against Civilians, and Riots/Protests. These groups con-

stitute three types of events recorded in the Armed Conflict Location & Event Data Project (ACLED) [45].

ACLED focuses on disaggregated conflict reports, or data points localized by time in units of days, space indicated by geographic coordinates, and actor identities on three types of conflicts. These data are aggregated by ACLED from news media organization or regional contacts. This kind of data that is disaggregated into discrete conflict points permits us to peer into the dynamics generating clusters of conflicts across time and space. This approach is much more informative than is relying on preaggregated clusters of events via sociopolitical criteria known as battles or wars. By systematically clustering such events into *conflict avalanches* relying only on spatiotemporal proximity, we reveal emergent scaling structure at large scales connecting many such conflict events. Such findings, consistent with predictions from the renormalization group, suggest that simple physical models could capture the dominant processes underlying the spread of armed conflict as described in the language of universality classes.

The central idea behind the renormalization group is the coarse-graining of a length scale that defines a mapping operation from one model to another. The coarse-graining operation describes a flow in the space of models that eventually leads to a fixed point, where separation of length scales leads to the emergence of characteristic, long-wavelength properties. Separating the basins of stable fixed points, or phases, are the critical manifolds corresponding to phase transitions. These correspond to unstable fixed points where the system becomes scale-invariant and the resulting power laws are described by a set of

critical exponents. In driven systems, these fixed points can become stable, e.g., the conservation of an order parameter under dynamics with infinitely mismatched time scales can lead to scale invariance, or self-organized criticality [18, 28]. In principle, critical behavior is defined in the thermodynamic limit, but real systems are finite, measurements are noisy, and systematic corrections like finite-size effects are unavoidable [8, 14]. Yet when we are close enough to a fixed point, we expect that a few relevant scaling variables dominate and a simple description of the system emerges that is independent of many microscopic details [52]. Such a prediction suggests a simple scaling hypothesis that we test with armed conflict data.

2.2.1 Scaling framework for armed conflict

We investigate data collected in the Armed Conflict Location & Event Data Project (ACLED) that aggregates events reported by news media and regional contacts from 1997–2016 [45]. The part of the data set on Africa is notable for its extent—covering two decades, thousands of kilometers, and $> 10^5$ events. We analyze three kinds of events in the data set: Battles involving two or more armed groups (K = 42,738), Violence Against Civilians in which armed groups attack the population (K = 39,127), and Riots/Protests (K = 37,582). Each identified event has a geographic coordinate, date, and number of fatalities. Like the canonical avalanche picture for nonequilibrium critical phenomena, we call clusters of events "conflict avalanches." Although we consider all three conflict types, we focus on the Battles (see Appendix E for other event types).

We cluster events into conflict avalanches by setting a separation length b and separation time a such that events that are within the specified distance and

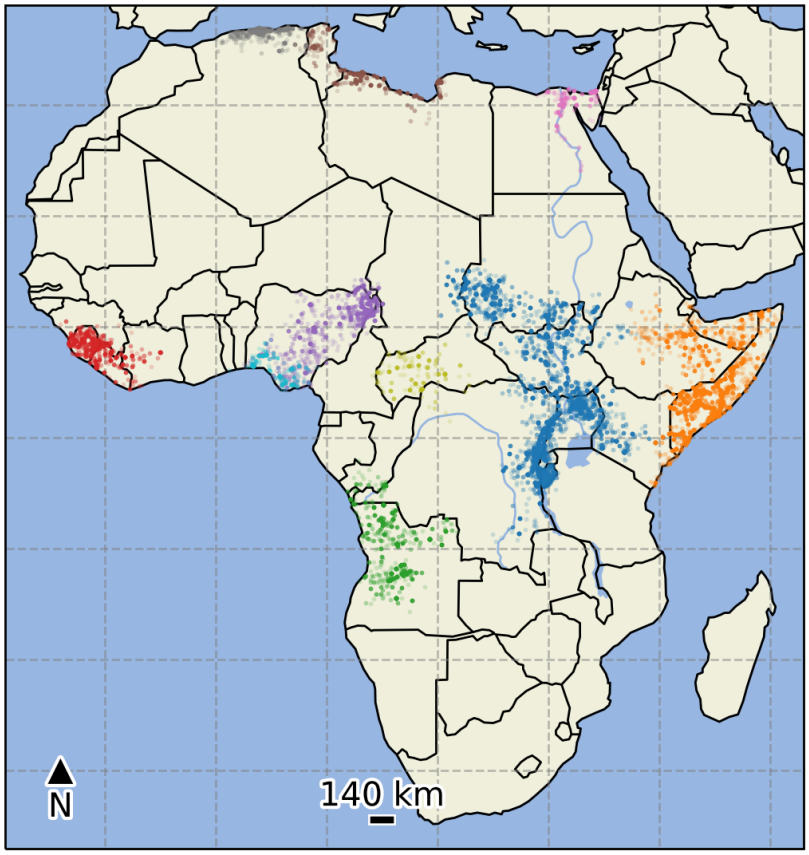


Figure 2.10: Battle conflict avalanches in Africa between 1997–2016 [45]. Spatial distribution of largest 10 conflict avalanches by size S for given separation scales $b = 140$ km and $a = 128$ days. Spatial density is highly non-uniform, largely confined to land, and typically denser near population centers.

time are grouped into the same avalanche (Appendix E.2), a procedure analogous to that done for neural avalanches [25, 44, 55]. As we vary these scales, the typical duration and geospatial extent of conflict avalanches change systematically, but for a large range of scales the observed statistics are remarkably consistent. For the following, we fix $b = 140$ km because it is sufficiently large that conflicts can percolate through a large network while sufficiently small that the system boundaries defined by geographic features (e.g. Sahara Desert, coastlines) do not significantly impact scaling (Appendix E.2). In Figure 2.10, we show the spatial distribution for the 10 largest avalanches by size for $b = 140$ km

and $a = 128$ days. A single example of a conflict avalanche spanning Libya and Tunisia lasting over 10^3 days with nearly 10^4 reported fatalities appears in Figure 2.11A along with its temporal profile. Thus, every conflict avalanche has a duration T in days, size measured by the number localized events or reports S, reported fatalities F, and geographic extent L in kilometers given by the maximally distant pair of events. This clustering operation, with only straightforward dependence on physical scales, defines a systematic way of constructing related sets of events, in contrast with notions of "battles" or "wars" which can depend on sociopolitical nuances.

As visible in Figures 2.10 and 2.11A, the spatial density of conflict is strongly nonuniform. Large conflicts tend to concentrate along high population areas: few occur in the Sahara Desert and only a handful are reported in the oceans. Conflict density also depends on other factors like the geography of country borders (e.g., Darfur). Not only do these geopolitical features impose boundaries on the propagation of conflict, but communication technology may render physical distance irrelevant for coordinated events. Considering the effects of strong spatial nonuniformity, pinning on geographic boundaries, and rapid long distance communication—analogous to effects that destroy scaling in physical systems—it would be surprising if the length scale L fit into a scaling description.

Since such effects are less relevant for time, we choose our scaling variable as the duration of avalanches T. Then, our scaling hypothesis predicts

$$S \sim T^{d_S/z}, \tag{2.9}$$

$$F \sim T^{d_F/z}, \tag{2.10}$$

and if including geographic extent L

$$L \sim T^{1/z} \tag{2.11}$$

with dynamical exponents d_S/z, d_F/z, and $1/z$ for size, fatalities, and geographic extent, respectively. The distributions of the scaling variables are likewise expected to scale simply

$$\begin{aligned} P(S) \sim S^{-\tau'}, \quad P(F) \sim F^{-\tau}, \\ P(L) \sim L^{-\nu}, \quad P(T) \sim T^{-\alpha}, \end{aligned} \tag{2.12}$$

The relations in Eqs 2.9–2.12 provide the basis for a scaling hypothesis of armed conflict that we test empirically.

We first measure the distributions of the scaling variables and find that their tails are well-described by power law distributions. Using a standard procedure for discovering the lower cutoff and exponent for the distribution like shown in Figure 2.11, we construct the distributions of the scaling variables (Figure 2.11B), and we find via a standard procedure that they are statistically indistinguishable from power laws [12]. The corresponding exponents appear in Figure 2.11C, where for the highlighted case of $a = 128$ days, we find $\tau' = 1.96 \pm 0.03$, $\tau = 1.65 \pm 0.08$, $\alpha = 2.44 \pm 0.13$, and $\nu = 2.78 \pm 0.21$.

If conflict avalanches grow in time, space, and magnitude in a self-similar manner, we expect that the dynamical exponents should be related to the power law exponents in a consistent way. To measure the dynamical exponents, we directly compare the scaling variables to determine $d_S/z = 2.0 \pm 0.3$, $d_F/z = 2.5 \pm 0.3$, and $1/z = 0.8 \pm 0.1$ as depicted in Figure 2.12. In a self-consistent framework, the

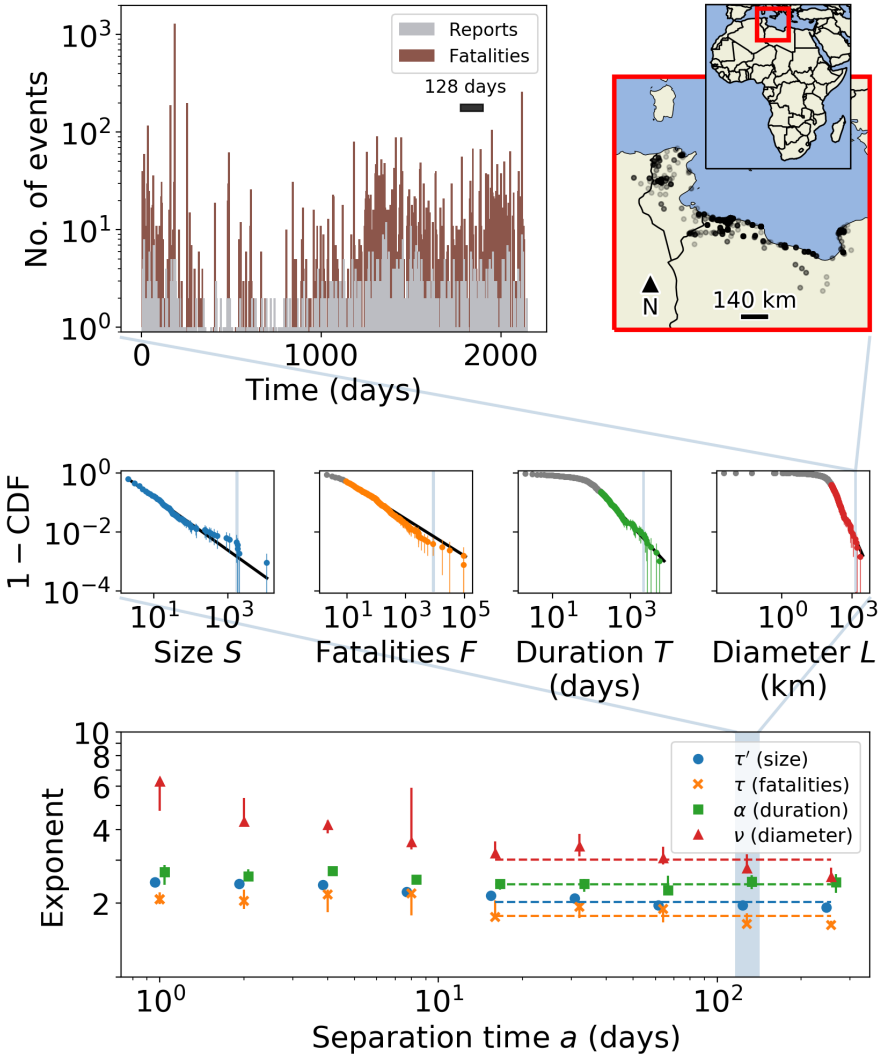


Figure 2.11: Overview of scaling for Battles. (A) A single conflict avalanche erupting across Tunisia and Libya from Feb. 2, 2011 til Dec. 27, 2016 with temporal profile on left and spatial distribution on right. This avalanche consists of $S = 1{,}717$ reports, $F = 8{,}569$ fatalities, lasts $T = 2{,}141$ days, and extends $L = 1{,}364$ km as highlighted in each graph in B in blue. (B) Complementary cumulative distribution functions for avalanche scaling variables given $a = 128$ days. Points below the lower cutoff in gray. Black lines indicate maximum likelihood fits, and error bars represent 90% bootstrapped confidence intervals. The data are statistically indistinguishable from power laws at the $p \geq 0.1$ significance level (Appendix E.3) [12]. (C) Exponents as a function of the separation time a. Dashed lines show the average exponent value for the last five points $16 \leq a \leq 256$ days.

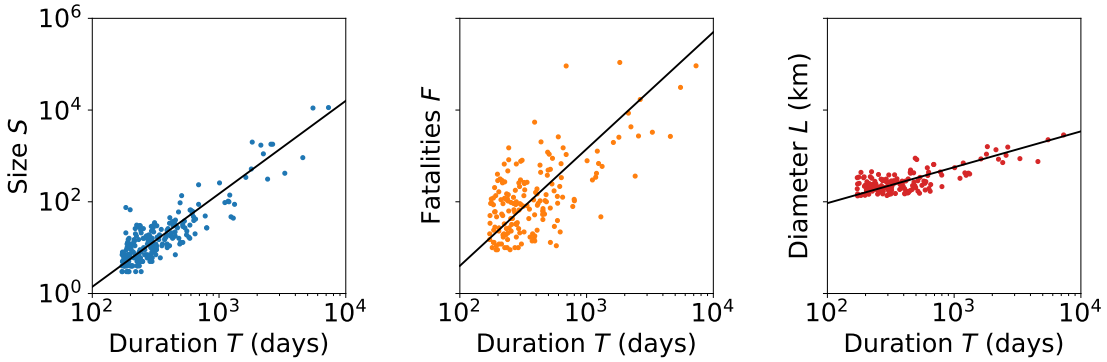


Figure 2.12: Dynamical scaling exponents measured from scaling relations are $d_S/z = 2.0 \pm 0.3$, $d_F/z = 2.5 \pm 0.3$, and $1/z = 0.8 \pm 0.1$.

measured exponents must satisfy the relations

$$\alpha - 1 = d_S(\tau' - 1)/z = d_F(\tau - 1)/z = (\nu - 1)/z. \tag{2.13}$$

These relations are satisfied within 90% bootstrap error intervals. Thus, the various features of conflict including, perhaps surprisingly, L are unified in a self-consistent fashion given a simple scaling description.

Self-similarity also predicts that the average evolution of each scaling variable within an avalanche approaches a universal profile at large scales. We test idea using the normalized trajectories of size $\langle s(t/T)/S \rangle$, fatalities $\langle f(t/T)/F \rangle$, and geographic extent $\langle l(t/T)/L \rangle$ that give the cumulative fraction of total events or extent by rescaled time t/T. For sizes s, at least one event occurs at $t = 0$ and $t = T$ by construction, so we must account for a $1/S$ "lattice" bias to obtain a collapse and a related bias for fatalities (Appendix E.5). Upon averaging, we find across avalanches with duration $T \geq a$ that the cumulative profiles largely overlap. Though we find notable variation between short and long conflict avalanches at short times $t/T < 10^{-1}$, indicative of a relevant scaling variable that depends on conflict size (insets in Figure 2.13), the profiles largely converge for $t/T \geq 10^{-1}$. This overlap between the temporal profiles indicates that the dy-

namics of conflict avalanches may be dominated by a scale-invariant process as is consistent with a scaling framework.

Notably, the statistical structure encoded in the exponent relations in Eq 2.13 and temporal profiles is largely preserved as we change the separation time a. In Figure 2.11, we show that the exponents stay close to their values in the highlighted example over an order of magnitude of $16 \leq a \leq 256$ days, and in Figure 2.13 the average temporal profiles hardly change across the matching range of a. In physical systems near a critical point, symmetries under rescaling are expected. In our case, increasing a does not exactly correspond to rescaling time but rather groups together events that are increasingly further apart into the same avalanche. Yet remarkably, we find that doubling a is statistically analogous to scaling T in that it largely preserves the exponents and temporal profiles across timescales from weeks to years, a result that reflects self-similarity in the timing of conflict events [43].

The temporal profiles hint at the underlying dynamics generating conflict avalanches. For comparison, we show profiles of canonical systems with self-similar avalanches like Barkhausen noise and an example of a neural culture. These tend to accelerate in the middle whereas average size and fatality profiles for conflict avalanches tend to evolve at a linear pace. Flat profiles can indicate dissipative effects that suppress large events as with demagnetizing fields in Barkhausen noise [41]. Yet, flattening is also a feature of both subcritical and supercritical cascades that spontaneously end—though such profiles will fail to collapse [27, 29]. Thus, the mapping between dynamics and profile is many-to-one, but we can rule out analogues of properties that,

	Size τ'	Fatalities τ	Diameter ν	Duration α	S vs. T d_S/z	F vs. T d_F/z	L vs. T $1/z$
Battles	1.96 1.91, 2.02	1.65 1.61, 1.87	2.78 2.60, 3.29	2.44 2.26, 2.67	2.0 1.7, 2.5	2.5 2.1, 3.2	0.78 0.64, 0.96
Forest fires 2D		2.14 2.11, 2.17	1.28 1.19, 1.37	1.27 1.20, 1.34		1.89 1.86, 1.92	0.96 0.94, 0.98
Percolation growth 2D		2.05	2.87	2.65		1.57	0.88
Barkhausen 2D*		2	2.09 1.91, 2.27	1.87 1.81, 1.93		1.55 1.51, 1.59	0.80 0.69, 0.91
ARW 2D		1.31		1.55			
Neural		2.10 2.09, 2.11		2.86 2.85, 2.87		1.85 1.82, 1.89	
Wars		1.53 1.46, 1.60					

Table 2.2: Scaling exponents for Battles conflict avalanches with those for physical, biological, and social systems. Critical exponents are shown for 2D forest fires [10], percolation growth (Appendix E.7), Barkhausen noise (random field Ising model) [42], and activated random walkers (Appendix E.7) [18]. For comparison, we show experimental neural avalanches [44] and the fatality exponent for interstate wars—defined sociopolitically in contrast to our conflict avalanches—from 1823–2003 [11]. Either exponent d_S or d_F could correspond to the conventional choice of exponents $1/\sigma\nu$ for fractal dimension. Where the exponent error intervals (shown in gray) overlap with those of Battles, we color the box light blue. Bigger table with more exponents shown in Appendix Table 4.1.

for example, generate asymmetric profiles such as eddy currents in magnetic materials [41], certain networks like in disassociated neural cultures [25], or variations in birth-death processes [27]. We also find that spatial extent grows in a strongly nonlinear and asymmetric fashion as shown in Figure 2.13C. This profile is closely described by the average linear extent of a convex hull of planar Brownian walkers [32, 46], perhaps related to properties of generalized diffusion models used to describe other conflict data sets [58]. More generally, these profiles are compatible with Markovian cascades on networks indicating that such dynamics may come to dominate in long conflict avalanches.

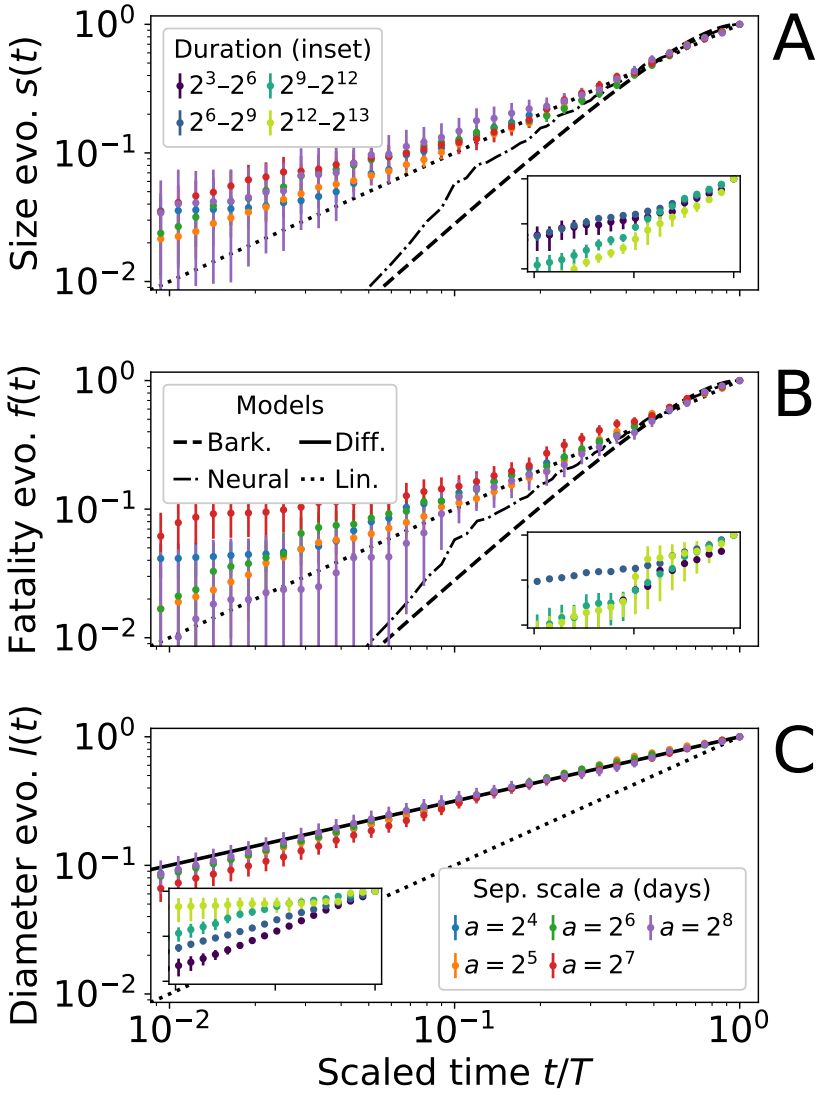


Figure 2.13: Temporal evolution measured by cumulative fraction of (A) sizes as number of reports, (B) fatalities, and (C) geographic extent along scaled time t/T. Profiles are averaged over all avalanches with duration $T \geq a$ days, the separation time scale. We compare conflicts with the normalized trajectory for experimental neural avalanches (dashed-dotted line, $K > 10^3$ [55]), Barkhausen noise $\int_0^{t/T} V(t')\,dt' = 3(t/T)^2 - 2(t/T)^3$ (dashed line [41]), diffusive growth $l(t/T) \propto (t/T)^{1/2}$ (solid line [46]), and linear growth $s(t/T) \propto t/T$ (dotted line). (insets) For separation time $a = 128$ days, we show that average profiles show substantial variation when binned by conflict duration. Error bars represent standard errors of the means.

Beyond temporal profiles, the measured exponents indicate how the spread of armed conflict is comparable to physical, biological, and social systems in Table 2.2. In agreement with our observations with temporal profiles, armed conflict shows differences with the cascade processes listed. Of particular note is the self-organized critical forest fire (FF) model that shows strong disagreement with duration exponent α. This model is oft-cited in the context of human conflict [9, 50]. In comparison, our measured exponents are similar to those for percolation growth, where time is measured by each shell of newly occupied sites for a cluster nucleated at the origin. Such similarity hints that the spread of armed conflict is comparable to growth processes on networks at the percolation threshold as appears to be the case with neural avalanches in zebrafish [44]. We note that the scaling of S vs. T is nearly quadratic for most of the listed processes, reflecting the fact that events happen faster in larger avalanches, one way of distinguishing small conflicts from larger ones early on. Furthermore, Battles, unlike the physical examples that are confined to the "lattice sites," exceed the spatial dimensions of the surface of the Earth: $d_F \gtrsim d_S \gtrsim 2$. This means that multiplicity of the events at conflict sites is a crucial feature of conflict avalanche dynamics similar to the recurrence of a neural avalanche on the same neuron. Thus, in this way we can use scaling exponents to systematically compare armed conflict with other physical processes relying on the formalism of universality classes from physics.

2.2.2 Discussion

The emergence of these large-scale symmetries is extraordinary. Such remarkable regularity presents an opportunity for prediction [53]. In particular, knowledge of the temporal profiles suggests one way of extrapolating from the begin-

ning of an ongoing conflict the potential human cost of the rest of the conflict before it ends. Scaling relations could be used to estimate missing data points like fatalities (which are especially difficult to measure), to detect anomalous statistics, or to help assess risk for nearby regions by showing how geographic extent scales with duration. These statistics are extracted from clusters of conflict events generated from simple physical scales, providing a well-defined, quantitative, and straightforwardly measured procedure as a complement to sociopolitical definitions of wars. Taken together, our results reveal a unified framework for conflict growth in which physical space and time scales constrain a social phenomenon. Universality and scaling laws have been found in a variety of social systems [4, 23], suggesting self-similarity and the renormalization group as means to understanding how physical constraints translate into emergent patterns at large scales. In this wider context, our findings hint at the intriguing possibility that emergent regularities reflect underlying physical principles that shape the evolution of armed conflict.

Chapter 2 references

[1] M J Allman, S Teki, and T D Griffiths. "Properties of the Internal Clock: First-and Second-Order Principles of Subjective Time". In: *Annu. Rev. Psychol.* (2014).

[2] R Barakat. "Sums of Independent Lognormally Distributed Random Variables". In: *J. Opt. Soc. Am., JOSA* (1976).

[3] R. E. Barlow and F. Proschan. *Mathematical Theory of Reliability*. Wiley, 1965.

[4] L. M. A. Bettencourt et al. "Growth, Innovation, Scaling, and the Pace of Life in Cities". en. In: *Proceedings of the National Academy of Sciences* 104.17 (Apr. 2007), pp. 7301–7306. ISSN: 0027-8424, 1091-6490. DOI: 10.1073/pnas.0610172104.

[5] D.T. Bishop and C. Cannings. "A Generalized War of Attrition". In: *Journal of Theoretical Biology* 70 (1978), pp. 85–124.

[6] A Bissonnette, H de Vries, and C P Van Schaik. "Coalitions in Male Barbary Macaques, Macaca Sylvanus: Strength, Success and Rules of Thumb". In: *Animal Behaviour* (2009).

[7] A. Burt and R. Trivers. *Genes in Conflict*. Cambridge: Harvard University Press, 2008.

[8] John L. Cardy, ed. *Finite Size Scaling*. Vol. 2. New York: North-Holland, 1988.

[9] Lars-Erik Cederman. "Modeling the Size of Wars: From Billiard Balls to Sandpiles". en. In: *American Political Science Review* 97.01 (Feb. 2003), pp. 135–150. ISSN: 0003-0554, 1537-5943. DOI: 10.1017/S0003055403000571.

[10] S. Clar, B. Drossel, and F. Schwabl. "Scaling Laws and Simulation Results for the Self-Organized Critical Forest-Fire Model". en. In: *Physical Review E* 50.2 (Aug. 1994), pp. 1009–1018. ISSN: 1063-651X, 1095-3787. DOI: 10.1103/PhysRevE.50.1009.

[11] Aaron Clauset. "Trends and Fluctuations in the Severity of Interstate Wars". en. In: *Science Advances* 4.2 (Feb. 2018), eaao3580. ISSN: 2375-2548. DOI: 10.1126/sciadv.aao3580.

[12] Aaron Clauset, Cosma Rohilla Shalizi, and M. E. J. Newman. "Power-Law Distributions in Empirical Data". en. In: *SIAM Review* 51.4 (Nov. 2009), pp. 661–703. ISSN: 0036-1445, 1095-7200. DOI: 10.1137/070710111.

[13] B. C. Daniels, D. C. Krakauer, and J. C. Flack. "Sparse Code of Conflict in a Primate Society". en. In: *Proceedings of the National Academy of Sciences* 109.35 (Aug. 2012), pp. 14259–14264. ISSN: 0027-8424, 1091-6490. DOI: 10.1073/pnas.1203021109.

[14] Bryan C. Daniels, David C. Krakauer, and Jessica C. Flack. "Control of Finite Critical Behaviour in a Small-Scale Social System". en. In: *Nature Communications* 8 (Feb. 2017), p. 14301. ISSN: 2041-1723. DOI: 10.1038/ncomms14301.

[15] Bryan C Daniels et al. "Quantifying Collectivity". In: *Current opinion in neurobiology* 37 (2016), pp. 106–113.

[16] Simon DeDeo, D Krakauer, and Jessica C Flack. "Evidence of Strategic Periodicities in Collective Conflict Dynamics". In: *Journal of The Royal Society Interface* 8.62 (July 2011), pp. 1260–1273.

[17] Simon DeDeo, David C Krakauer, and Jessica C Flack. "Inductive Game Theory and the Dynamics of Animal Conflict". In: *PLoS Comput Biol* q-bio.PE.5 (June 2010), e1000782.

[18] Ronald Dickman et al. “Paths to Self-Organized Criticality”. en. In: *Brazilian Journal of Physics* 30.1 (Mar. 2000), pp. 27–41. ISSN: 0103-9733. DOI: 10.1590/S0103-97332000000100004.

[19] P.S. Dodds and D.J. Watts. “A Generalized Model of Social and Biological Contagion”. In: *Journal of Theoretical Biology* 232 (2005), pp. 587–604.

[20] J. C. Flack. “Multiple Time-Scales and the Developmental Dynamics of Social Systems”. In: *Philosophical Transactions of the Royal Society B: Biological Sciences* 367.1597 (July 2012), pp. 1802–1810. DOI: 10.1098/rstb.2011.0214.

[21] Jessica C. Flack, Frans B M de Waal, and David C. Krakauer. “Social Structure, Robustness, and Policing Cost in a Cognitively Sophisticated Species.” In: *The American naturalist* 165.5 (May 2005), E126–39.

[22] Jessica C. Flack et al. “Policing Stabilizes Construction of Social Niches in Primates”. In: *Nature* 439.7075 (Jan. 2006), pp. 426–429.

[23] Santo Fortunato and Claudio Castellano. “Scaling and Universality in Proportional Elections”. en. In: *Physical Review Letters* 99.13 (Sept. 2007), p. 138701. ISSN: 0031-9007, 1079-7114. DOI: 10.1103/PhysRevLett.99.138701.

[24] S.A. Frank. “Repression of Competition and the Evolution of Cooperation”. In: *Evolution* 57 (2003), pp. 693–705.

[25] Nir Friedman et al. “Universal Critical Dynamics in High Resolution Neuronal Avalanche Data”. en. In: *Physical Review Letters* 108.20 (May 2012), p. 208102. ISSN: 0031-9007, 1079-7114. DOI: 10.1103/PhysRevLett.108.208102.

[26] John Gibbon et al. “Toward a Neurobiology of Temporal Cognition: Advances and Challenges”. In: *Current Opinion in Neurobiology* 7.2 (Apr. 1997), pp. 170–184.

[27] James P. Gleeson and Rick Durrett. “Temporal Profiles of Avalanches on Networks”. en. In: *Nature Communications* 8.1 (Dec. 2017), p. 1227. ISSN: 2041-1723. DOI: 10.1038/s41467-017-01212-0.

[28] G. Grinstein. “Generic Scale Invariance and Self-Organized Criticality”. en. In: *Scale Invariance, Interfaces, and Non-Equilibrium Dynamics*. Ed. by Alan McKane et al. Vol. 344. Boston, MA: Springer US, 1995, pp. 261–293. DOI: 10.1007/978-1-4899-1421-7_11.

[29] Jason Hindes and Ira B. Schwartz. “Epidemic Extinction and Control in Heterogeneous Networks”. en. In: *Physical Review Letters* 117.2 (July 2016), p. 028302. ISSN: 0031-9007, 1079-7114. DOI: 10.1103/PhysRevLett.117.028302.

[30] Morten Jerven. *Poor Numbers: How We Are Misled By African Development Statistics And What To Do About It*. Ed. by Peter J. Katzenstein. Ithaca: Cornell University Press, 2013.

[31] M. J. Keeling and B.T. Grenfell. “Disease Extinction and Community Size: Modeling the Persistence of Measles”. In: *Science* 275 (1997), pp. 65–67.

[32] Mark Kot, Mark A. Lewis, and P. van den Driessche. “Dispersal Data and the Spread of Invading Organisms”. en. In: *Ecology* 77.7 (Oct. 1996), pp. 2027–2042. ISSN: 00129658. DOI: 10.2307/2265698.

[33] D.C. Krakauer and A Mira. “Mitochondria and Germ-Cell Death”. In: *Nature* 400 (1999), pp. 125–126.

[34] D.C. Krakauer, K. Page, and J.C.F Flack. “The Immuno-Dynamics of Conflict Intervention in Social Systems”. In: *PLoS One* 6 (2011), e22709.

[35] L. Lafore. *The Long Fuse: An Interpretation of the Origins of World War I, Second Edition*. Waveland Press, 1997. ISBN: 978-1-4786-0933-9.

[36] Edward D. Lee et al. "Collective Memory in Primate Conflict Implied by Temporal Scaling Collapse". en. In: *Journal of The Royal Society Interface* 14.134 (Sept. 2017), p. 20170223. ISSN: 1742-5689, 1742-5662. DOI: 10.1098/rsif.2017.0223.

[37] Edward D. Lee et al. "Emergent Regularities and Scaling in Armed Conflict Data". en. In: *arXiv:1903.07762 [cond-mat, physics:nlin, physics:physics, q-bio]* (Mar. 2019). arXiv: 1903.07762 [cond-mat, physics:nlin, physics:physics, q-bio].

[38] Edward Lee et al. "Capturing Collective Conflict Dynamics with Sparse Social Circuits". In: *arXiv:1406.7720 [physics]* (June 2014). arXiv: 1406.7720 [physics].

[39] P A Lewis and R C Miall. "The Precision of Temporal Judgement: Milliseconds, Many Minutes, and Beyond". In: *Philosophical Transactions of the Royal Society of London B: Biological Sciences* 364.1525 (July 2009), pp. 1897–1905.

[40] J. Maynard-Smith. *Evolution and the Theory of Games*. Cambridge University Press, 1982.

[41] Stefanos Papanikolaou et al. "Universality beyond Power Laws and the Average Avalanche Shape". en. In: *Nature Physics* 7.4 (Apr. 2011), pp. 316–320. ISSN: 1745-2473, 1745-2481. DOI: 10.1038/nphys1884.

[42] Olga Perkovic, Karin A. Dahmen, and James P. Sethna. "Disorder-Induced Critical Phenomena in Hysteresis: A Numerical Scaling Analysis". In: *arXiv:cond-mat/9609072* (Sept. 1996). arXiv: cond-mat/9609072.

[43] S. Picoli et al. “Universal Bursty Behaviour in Human Violent Conflicts”. en. In: *Scientific Reports* 4.1 (May 2015), p. 4773. ISSN: 2045-2322. DOI: 10.1038/srep04773.

[44] Adrián Ponce-Alvarez et al. “Whole-Brain Neuronal Activity Displays Crackling Noise Dynamics”. en. In: *Neuron* 100.6 (Dec. 2018), 1446–1459.e6. ISSN: 08966273. DOI: 10.1016/j.neuron.2018.10.045.

[45] Clionadh Raleigh et al. “Introducing ACLED: An Armed Conflict Location and Event Dataset: Special Data Feature”. en. In: *Journal of Peace Research* 47.5 (Sept. 2010), pp. 651–660. ISSN: 0022-3433, 1460-3578. DOI: 10.1177/0022343310378914.

[46] Julien Randon-Furling, Satya N. Majumdar, and Alain Comtet. “Convex Hull of N Planar Brownian Motions: Exact Results and an Application to Ecology”. en. In: *Physical Review Letters* 103.14 (Sept. 2009), p. 140602. ISSN: 0031-9007, 1079-7114. DOI: 10.1103/PhysRevLett.103.140602.

[47] L.F. Richardson. *Statistics of Deadly Quarrels*. Boxwood Press, 1960.

[48] Lewis F. Richardson. “Frequency of Occurrence of Wars and Other Fatal Quarrels”. In: *Nature* 148 (Nov. 1941), p. 598.

[49] Lewis F. Richardson. “Variation of the Frequency of Fatal Quarrels With Magnitude”. In: *Journal of the American Statistical Association* 43.244 (1948), pp. 523–546. ISSN: 0162-1459. DOI: 10.2307/2280704.

[50] D. C. Roberts and D. L. Turcotte. “Fractality and Self-Organized Criticality of Wars”. In: *Fractals* 6.4 (1998), pp. 351–357.

[51] James P Sethna. *Entropy, Order Parameters, and Complexity*. en. Vol. 14. Oxford Master Series Statistical, Computational, and Theoretical Physics. Oxford University Press, 2006.

[52] James P. Sethna, Karin A. Dahmen, and Christopher R. Myers. "Crackling Noise". En. In: *Nature* 410.6825 (Mar. 2001), p. 242. ISSN: 1476-4687. DOI: 10.1038/35065675.

[53] Michael Spagat, Neil F. Johnson, and Stijn van Weezel. "Fundamental Patterns and Predictions of Event Size Distributions in Modern Wars and Terrorist Campaigns". en. In: *PLoS ONE* 13.10 (Oct. 2018). Ed. by Kristian Skrede Gleditsch, e0204639. ISSN: 1932-6203. DOI: 10.1371/journal.pone.0204639.

[54] S.C. Stearns. "The Selection-Arena Hypothesis". In: *Experientia Suppl.* 55 (1987), pp. 337–49.

[55] Nicholas M. Timme et al. "Criticality Maximizes Complexity in Neural Tissue". en. In: *Frontiers in Physiology* 7 (Sept. 2016). ISSN: 1664-042X. DOI: 10.3389/fphys.2016.00425.

[56] C P Van Schaik, S A Pandit, and E R Vogel. "Toward a General Model for Male-Male Coalitions in Primate Groups". In: *Cooperation in Primates and Humans*. Springer Berlin Heidelberg, 2006, pp. 151–171.

[57] Carl von Clausewitz. *On War*. Ed. by Michael Howard and Peter Paret. Princeton University Press, 1984.

[58] A. Zammit-Mangion et al. "Point Process Modelling of the Afghan War Diary". en. In: *Proceedings of the National Academy of Sciences* 109.31 (July 2012), pp. 12414–12419. ISSN: 0027-8424, 1091-6490. DOI: 10.1073/pnas.1203177109.

CHAPTER 3

COORDINATION OF HUMAN MOTION

How do dancing partners synchronize their motions? How does a jazz ensemble keep time? How do rowers on a crew team push and pull simultaneously? Such coordination involves the interplay between physical coupling with the environment, through perceptual information [49, 50], and internal dynamical models of self and environment. In previous chapters, we investigated the collective behavior of many interacting individuals, but in this chapter we focus down to a single individual in order to measure how a person responds to changes in the sensory environment.

Many physical models of interaction between agents synchronizing their behavior involve generic parameters characterizing the magnitude and perhaps the variation in interaction strength between agents. A classic example is the Kuramoto model, where the phases of many oscillators are coupled [1, 25]. The idea, of course, is that the simplified representation yields insight, whereas a complicated model would be intractable and perhaps too complex for distilling useful principles of behavior. Human interactions, however, display remarkable diversity depending on the environment, the cognitive state of actors, physical capabilities, etc. One surprising discovery in the study of interpersonal coordination is the ability of strangers to synchronize spontaneous motions precisely without verbal communication, exceeding what might be expected from limits of behavioral response time [30]. Accurate synchronization of behavior depends on previous experience [47] as well as learning and adaptation [45]. Given observations like these, how we should appropriately reduce the variety of interper-

sonal interactions into generic parameters when building models of interacting human systems?

Answering this question even in the domain of motion has been difficult because experiments that capture precisely the entire human posture—needless to say the cognitive state—has been difficult. As a result, many experiments are based on clever but necessarily reduced experimental paradigms exploring a limited range of environments. In the last few years, however, inexpensive commercially technologies have become available, making it possible to determine the motion of the human body with relatively good accuracy, temporal precision, and ease. Furthermore, the advent of virtual reality technology opens the door to directly controlling the visual and auditory fields in a much wider variety of ways [15, 35]. We combine multiple commercially-available technologies to characterize how subjects response to changes in their sensory environment as they perform a mirroring task [26].

This work presents an initial foray into this area of interpersonal coordination including, in large part, the development of the experimental apparatus. With such a tool, we propose that the complexity of sensory input data, the entire visual and auditory fields, along with the that of the high-dimensional motion-capture data naturally leads to a statistical framework for studying the information that is encoded between the environmental and translated into behavior [2, 40, 42]. Though this approach is evocative of the study of model biological systems, such approaches remain yet undeveloped in the domain of human behavior.

3.1 Audio cues enhance mirroring of arm motion when visual cues are scarce

Successful coordination of human motion in a group is crucial for many tasks including dance, team sports, or music ensembles [11, 23, 31, 43]. In all these cases, it is essential that the individual extract information from the local environment [5, 6] to maintain coordination with others. When input to a sensory channel is disrupted systematically, however, how do individuals compensate for such disruption? In the context of sensorimotor integration for reaching tasks, this question has been well-studied [14, 37, 49]. Here, we study the transition from coordinated to uncoordinated behavior using an experimental apparatus that manipulates the visual and auditory fields, measures the dynamic motions of individuals, and quickly maps out performance across large regions of parameter space. To determine the relationship between available visual information and a subject's ability to mirror accurately, we asked 35 subjects to mirror the hand motions of a pre-recorded avatar while we changed the rate with which and duration during which the avatar was visible. Next, we measured changes in performance when the subjects were given supplementary audio cues that mapped velocity of motion to frequency, practice training rounds, or both training and audio cues. Using these data, we find that audio enhances performance at fast time scales while the combination of both audio and training affects the dynamics of coordination performance in a characteristic way that may be detectable in other experiments.

Pitch-based auditory cues provide an informative, intuitive, and commonly used approach for representing kinematic and kinetic measurements in human

motion [9, 46, 52]. When used as feedback, audio cues can enhance performance at motor tasks across a variety of contexts including learning a cyclic motion [10, 13, 52] and interpersonal coordination [16, 23, 38]—more generally perceptual coupling including other sensory modalities like vision and touch have been shown to enhance interpersonal coordination (see reference [39] for a review). In the case where subjects are trying to learn a new motion, evidence suggests that feedback based on the target motion is more effective than that based on the subject's own motion [10, 16]. Along these lines, we represent the target hand motion of the avatar that the subject is mirroring using a simple proportional pitch mapping based on speed. Given that these auditory cues provide complementary auditory information along with a visual of the avatar, we expect that performance at mirroring the avatar will be enhanced.

Beyond perceptual coupling, higher-level planning processes may play a role in learning how to mimic others' motions [44], an aspect that we study by training some of the participants in practice trials. In various studies of interpersonal coordination, there is evidence that anticipatory motor activation might help individuals respond to the motion of others [21, 24]. One experiment showed that even imagining the motion of another subject prior to motion helped to synchronize behavior [43]. Here, we explore how the provision of auditory cues might compare to the benefits of a training round where individuals have a chance to practice mirroring an avatar. We run two variations of the conditions where in one participants do not have the opportunity to practice the task and in the other they do. This variation allows us to measure interplay between audio cues and practice with the task. In some experiments exploring the effect of audio cues, similar kinds of practice rounds precede measurement [10], whereas

in others subjects only witness the task before immediately proceeding to performance [13]. We would expect that performance at the mirroring task would improve when subjects receive either training or audio cues in the absence of reliable visual cues, a prediction that would be consistent with other results in the literature [10, 13, 30].

3.1.1 Experimental setup

In our experiments, subjects wore virtual reality goggles and a motion capture suit, stood face-to-face with an avatar that played a pre-recorded sequence of aperiodic motions generated by an experimenter, and were instructed to mirror the motion of the avatar's hand as shown in Figure 3.1A. Each experimental sequence consisted of 16 sequential 30 s trials with varying difficulty. To control the difficulty, we took windows of duration $1/2\,\mathrm{s} \leq \tau \leq 2\,\mathrm{s}$ and only showed the avatar for a contiguous visible fraction $0.1 \leq f \leq 1$ of the window (f is analogous to the duty cycle). In the first and last trials, the avatar was visible at all times, so $f = 1$. After a 16 trial sequence with a randomly chosen hand, the subjects repeated another sequence with the other hand for a different set of motions.

We constructed the experimental system by combining a variety of open source and commercially available software and hardware components coordinated by a custom-built program to coordinate the experiment and run the analysis. We show in Figure 3.2 the architecture of the experimental apparatus, where the components are indicated by boxes and direction of communication between the components indicated by the arrows (e.g., to coordinate a change in the visual stimulus from the Python backend, a command would have to traverse

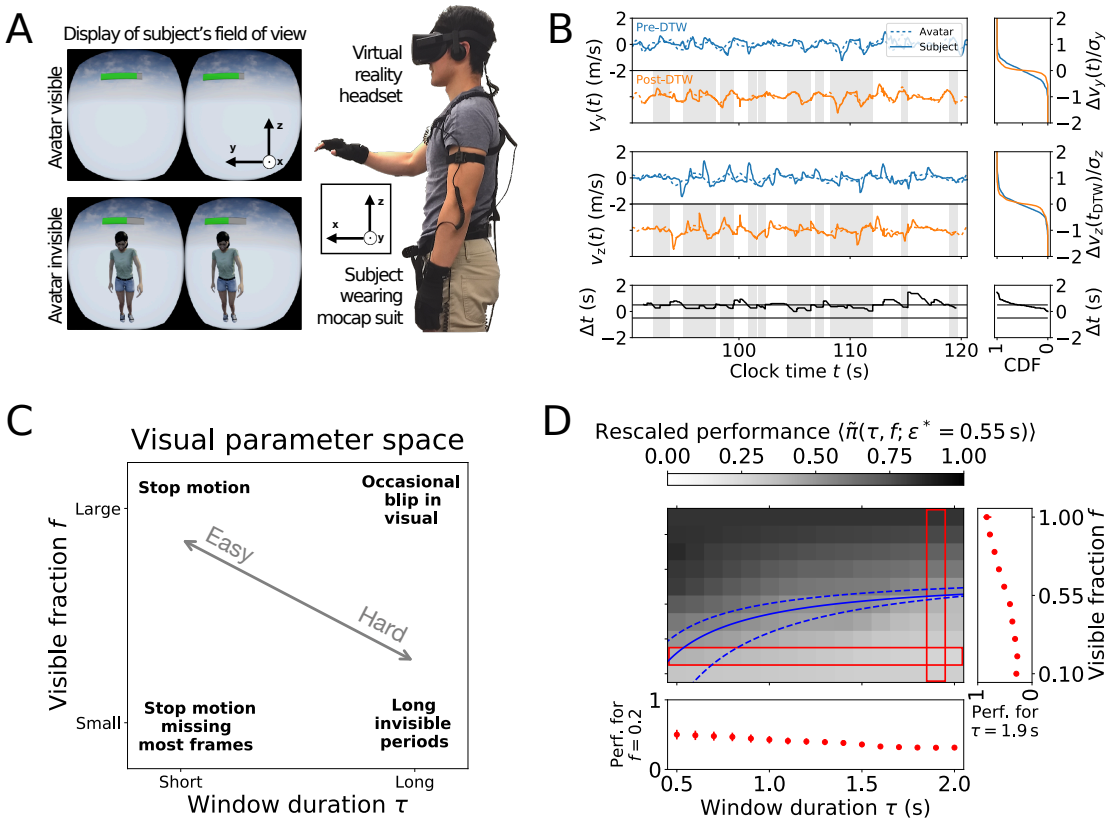


Figure 3.1: (A) Experimental setup showing a subject wearing virtual reality goggles and the motion capture suit along with his field of view when the avatar is invisible and visible. The green bar indicates current performance. (B) Dynamic time warping (DTW) aligns the measured velocities (blue) along the y and z axes. After DTW (orange), we identify runs of successful tracking (gray), and the fraction of the trial that these regions span is the estimated performance $\hat{\pi}$. DTW, as expected, reduces velocity error (normalized by standard deviation) as shown on right and returns time delays with distribution on bottom right. (C) Parameter space diagram. The four corners represent the extremes of the parameter space. We label where the visual representation of the avatar blinks on and off quickly as stop motion animation and f determines the fraction of time, akin to the duty cycle, during which the avatar is visible. (D) Rescaled performance landscape $\langle \tilde{\pi}(\tau, f) \rangle_{\epsilon^*=0.55\,\mathrm{s}}$ aggregated across all subjects in the Train+Audio condition (M = 15). One-dimensional cuts, outlined by red rectangles, are shown on the sides with predicted uncertainties. Error bars are one standard over the rescaled landscapes. Solid blue line traces the relation in Eq 3.1 fit to the level curve in performance given by $\langle \tilde{\pi} \rangle = 1/2$. Dashed blue lines indicate fits to rescaled landscapes one standard deviation above and below the shown mean landscape.

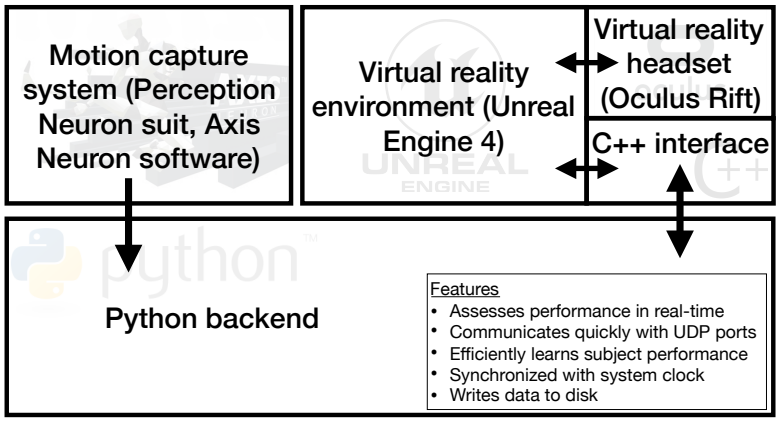


Figure 3.2: Architecture of the experimental apparatus. The boxes represent separate software programs and the arrows represent directions of information transfer for coordinating the programs. The Python backend collects data from the Perception Neuron (PN) motion capture system and compares it with prerecorded motion of the avatar that is displayed in the Oculus Rift virtual reality headset in the environment designed using Unreal Engine 4 (UE4). Subject performance is then assessed and the result used to train the learning algorithm that determines the next set of visibility parameters. Those parameters are communicated back to UE4 for the next 30 s trial.

both the C++ interface and the Unreal Engine 4 program). We used the Perception Neuron motion capture suit developed by Noitom to record subject motion and we discuss the limitations of this device in further detail in Appendix F.2. For generating the avatar stimulus, we built a 3-dimensional environment with Unreal Engine 4, a game development software suite, that the subject experienced through the Oculus Rift Virtual Reality headset. In order to coordinate all these components, we wrote a C++ interface to communicate with the central control program written in Python that kept track of the state of the other interfaces. By coordinating multiple sensory components with a customizable interface in Python, this architecture presents a flexible, modularized research tool that can respond in realtime to the behavior of the subject.

All subjects were informed about the purpose and goal of the study at the be-

ginning of the experiment and gave consent. After a preliminary survey about experience in sports or performing arts and questions about any conditions that would exclude them from the study (including vision, hearing, and arm motion problems and history of poor experience with virtual reality headsets), they were shown how to use the motion capture suit and virtual reality headset comfortably. The subject was familiarized with the mirror game outside of the virtual reality environment through two quick practice rounds (one hand at a time) with the researcher. Subjects were then instructed to "mirror [simultaneously] the motion, or velocity, of the avatar" where the word "simultaneously" was included in the training conditions because it was unclear if all subjects understood what was implied by mirroring in the untrained condition. When audio cues were used, they were also told, "Try to use the sound to predict the motion of the avatar's hand." Immediately previous to the start of the mirroring task, they were reminded visually by a floating script to "Mirror the hand." Periodically throughout the trial, the comfort of subjects in the virtual environment was assessed verbally. At the end of the experiment, all subjects filled out a post-experiment survey to assess the comfort of the suit and virtual headset, importance of fatigue, clarity of instructions, and to check if they had been following instructions.

3.1.2 Experimental results

We assess how well subjects mirrored the motion of the avatar by comparing the two-dimensional velocity trajectory of the subject $\vec{v}_{\mathrm{s}}(t)$ with the avatar's $\vec{v}_{\mathrm{a}}(t)$. We show an example for a single 30 s trial in Figure 3.1B, where we measure the velocity of the subject's hand (solid blue line) in the mirror plane separating the subject from the avatar. The plane corresponds to the y and z axes as defined

in Figure 3.1A.[1] By inspecting the coarse features of the trajectories, we observe that the subject captures much of the lower-frequency motions of the avatar but only after a varying temporal delay. To account for these delays, we use a standard algorithm for aligning two trajectories with local temporal modulation called dynamic time warping (DTW) [18, 29, 36]. We regularize the alignment problem so that solutions where the subject is more than $1/2\,\mathrm{s}$ ahead or $3/2\,\mathrm{s}$ behind are penalized to avoid pathologies that can arise from periodic motion (see Appendix F.2). We show an example of the time warped velocity trajectories (green solid and dashed lines) in Figure 3.1B. After DTW, the velocity difference normalized by the typical size of the velocity fluctuations of the avatar, σ_{a}, is narrower than the unaligned distribution. This narrowing indicates that accounting for temporal delays ϵ (black line in Figure 3.1B) substantially improves feature matching between the curves. Since close mirroring corresponds to minimal delay, we use the distribution of delays found from aligning the curves as a measure of how well subjects mirrored the avatar; results are similar if we also consider the direction of the velocity vector (see Appendix F.4).

After alignment with DTW, we summarize mirroring performance with the estimated fraction of time that a subject is able to stay within a time threshold ϵ^* given the window duration τ and visible fraction f, $\hat{\pi}(\tau, f; \epsilon^*)$, which can only vary from 0 to 1. When the subject is consistently within a time delay of ϵ^* (as indicated by the shaded regions in Figure 3.1B), the estimated performance measure $\hat{\pi} \approx 1$. With a short threshold ϵ^*, high performing subjects must mirror the avatar very closely with few deviations in both timing and velocity—we find that dissimilar trajectories lead to strong temporal variability with DTW. On the

[1] We do not consider the x axis which points from the subject to the avatar. This axis is particularly problematic for the motion capture system that we used and we found that timing errors could be significant (see Appendix F.2).

other hand with large ϵ^*, slower reaction times and bigger corrections will not affect the value of the performance. Thus, we vary ϵ^* to probe variation in how closely subjects mirror the avatar.

Given a particular value of the time threshold ϵ^*, we use Gaussian process regression to model a single subject's performance landscape using the 16 trials as training data to interpolate the unmeasured points [4, 33]. These 16 data points represent a sparse sample of 160 discretized grid points. During an experiment, we chose these points by updating a Gaussian process model on previous trials and selecting points of maximum predicted uncertainty to explore quickly the performance landscape. After the experiments, we combined all subjects into another multi-subject Gaussian process that captures subject-specific variation and shared structure; this model agrees closely with the data. Checking with a leave-one-out cross validation procedure, we find that the multi-subject model works well as measured by the strong correlation of the prediction with the test point ($\rho = 0.95$) across all experiments (see Appendixs F.5 and F.6).

Looking across subjects, we find that performance varies with both visibility parameters. To show this trend, we combine performance landscapes across subjects after normalizing them to be centered about the same midpoint of performance (see Appendix F.6). We show an example for $\epsilon^* = 0.55$ s—a few times the fastest motor response time for humans [41]—in Figure 3.1C. At $f = 1$, the avatar is always visible and subject performance is the highest. As avatar visibility is reduced by decreasing the fraction visible f, we observe poorer performance. We also tend to observe better performance at shorter window intervals τ. The variation with τ and f shows systematic trends in performance across subjects

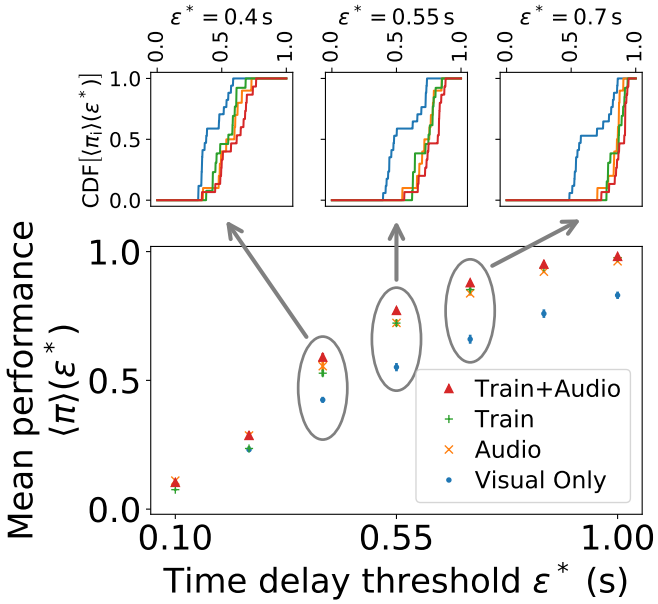


Figure 3.3: Mean of the predicted performance landscape as a function of the time error threshold ϵ^*. (bottom) The mean over performance landscapes $\langle\pi\rangle_{\epsilon^*} = \frac{1}{M}\sum_i \frac{1}{145}\sum_{(\tau,f)} \pi_i(\tau, f; \epsilon^*)$. At the physical limit of reaction times of small ϵ^*, performance converges. For large ϵ^*, performance again converges except for untrained individuals who do poorly in general. For 0.1 s $\leq \epsilon^* \leq$ 0.4 s, mean performance in Audio becomes better than Train, evidence that audio enhances performance at faster time scales. Error bars are two standard deviations of the mean by subject normalized by the square root of the number of subjects and samples $\sqrt{16N}$ and are small, reflecting precise estimates. (top row) Variation in performance across subjects by cumulative distribution functions (CDF) of the average performance per subject i, $\langle\pi_i\rangle(\epsilon^*) = \frac{1}{145}\sum_{(\tau,f)} \pi_i(\tau, f; \epsilon^*)$.

in this mirroring task.

We characterize the typical form of the transition between high and low performance by inspecting the level contours of the aggregated performance landscape in detail in Figure 3.1D. A simple parameterization for the level contours is the nonlinear, inverse relation

$$f = a - b/\tau, \tag{3.1}$$

where a and b are constants. This form captures the fact that for large τ perfor-

mance must become a linear function of f—because performance becomes an average between long visible and invisible windows—and captures the intuition that as τ decreases subjects do better because the rapid, intermittent views simulate stop motion animation. Fitting to the level contour $\tilde{\pi} = 1/2$ on the aggregated landscape, we find that nearly all the landscapes we consider are well captured by Eq 3.1 and better than by a linear relation between f and τ. The results of the best fit parameters are shown on the landscape in Figure 3.1 (blue line, see Appendix F.7). Thus, the shape of the transition region shows that faster windows typically increase the range of f where good performance is accessible across subjects.

To determine if audio cues can affect performance, we introduce the Audio experimental condition where subjects hear a tone whose frequency increases with the speed of the avatar's hand (see Section 3.2). Though the tone does not provide directional information, it can be used to deduce when the avatar is making long sweeping motions or changing directions. We compare Audio with the Train condition, where subjects first undergo a 5 minute, practice version of the experiment. Finally, we combine these two changes in the Train+Audio condition in which subjects are reminded to use and coached on how to use the audio signals. This schema gives four different experimental conditions with N subjects and M unique subject and hand combinations: Visual Only ($N = 10, M = 17$), Audio ($N = 10, M = 10$), Train ($N = 7, M = 13$), and Train+Audio ($N = 8, M = 15$).[2]

The presence of audio and training enhances average performance taken over

[2]Some pairs of subjects and hands were not considered because of errors in the code.

the predicted performance landscapes across subjects $\langle \pi \rangle (\epsilon^*)$. Since this measure depends on ϵ^*, we lower ϵ^* to assess how well subjects track the motion of the avatar at shorter time scales. We expect to find that the points converge at large and very small ϵ^* corresponding to the regimes of generous time delay where subjects do equally well and the limits of human reaction time where subjects perform equally poorly, respectively. Across nearly the entire range shown, where ϵ^* varies from 0.1 s to 1 s, we find large improvement from the Visual Only to all other conditions as shown in Figure 3.3. When comparing the other conditions with each other in the intermediate regime (circled regions), we observe significant differences in mean performance of up to $\sim 25\%$ with the highest performance consistently in the Train+Audio condition. Interestingly, at $\epsilon^* \approx 1/2$ s the order of mean performance of the Train and Audio conditions reverses suggesting that audio cues enhance mean performance more at shorter time scales. This result is consistent with studies showing that human reaction times to audio cues are faster than reactions to visual cues [41], if effects of training were visually mediated, or if training engaged higher-level anticipatory motor responses acting at slower time scales [21, 24]. Collectively, these data demonstrate that for time scales spanning up to four times the human motor response time, from 200–800 ms, there is notable variation in performance depending on whether or not audio cues and training are provided.

To gain insight into what distinguishes good performers across conditions, we investigate the dynamics of how individuals mirror the avatar. We inspect runs of successful mirroring that are indicated by the shaded regions in Figure 3.1B. Each of these runs has a duration t. When subjects are able to mirror the avatar closely, they show two kinds of dynamics: either long runs of close mirroring

or a dearth of immediate failures (see Appendix F.8). We map where these behaviors appear in the parameter space given the condition of high performance $\hat{\pi} \geq 1/2$ (Figure 3.4). We plot in blue where at least one high performance trial appears on the performance landscape for the Train+Audio condition. We plot in red where one high performance trial appears in the Audio or Train conditions. Where blue and red overlap, we color the grid gray. For this comparison we ignore the Visual Only condition where average performance is clearly poor. At $\epsilon^* = 0.7\,\text{s}$, we inspect the region in the bottom right corner where visual gaps are the largest. To identify this region, we draw a line with the form of Eq 3.1 for which it is significantly more probable to find a high performance trial from Train+Audio ($p = 0.18 \pm 0.07$) than from Train or Audio ($p = 0.07 \pm 0.04$) as given by probabilities estimated with the Jeffreys prior and 90% confidence intervals. As we decrease ϵ^* to $1/2\,\text{s}$, this region is still significantly dominated by Train+Audio high performance trials. This effect is no longer significant once $\epsilon^* = 0.4\,\text{s}$, where high performance trials are rare across all conditions. This asymmetry in the distribution across parameter space spanned by high performance trials shows that for a limited range of ϵ^* some subjects in the Train+Audio condition are able to maintain stable mirroring under more difficult scenarios than subjects from the other experimental conditions.

3.1.3 Discussion

How might the few high performers in Train+Audio that do well across the extended parameter range be doing better? One explanation is that they reflect natural variation in the population and are affected by neither training nor audio cues [31]. Although this is a possibility we could only rule out completely by testing the same subjects *de novo* under multiple conditions, the significant in-

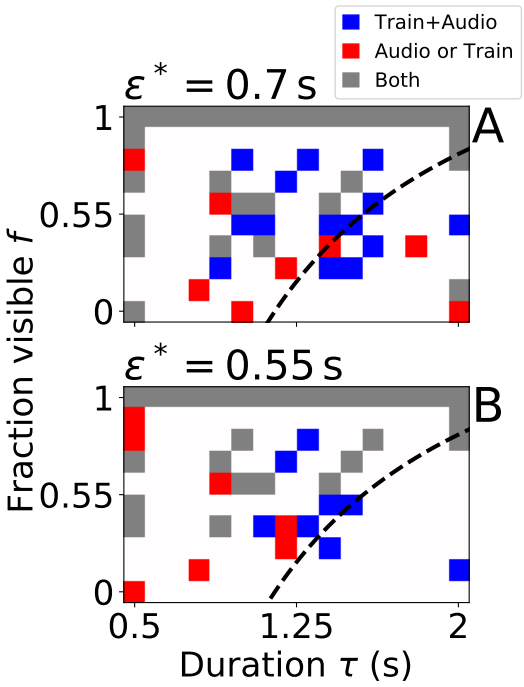


Figure 3.4: Combination of audio cues and training expands the region of parameter space accessible to stable mirroring in the regime of $\sim 1/2$ s. Red indicates where at least one sticky or robust trial with high performance ($\hat{\pi} \geq 1/2$) appears in the Audio or Train conditions. Blue indicates the same for the Train+Audio condition. Gray indicates where red and blue overlap. Black dashed lines denote areas below which there are significantly more trials in the Train+Audio condition than the other two conditions using Jeffreys prior with $p < 0.1$ and parameterized by the functional form given in Eq 3.1.

crease in typical performance over many subjects implies that our experimental conditions are changing behavior. Thus, these dynamic signatures provide evidence that the highest performing individuals are better at learning to use information from the audio cues to mirror the avatar, enabling them to perform well in regions of parameter space inaccessible to low performers. Similar examples of high performers have been identified in a number of experiments involving individuals mirroring the behavior of another subject between a leader and a follower and in joint coordination without a designated leader [8, 30, 47]. The enhanced stability of mirroring runs that we observe is consistent with studies

showing that short training sessions reduce error rates and temporal variability in motor tasks [38]. We find that the observed dynamics have signature distributions of temporal variation that have not been explored in other experiments though studies of have shown that dancers show stronger coherence when following new motions versus non-dancers over a range of time scales [47]. Our experiments also reveal that high performance is facilitated by audio cues, consistent with prior work showing that auditory information enhanced entrainment [34] as well as performance at joint coordination between individuals conducting complementary actions [16, 22]. In the context of these previous studies, our results suggest that people can be trained to use audio cues to perform coordination tasks in regimes where visual cues are sparse.

Our results show that even simple low-dimensional, scalar representations of three-dimensional motion using pitch can enhance the ability to mirror. Similarly, one experiment showed that sonification of a cyclic target motion along with sonification of the subject's own motion could have a beneficial effect on learning greater than natural sounds of motion [13]. Other experiments likewise support the observation that auditory cues can enhance learning of motor tasks [16]. Furthermore, pitch information has been found to substitute for missing visual information in motor perception, consistent with our finding of enhanced performance [12]. Interestingly, subjects responding freely to music showed no association between pitch and speed of hand in one experiment, but our results show that such a mapping can help when subjects are instructed to do so. Indeed, humans are adept at recognizing abstracted motion in various representations through both visual and auditory modalities [20, 46] (see reference [46] for an overview). Flexibility in how human motion can be encoded suggests that

our approach may be just one way of generating helpful auditory cues for mirroring tasks [3, 46]. More broadly, frequency coding of motion is a commonly used approach for representing kinematic and motion values in human experiments [9]. In contrast with experiments using multidimensional encoding of motion [16, 46], we give a simple representation only mapping the speed of the avatar's hand to the frequency of a pure tone such that faster speed corresponds to higher frequency.

We find that overall performance in the Audio and Train conditions are quite similar despite small but significant differences in the time scales at which performance is improved. One explanation for this difference is that response to visual stimuli are slower than to auditory stimuli as measured by reaction times [41], assuming that performance in the Train condition is visually mediated. Indeed, many subjects in the Visual Only condition had a difficult time responding to changes in direction of the avatar's hand, showing considerable latency that lessened with the provision of auditory cues. Another possible explanation for the difference between Audio and Train is that training engages higher-level cognitive processes that act at slower time scales. Indeed, subjects in the Train condition had verbal reinforcement and a brief conversation to talk through the task with the experimenter, perhaps engaging higher-level cognitive functions for motor planning and social context [21, 24, 50]. Furthermore, familiarity with a motor task can improve performance [7, 19, 32]. We note that the presence of longer time scales over which subjects are able to mirror the avatar well in the Train condition compared to the Visual Only condition suggests that subjects are engaging cognitive processes with commensurate time scales going beyond visual reaction times. When we additionally include audio in the Train+Audio

condition, stable trajectories appear over many seconds and many changes in direction, posing an interesting question: in which aspects is time-consuming training substitutable with intuitive perceptual cues?

Our study is just one example of how virtual reality technology combined with a set of statistical learning tools can advance the study of human behavior [27]. An analogous toolkit has been used in cognitive neuroscience where control over the sensory apparatus in model systems has led to significant advances in the understanding of cognitive mechanisms [2, 28, 42]. To adapt this approach for human subjects, we used learning techniques to cover quickly a large parameter range across four different conditions over an order of magnitude in visual duration and another in invisible duration. Similar expansive experiments for mapping multiple conditions and parameters could be used to explore the efficacy of machine-human interfaces [17, 48], determine parameters for athletic performance, and diagnose motor or cognitive conditions with characteristic dynamics [51]. In the context of this study, this combination of techniques has been used to illustrate how visual perception can be augmented with audio signals to enhance coordination. Such developments could prove useful for medical teams synchronizing different tasks, enhancing the fluidity of human-robot interactions, or even learning to improve one's tango.

3.2 Methods

For aligning the velocity trajectories, we use dynamic time warping (DTW) with a cost function for the trajectory comparing times with indices i and j,

$$g(\mathrm{i},\mathrm{j}) = \begin{cases} 0, & |t_\mathrm{i} - t_\mathrm{j} + 1/2| < 1 \\ |t_\mathrm{i} - t_\mathrm{j} + 1/2|^6, & |t_\mathrm{i} - t_\mathrm{j} + 1/2| \geq 1 \end{cases} \tag{3.2}$$

To control the strength of this regularization, we set the coefficient of g to be $\lambda = 10^{-3}$ in the minimized objective function (see Appendix F.2). We first use FastDTW which can calculate the time warp in nearly linear time instead of quadratic time [36]. If the found trajectory ventures outside of the bounding interval $\Delta t \in [-1/2\,\mathrm{s}, 3/2\,\mathrm{s}]$, we then solve the problem using our own (slower) implementation including the regularization specified in Eq. 3.2. We find that about 60% of the untrained trials were regularized whereas only 35% of the trained trials were. We might expect this difference because untrained individuals typically do not replicate the trajectory of the avatar as well and the algorithm is more prone to misaligning stretches of motion.

To measure mirroring error, we measure the fraction of time that the subject is within some time delay ϵ^* measured from alignment with DTW.

$$\hat{\pi}(\tau, f; \epsilon^*) \;\; = \tfrac{1}{\tilde{T}+2}\left(1 + \textstyle\sum_{\tilde{t}}^{\tilde{T}} \Theta\left[\epsilon^* - \left|\epsilon\left(\tilde{t}\right)\right|\right]\right) \tag{3.3}$$

which is regularized by the Laplace counting estimator. The indicator function, given by the Heaviside theta function $\Theta(x \geq 0) = 1$ and $\Theta(x < 0) = 0$, counts when the subject is within or beyond the error threshold. We use the warped time $\tilde{t}$ and normalize by the length of the warped trajectory $\tilde{T}$.

The distributions of durations of mirroring runs are given by three classes: an

exponential, a "sticky" gamma-like function with a dearth of the shortest decay times, and a heavy-tailed "robust" distribution. Although the exponential decay is a signature of a memoryless process, the remaining two distributions suggest that the dynamics of how subjects are tracking the motion of the avatar are generated from a history-dependent process.

The "sticky" distribution is described by the complementary cumulative distribution function (CDF) of decay times, otherwise known as the survival function, as a function of a single rate constant K

$$1 - \mathrm{CDF}(t') \quad = e^{-Kt'} \sum_{\mathrm{n}=0}^{N} \frac{K^{\mathrm{n}} t'^{\mathrm{n}}}{\mathrm{n}!} \tag{3.4}$$

In the limit of $N \to \infty$, we recover the gamma distribution. We find that the measured values of N as calculated with maximum likelihood are concentrated at smaller values. Over 50% of the observed values are smaller than or equal to 5 when $\epsilon^* = 1/2\,\mathrm{s}$, suggesting that enhanced dynamical stability corresponding to the "sticky" distribution is slight. The "robust" distribution describes the first passage time for simple diffusion,

$$1 - \mathrm{CDF}\,(t') \quad = 1 - \sqrt{\frac{\alpha}{\pi}} \int_{1/30}^{t'=t\alpha^*/\alpha} t^{-3/2} e^{-\alpha/t}\, dt. \tag{3.5}$$

Here, the lower limit is important and is given by our interpolation of the velocity trajectories at 30 Hz.

For more details, see Appendix F.

Chapter 3 references

[1] Juan A. Acebrón et al. "The Kuramoto Model: A Simple Paradigm for Synchronization Phenomena". en. In: *Reviews of Modern Physics* 77.1 (Apr. 2005), pp. 137–185. ISSN: 0034-6861, 1539-0756. DOI: 10.1103/RevModPhys.77.137.

[2] Dmitriy Aronov and David W Tank. "Engagement of Neural Circuits Underlying 2D Spatial Navigation in a Rodent Virtual Reality System". In: *Neuron* 84.2 (Oct. 2014), pp. 442–456.

[3] Aurélie Bidet-Caulet et al. "Listening to a Walking Human Activates the Temporal Biological Motion Area". In: *NeuroImage* 28.1 (Oct. 2005), pp. 132–139.

[4] Christopher M Bishop. *Pattern Recognition and Machine Learning*. Singapore: Springer Verlag, Aug. 2006.

[5] Rafal Bogacz et al. "The Physics of Optimal Decision Making: A Formal Analysis of Models of Performance in Two-Alternative Forced-Choice Tasks." In: *Psychol Rev* 113.4 (2006), pp. 700–765.

[6] B W Brunton, M M Botvinick, and C D Brody. "Rats and Humans Can Optimally Accumulate Evidence for Decision-Making". In: *Science* 340.6128 (Apr. 2013), pp. 95–98.

[7] Beatriz Calvo-Merino et al. "Seeing or Doing? Influence of Visual and Motor Familiarity in Action Observation". In: *CURBIO* 16.19 (Oct. 2006), pp. 1905–1910.

[8] Baptiste Caramiaux et al. "Individuality in Piano Performance Depends on Skill Learning". In: *MOCO '17*. ACM, June 2017.

[9] Gaël Dubus and Roberto Bresin. "A Systematic Review of Mapping Strategies for the Sonification of Physical Quantities". In: *PLoS ONE* 8.12 (Dec. 2013), e82491.

[10] John R Dyer, Paul Stapleton, and Matthew Rodger. "Transposing Musical Skill: Sonification of Movement as Concurrent Augmented Feedback Enhances Learning in a Bimanual Task". In: *Psychol Res* 81 (2017), pp. 850–862.

[11] John RG Dyer et al. "Consensus Decision Making in Human Crowds". In: *Animal Behav* 75.2 (2008), pp. 461–470.

[12] Alfred O Effenberg and Gerd Schmitz. "Acceleration and Deceleration at Constant Speed: Systematic Modulation of Motion Perception by Kinematic Sonification". In: *Ann NY Acad Sci* 1425.1 (Aug. 2018), pp. 52–69.

[13] Alfred O Effenberg et al. "Movement Sonification: Effects on Motor Learning beyond Rhythmic Adjustments". In: *Front Hum Neurosci* 10.149 (May 2016), p. 67.

[14] S J Goodbody and D M Wolpert. "The Effect of Visuomotor Displacements on Arm Movement Paths". In: *J Stat Phys* 127 (Mar. 1999), pp. 213–223.

[15] Jonatan Hvass et al. "Visual Realism and Presence in a Virtual Reality Game". en. In: *2017 3DTV Conference: The True Vision - Capture, Transmission and Display of 3D Video (3DTV-CON)*. Copenhagen: IEEE, June 2017, pp. 1–4. ISBN: 978-1-5386-1635-2. DOI: 10.1109/3DTV.2017.8280421.

[16] Tong-Hun Hwang et al. "Effect- and Performance-Based Auditory Feedback on Interpersonal Coordination". In: *Front Psychol* 9 (Mar. 2018), p. 123.

[17] T Iqbal, M J Gonzales, and Laurel D Riek. "Joint Action Perception to Enable Fluent Human-Robot Teamwork". In: *Proc of the 24th IEEE Int Symp on Robot and Human Commun*. IEEE, 2015, pp. 400–406.

[18] F Itakura. "Minimum Prediction Residual Principle Applied to Speech Recognition". In: *IEEE Trans Acoust, Speech, Signal Process* 23.1 (Feb. 1975), pp. 67–72.

[19] Marc Jeannerod. "Neural Simulation of Action: A Unifying Mechanism for Motor Cognition". In: *NeuroImage* 14.1 (July 2001), S103–S109.

[20] G Johansson. "Visual Perception of Biological Motion and a Model for Its Analysis". In: *J Stat Phys* 14 (1973), pp. 201–211.

[21] James M Kilner et al. "Motor Activation Prior to Observation of a Predicted Movement". In: *Nat Neurosci* 7.12 (Dec. 2004), pp. 1299–1301.

[22] Günther Knoblich and Jerome Scott Jordan. "Action Coordination in Groups and Individuals: Learning Anticipatory Control." In: *J Exp Psychol Learn Mem Cogn* 29.5 (2003), pp. 1006–1016.

[23] Ivana Konvalinka et al. "Follow You, Follow Me: Continuous Mutual Prediction and Adaptation in Joint Tapping:" in: *Q J Exp Psychol* 63.11 (Nov. 2010), pp. 2220–2230.

[24] Dimitrios Kourtis, Natalie Sebanz, and Günther Knoblich. "Favouritism in the Motor System: Social Interaction Modulates Action Simulation". In: *Biol Lett* 6.6 (Dec. 2010), pp. 758–761.

[25] Yoshiki Kuramoto and Ikuko Nishikawa. "Statistical Macrodynamics of Large Dynamical Systems. Case of a Phase Transition in Oscillator Communities". en. In: *Journal of Statistical Physics* 49.3-4 (Nov. 1987), pp. 569–605. ISSN: 0022-4715, 1572-9613. DOI: 10.1007/BF01009349.

[26] Edward D. Lee, Edward Esposito, and Itai Cohen. "Audio Cues Enhance Mirroring of Arm Motion When Visual Cues Are Scarce". en. In: *Journal of The Royal Society Interface* 16.154 (May 2019), p. 20180903. ISSN: 1742-5689, 1742-5662. DOI: 10.1098/rsif.2018.0903.

[27] Matthias Minderer et al. "Neuroscience: Virtual Reality Explored". In: *Nature* 533.7603 (May 2016), pp. 324–325.

[28] Ari S Morcos and Christopher D Harvey. "History-Dependent Variability in Population Dynamics during Evidence Accumulation in Cortex". In: *Nat Neurosci* 19.12 (Dec. 2016), pp. 1672–1681.

[29] Meinard Müller. *Information Retrieval for Music and Motion*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2007.

[30] L Noy, E Dekel, and U Alon. "The Mirror Game as a Paradigm for Studying the Dynamics of Two People Improvising Motion Together". In: *PNAS* 108.52 (Dec. 2011), pp. 20947–20952.

[31] Nadine Pecenka and Peter E Keller. "The Role of Temporal Prediction Abilities in Interpersonal Sensorimotor Synchronization". In: *Exp Brain Res* 211.3-4 (Mar. 2011), pp. 505–515.

[32] Narender Ramnani and R Christopher Miall. "A System in the Human Brain for Predicting the Actions of Others". In: *Nat Neurosci* 7.1 (Jan. 2004), pp. 85–90.

[33] C E Rasmussen and C K I Williams. *Gaussian Processes for Machine Learning*. Cambridge: MIT Press, 2006.

[34] Michael J Richardson et al. "Rocking Together: Dynamics of Intentional and Unintentional Interpersonal Coordination". In: *Hum Mov Sci* 26.6 (Dec. 2007), pp. 867–891.

[35] Daniel Roth et al. “Avatar Realism and Social Interaction Quality in Virtual Reality”. en. In: *2016 IEEE Virtual Reality (VR)*. Greenville, SC, USA: IEEE, Mar. 2016, pp. 277–278. ISBN: 978-1-5090-0836-0. DOI: 10.1109/VR.2016.7504761.

[36] Stan Salvador and Philip Chan. “Toward Accurate Dynamic Time Warping in Linear Time and Space”. In: *Intell Data Anal* 11.5 (2007), pp. 561–580.

[37] Robert A Scheidt et al. “Interaction of Visual and Proprioceptive Feedback During Adaptation of Human Reaching Movements”. In: *J Neurophysiol* 93.6 (June 2005), pp. 3200–3213.

[38] Rebecca Scheurich, Anna Zamm, and Caroline Palmer. “Tapping Into Rate Flexibility: Musical Training Facilitates Synchronization Around Spontaneous Production Rates”. In: *Front Psychol* 9 (Apr. 2018), pp. 1–13.

[39] Richard C Schmidt and Michael J Richardson. “Dynamics of Interpersonal Coordination”. In: *Coordination: Neural, Behavioral and Social Dynamics*. Ed. by A Fuchs, V K Jirsa, and Jirsa. Springer, Berlin, Heidelberg, 2008, pp. 281–308.

[40] Elad Schneidman et al. “Weak Pairwise Correlations Imply Strongly Correlated Network States in a Neural Population”. en. In: *Nature* 440.7087 (Apr. 2006), pp. 1007–1012. ISSN: 0028-0836, 1476-4687. DOI: 10.1038/nature04701.

[41] Jose Shelton and Gideon Praveen Kumar. “Comparison between Auditory and Visual Simple Reaction Times”. In: *NM* 1.1 (2010), pp. 30–32.

[42] John R Stowers et al. “Virtual Reality for Freely Moving Animals”. In: *Nat Meth* 14.10 (Oct. 2017), pp. 995–1002.

[43] C Vesper, G Knoblich, and N Sebanz. "Our Actions in My Mind: Motor Imagery of Joint Action". In: *Neuropsychologia* 55 (2014), pp. 115–121.

[44] Cordula Vesper et al. "Are You Ready to Jump? Predictive Mechanisms in Interpersonal Coordination." In: *J Exp Psychol Hum Percept Perform* 39.1 (2013), pp. 48–61.

[45] Cordula Vesper et al. "Joint Action Coordination through Strategic Reduction of Variability". en. In: *Proceedings of the Annual Meeting of the Cognitive Science Society*. Vol. 35. 2013, pp. 1522–1527.

[46] Pia M Vinken et al. "Auditory Coding of Human Movement Kinematics". In: *Multisens Res* 26.6 (2013), pp. 533–552.

[47] A Washburn et al. "Dancers Entrain More Effectively than Non-Dancers to Another Actor's Movements". In: *Front Hum Neurosci* 8 (2014), pp. 1–14.

[48] Christopher D Wickens. "Multiple Resources and Performance Prediction". In: *TIES* 3.2 (Jan. 2002), pp. 159–177.

[49] D M Wolpert, Z Ghahramani, and M I Jordan. "An Internal Model for Sensorimotor Integration". In: *Science* 269.5232 (1995), pp. 1880–1882.

[50] Daniel M Wolpert, Kenji Doya, and Mitsuo Kawato. "A Unifying Computational Framework for Motor Control and Social Interaction". In: *Phil Trans R Soc Lond B* 358.1431 (Feb. 2003), pp. 593–602.

[51] Di Wu et al. "A Biomarker Characterizing Neurodevelopment with Applications in Autism". In: *Sci Rep* 8.614 (Jan. 2018), pp. 1–14.

[52] W Young, M Rodger, and CM Craig. "Perceiving and Reenacting Spatiotemporal Characteristics of Walking Sounds". In: *J Exp Psychol Hum Percept Perform* 39.2 (2012), pp. 464–476.

CHAPTER 4
APPENDICES

A Appendix for Chapter 1.1

A.1 Principle of maximum entropy

The maximum entropy formulation is described in detail in many places and the interested reader is referred to references [7], [45], [58], and [46] for additional material beyond what is presented here. Here, we will present a brief introduction to the maxent formulation and discuss some of the numerical issues and techniques summarized in reference [46].

Maximum entropy refers to a method of building statistical models using the concept of information entropy as described by Shannon. In his seminal paper, Shannon proved that there was a unique measure of uncertainty S for a probability distribution p_i over configurations indexed by i given the three axioms about continuity of entropy with respect to p_i, monotonicity with the number of independent configurations N, and decomposition of uncertainty when configurations are grouped together.[1] This measure is specified up to a positive constant K

$$S = -K \sum_i p_i \log p_i. \tag{A1}$$

We recognize this as the thermodynamic entropy when the probability of all microstates is equal and $S = k_B \log N$, in units of Boltzmann's constant.

Shannon was working on a theory of communication and considering the predictability of a future sequence of bits that were to follow in a message. In this context, the information entropy, or entropy hereon, is a measure of the surprise of the next symbol in the sequence given the statistics of the ensemble. For example, in the English alphabet it would be very surprising to find a "z" as the next letter in comparison with the very frequent "e." This surprise is inversely related to the predictability of a sequence.

Consider, as another example, the biased coin with probability of heads p and of tails $1 - p$. When the coin is fair and $p = 1/2$, there is no way of doing better

[1] This was later generalized in the Shannon-Khinchine entropy by weakening one of Shannon's assumptions about arbitrarily large N [42].

than chance at predicting what the next coin flip will return, and the entropy is maximized $S = \log 2$ nats (natural units of information). As soon as any bias is present, the entropy decreases, and there exists some statistical structure in the data by which we can improve our prediction of the next coin flip.

This simple example demonstrates that the entropy is maximized when the probability of all configurations is equal, reflecting the express intent to measure the amount of statistical structure encoded in the probability distribution. By maximizing the entropy, we can squeeze all the available structure in our model while constraining it to fit features that we believe are important. This is the maximum entropy principle [37].

Given a probability distribution $p(s)$ of configurations s, observables $f_{\mathrm{n}}(s)$ that we are interested in fitting, and Langrangian multipliers θ_{n}, we maximize the functional

$$\mathcal{L}[p(s)] = S[p(s)] + \sum_{\mathrm{n}} \theta_{\mathrm{n}} \left(\langle f_{\mathrm{n}} \rangle_{\mathrm{data}} - \langle f_{\mathrm{n}} \rangle \right) + \theta' \left(1 - \sum_{s} p(s) \right). \tag{A2}$$

The term with the Lagrangian multiplier θ' serves to normalize the distribution such that it sums to unity. Just to be explicit with what we are summing over,

$$\begin{aligned} = & - \sum_{s} p(s) \log p(s) + \sum_{\mathrm{n}} \theta_{\mathrm{n}} \left(\frac{1}{K} \sum_{\mathrm{k}=1}^{K} p_{\mathrm{data}}(s) f_{\mathrm{n}}(s) - \sum_{s} p(s) f_{\mathrm{n}}(s) \right) + \\ & \theta' \left(1 - \sum_{s} p(s) \right) \end{aligned} \tag{A3}$$

where we have K data points. By maximizing this quantity, we find the maxent model, the Boltzmann distribution, the exponential family, or log-linear models (depending on your field)

$$p(s) = e^{-E(s)} \Big/ Z \tag{A4}$$

with partition function for normalization

$$Z = \sum_{s} e^{-E(s)} \tag{A5}$$

and energy function

$$E(s) = - \sum_{\mathrm{n}} \theta_{\mathrm{n}} f_{\mathrm{n}}(s). \tag{A6}$$

Thus, we might think of statistical physics models to be the minimal models consistent with the data, given well-chosen constraints that fix the averages of observables.

A.2 Numerical solution

Given the maxent model from Eq A4, how do we find the values of the parameters that are consistent with the contraints that we have chosen? This is known as the inverse problem since we have the observables and must find the parameters instead of the usual problem of finding the behavior of the model given the parameters. In general, this is a difficult problem.

To find the parameters λ_n that match the constraints $\langle f_n \rangle_{\text{data}}$, we can minimize the Kullback-Leibler divergence between the model and the data [20]

$$D_{\text{KL}}(p_{\text{data}} \| p) = \sum_s p_{\text{data}} \log \left(\frac{p_{\text{data}}(s)}{p(s)} \right) \tag{A7}$$

$$\frac{\partial D_{\text{KL}}}{\partial \lambda_n} = \sum_s p_{\text{data}}(s) \frac{\partial(-E(s) - \log Z)}{\partial \lambda_n} = 0$$

$$\implies \langle f_n \rangle_{\text{data}} = \langle f_n \rangle \,. \tag{A8}$$

In other words, the parameters of the maxent model are the ones that minimize the information theoretic "distance" to the distribution of the data by matching the constraints. Note that these parameters are not fit in the conventional sense where there is arbitrariness in the range of possible values. Once the constraints have been chosen, there is a single maxent solution with no free parameters.

With recent interest in this problem, there have been a number of clever algorithms that have been suggested for efficient numerical approximation of the solution [58]. To provide a simple interface to a number of algorithms, we developed the Convenient Interface to Inverse Ising (ConIII pronounced CONee), an open-source Python project for solving pairwise and higher-order maxent models and a base for future extension [46]. The algorithms that are implemented as part of ConIII include Monte Carlo histogram, pseudolikelihood, minimum probability flow, a regularized mean field method, and a cluster expansion method. Our goal was to make a variety of maximum entropy techniques accessible to those unfamiliar with the techniques and accelerate workflow for users. We discuss the implemented methods briefly in turn below.

Enumeration

The naïve approach that only works for small systems is to write out the equations from Eq A8 and solve them numerically. After writing out all K equations,

$$\langle f_n \rangle = -\frac{\partial \ln Z}{\partial \lambda_n} = \langle f_n \rangle_{\text{data}} \,, \tag{A9}$$

we can use any standard root-finding algorithm to find the parameters λ_n. This approach, however, involves enumerating all states of the system, whose number grows exponentially with system size.

For the Ising model, writing down the equations has a number of steps $O(K^2 2^N)$, where K is the number of constraints and N the number of spins. Each evaluation of the objective in the root-finding algorithm will be of the same order. For relatively small systems, around $N \leq 15$, this approach is feasible on a typical desktop computer and is a good way to test the results of a more sophisticated algorithm.

Monte Carlo Histogram (MCH)

Perhaps the most straightforward though expensive computational approach is Monte Carlo Markov Chain (MCMC) sampling. A series of states sampled from a proposed $p(s)$ is produced by MCMC to approximate $\langle f_n \rangle$ and determine how close we are to matching $\langle f_n \rangle_{\text{data}}$. The parameters are then adjusted using a learning rule, and both sampling and learning are repeated until a stopping criterion is met. This can be combined with a variety of approximate gradient descent methods to reduce the number of sampling steps by predicting how the distribution will change if we modify the parameters slightly. The particular technique implemented in ConIII is the Monte Carlo Histogram (MCH) method [11].

Since the sampling step is expensive, the idea behind MCH is to reuse a sample for more than one gradient descent step [11]. Given that we have a sample with probability distribution $p(s)$ generated with parameters λ_n, we would like to estimate the proposed distribution $p'(s)$ from adjusting our parameters $\lambda'_n = \lambda_n + \Delta\lambda_n$. We can leverage our current sample to make this extrapolation.

$$p' = \frac{p'}{p} p \tag{A10}$$

$$p'(s) = \frac{Z}{Z'} e^{\sum_n \Delta\lambda_n f_n(s)} p(s) \tag{A11}$$

To estimate the average,

$$\sum_s p'(s) f_n(s) = \frac{Z}{Z'} \sum_s p(s) e^{\sum_n \Delta\lambda_n f_n(s)} f_n(s) \tag{A12}$$

To be explicit about the fact that we only have a sampled approximation to p, we replace p with the sample distribution.

$$\langle f_n \rangle' = \frac{Z}{Z'} \left\langle e^{\sum_n \Delta\lambda_n f_n(s)} f_n(s) \right\rangle_{\text{sample}} \tag{A13}$$

Likewise, the ratio of the partition function can be estimated

$$\frac{Z}{Z'} \approx 1 \Big/ \left\langle e^{\sum_n \Delta\lambda_n f_n(s)} \right\rangle_{\text{sample}} \tag{A14}$$

At each step, we update the Lagrangian multipliers $\{\lambda_n\}$ while being careful to stay within the bounds of a reasonable extrapolation. One suggestion is to update the parameters with some inertia [79]

$$\Delta\lambda_n(t+1) = \Delta\lambda_n(t) + \epsilon \Delta\lambda_n(t-1) \tag{A15}$$

$$\Delta\lambda_n(t) = \eta \left(\langle f_n \rangle' - \langle f_n \rangle \right) \tag{A16}$$

This has a fixed point at the correct parameters.

In practice, MCH can be difficult to tune properly and one must check in on the progress of the algorithm often. One issue is choosing how to set the learning rule parameters η and ϵ. One suggestion for η is to shrink it as the inverse of the number of iterations [79]. Another issue is that parameters cannot be changed by too much when using the MCH approximation step or the extrapolation to λ'_n will be inaccurate and the algorithm will fail to converge. In ConIII, this can be controlled by setting a bound on the maximum possible change in each parameter $\Delta\lambda_{\max}$ and restricting the norm of the vector of change in parameters $\sum_k \sqrt{\Delta\lambda_n^2}$. Another issue is setting the parameters of the MCMC sampling routine. Both the burn time (the number of iterations before starting to sample) and sampling iterations (number of iterations between samples) must be large enough that we are sampling from the equilibrium distribution. Typically, these are found by measuring how long the energy or individual parameter values remain correlated as MCMC progresses. The parameters may need to be updated during the course of MCH because the sampling parameters may need to change with the estimated parameters of the model. For some regimes of parameter space, samples are correlated over long times and alternative sampling methods like Wolff or Swendsen-Wang would vastly reduce time to reach the

equilibrium distribution although these are not included in the current release of ConIII. We do not discuss these sampling details here, but see references [50, 56] for examples.

The main computational cost for MCH lies in the sampling step. For each iteration of MCH, the runtime is proportional to the number of samples K, number of MCMC iterations T, and the number of constraints for the Ising model N^2, $O(TKN^2)$, whereas the MCH estimate is relatively quick $O(tKN^2)$ because the number of MCH approximation steps needed to converge is much smaller than the number of MCMC sampling iterations $t \ll T$.

Pseudolikelihood

The pseudolikelihood approach is an analytic approximation to the likelihood that drastically reduces the computational complexity of the problem and is exact as $N \to \infty$ [3]. We calculate the conditional probability of each spin s_i given the rest of the system $\{s_{j\neq i}\}$

$$p\left(s_i|\{s_{j\neq i}\}\right) = \left(1 + e^{-2s_i\left(h_i+\sum_{j\neq i} J_{ij}s_j\right)}\right)^{-1} \tag{A17}$$

Taking the logarithm, we define the approximate log-likelihood by summing over data points indexed by r:

$$f(h_i, \{J_{ij}\}) = \sum_{r=1}^{R} \ln p\left(s_i^{(r)} \middle| \{s_{j\neq i}\}^{(r)}\right). \tag{A18}$$

In the limit where the ensemble is well sampled, the average over the data can be replaced by an average over the ensemble:

$$f(h_i, \{J_{ij}\}) = \sum_{s} \ln p\left(s_i|\{s_{j\neq i}\}\right) p\left(s; h_i, \{J_{ij}\}\right). \tag{A19}$$

To find the point of maximum likelihood for a single spin s_i, we calculate the analytical gradient and Hessian, $\partial f/\partial J_{ij}$ and $\partial^2 f/\partial J_{ij}\partial J_{i'j'}$ for a Newton conjugate-gradient descent method. After maximizing likelihood for all spins, the maximum likelihood parameters may not satisfy the symmetry $J_{ij} = J_{ji}$. We impose the symmetry by insisting that

$$J'_{ij} = (J_{ij} + J_{ji})/2. \tag{A20}$$

Pseudolikelihood is extremely fast and often surprisingly accurate. Each calculation of the gradient is order $O(RN^2)$ and Hessian $O(RN^3)$, which must be done for all N. With analytic forms for the gradient and Hessian, the conjugate-gradient descent method tends to converge quickly.

Minimum Probability Flow (MPF)

Minimum probability flow involves analytically approximating how the probability distribution *changes* as we modify the *configurations* [73, 74]. In the methods so far mentioned, the approach has been to maximize the objective (the likelihood function) by immediately taking the derivative with respect to the parameters. With MPF, we first posit a set of dynamics that will lead the data distribution to equilibrate to that of the model. When these distributions are equivalent, then there is no "probability flow" between them. This technique is closely related to score matching, where we instead have a continuous state space and can directly take the derivative with respect to the states without specifying dynamics [35].

First note that Monte Carlo dynamics (satisfying ergodicity and detailed balance) would lead to equilibration to the stationary distribution. One such transition matrix suggested in reference [73] is

$$\dot{p}_{\mathrm{s}} = \sum_{\mathrm{s}' \neq \mathrm{s}} \Gamma_{\mathrm{ss}'} p_{\mathrm{s}'} - \sum_{\mathrm{s}' \neq \mathrm{s}} \Gamma_{\mathrm{s}'\mathrm{s}} p_{\mathrm{s}} \tag{A21}$$

$$\Gamma_{\mathrm{ss}'} = g_{\mathrm{ss}'} \exp\left[\frac{1}{2}\left(E_{\mathrm{s}'} - E_{\mathrm{s}}\right)\right] \tag{A22}$$

with transition probabilities $\Gamma_{\mathrm{ss}'}$ from state s′ to state s. The connectivity matrix $g_{\mathrm{ss}'}$ specifies whether there is edge between states s and s′ such that probability can flow between them. By choosing a sparse $g_{\mathrm{ss}'}$ while not breaking ergodicity, we can drastically reduce the computational cost of computing this matrix.

Imagine that we start with the distribution over the states as given by the data and run the Monte Carlo dynamics. When data and model distributions are different, probability will flow between them and indicate that the parameters must be changed. By minimizing a derivative of the Kullback-Leibler divergence, we measure how the difference between the model and the states in the data $\mathcal{D}$ changes when the dynamics are run for an infinitesimal amount of time.

$$L(\{\lambda_{\mathrm{n}}\}) \equiv \partial_t D_{\mathrm{KL}}(p^{(0)} \| p^{(t)}(\{\lambda_{\mathrm{n}}\})) = \sum_{\mathrm{s} \notin \mathcal{D}} \dot{p}_{\mathrm{s}}(\lambda_{\mathrm{n}}) \tag{A23}$$

The idea is that this derivative is also minimized with optimal parameters: the MPF algorithm looks for a minimum of the objective function L.

For the Ising model, each evaluation of the objective function where $\Gamma_{\mathrm{ss}'}$ connects each data state with G neighbors has runtime $O(RGN^2)$. In a large fully

connected system, $G \sim 2^N$ would be prohibitively large so a sparse choice is necessary.

Regularized mean-field method

One attractively simple and efficient approach uses a regularized version of mean-field theory. In the inverse Ising problem, mean-field theory is equivalent to treating each binary individual as instead having a continuously varying state (corresponding to its mean value). The inverse problem then turns into simply inverting the correlation matrix C [19]:

$$J_{ij}^{\text{mean-field}} = -\frac{(C^{-1})_{ij}}{\sqrt{p_i(1-p_i)p_j(1-p_j)}}, \tag{A24}$$

where

$$C_{ij} = \frac{p_{ij} - p_i p_j}{\sqrt{p_i(1-p_i)p_j(1-p_j)}}, \tag{A25}$$

and where p_i corresponds to the frequency of individual i being in the active (+1) state and p_{ij} is the frequency of the pair i and j being simultanously in the active state.

A simple regularization scheme in this case is to discourage large values in the interaction matrix J_{ij}. This corresponds to putting more weight on solutions that are closer to the case with no interactions (independent individuals). A particularly convenient form adds the following term, quadratic in J_{ij}, to the negative log-likelihood:

$$\gamma \sum_i \sum_{i<j} J_{ij}^2 p_i(1-p_i)p_j(1-p_j). \tag{A26}$$

In this case, the regularized version of the mean-field solution in Eq A24 can be solved analytically, with the slowest computational step coming from the inversion of the correlation matrix. For details, see references [4, 21].

The idea is then to vary the regularization strength γ to move between the non-interacting case ($\gamma \to \infty$) and the naively calculated mean-field solution Eq A24 ($\gamma \to 0$). While there is no guarantee that varying this one parameter will produce solutions that are good enough to "fit within error bars," this approach has been successful in at least one case of fitting social interactions [21].

The inversion of the correlation matrix is relatively fast, bounded by $O\left(N^3\right)$. Finding the optimal γ, involves Monte Carlo sampling from the model distribution, which has computational cost similar to MCH. It is, however, much more efficient because we are only optimizing a single parameter.

Cluster expansion

Adaptive cluster expansion [4, 19] iteratively calculates terms in the cluster expansion of the entropy S:

$$S - S_0 = \sum_{\Gamma} \Delta S_{\Gamma}, \tag{A27}$$

where the sum is over clusters Γ and in the exact case includes all $2^N - 1$ possible nonempty subsets of individuals in the system. In the simplest version of the expansion, one expands around $S_0 = 0$. In some cases it can be more advantageous to expand around the independent individual solution or one of the mean-field solutions described in the previous section [4].

The inverse Ising problem is solved independently on each of the clusters, which can be done exactly when the clusters are small. These results are used to construct a full interaction matrix J_{ij}. The expansion starts with small clusters and expands to use larger clusters, neglecting any clusters whose contribution ΔS_{Γ} to the entropy falls below a threshold. To find the best solution that does not overfit, the threshold is initially set at a large value and then lowered, gradually including more clusters in the expansion, until samples from the resulting J_{ij} fit the desired statistics of the data sufficiently well.

The runtime will depend on the size of clusters included in the expansion. If the expansion is truncated at clusters of size n, the worst-case runtime would be $O\left(\binom{N}{n} 2^n\right)$. The point is that S can often be accurately estimated even when $n \ll N$.

B Appendix for Chapter 1.2

B.1 Missing data

For some data sets, missing data is a problem. Sometimes missing data is a genuine additional state (e.g., self-recusal in a court, absenteeism in a legislature) and sometimes it is a fault of the experimentalist (e.g., equipment miscalibration, he fell asleep). As one strategy for handling missing spins or votes, we might choose to model the distribution of missing votes such that we assign some probability to each possible way data points may be missing. This step, however, will drastically increase the state space and grow the complexity of the model.

As another strategy, we might try to infer what the missing data point could have been had the data point not been missing by conditioning on the statistics of the rest of the ensemble. This, of course, assumes that the fact that a voter was missing does not substantially alter the pattern of correlations between the remaining voters. Then, the data is a view of the marginalized distribution,

$$r_{\text{data}}(s') = \sum_{s \setminus s'} p_{\text{data}}(s) \tag{B1}$$

where we have marginalized over the configuration of all spins s except for the subset that were observed s'. The same marginalization for the maximum entropy model is

$$r(s') = \sum_{s \setminus s'} p(s) \tag{B2}$$

$$= \sum_{s \setminus s'} e^{-E(s)}/Z \tag{B3}$$

By taking subsets of spins from the set of all spins, we can minimize the KL

divergence between the marginal distributions

$$D_{\mathrm{KL}}[(r_{\mathrm{data}}(s')\|r(s')] = \sum_{s'} r_{\mathrm{data}}(s') \log\left(\frac{r_{\mathrm{data}}(s')}{r(s')}\right) \tag{B4}$$

$$\frac{\partial D}{\partial \lambda_{\mathrm{k}}} = -\sum_{s'} r_{\mathrm{data}}(s') \frac{\partial\left[\log\left(\sum_{s\setminus s'} e^{-E(s)}\right) - \log Z\right]}{\partial \lambda_k} \tag{B5}$$

$$0 = \sum_{s'} r_{\mathrm{data}}(s') \frac{\sum_{s\setminus s'} f_{\mathrm{k}}(s) e^{-E(s)}}{\sum_{s\setminus s'} e^{-E(s)}} - \langle f_{\mathrm{k}}\rangle \tag{B6}$$

$$0 = \langle\langle f_{\mathrm{k}}\rangle_s\rangle_{\mathrm{data}} - \langle f_{\mathrm{k}}\rangle \tag{B7}$$

$$\implies \langle\langle f_{\mathrm{k}}\rangle_s\rangle_{\mathrm{data}} = \langle f_{\mathrm{k}}\rangle \tag{B8}$$

In contrast with the original objective in Eq A8 where we just try to match the constraints as measured from the data, the objective for matching marginal distributions introduces a weighted average across this subset of spins accounting for the frequency with they are observed to appear in the data. By assuming that the marginalized distribution comes from the model, we no longer have a fixed average from the data that we converge to as in Eq A8. Instead, the target moves as we change the parameters, this is no longer guaranteed to have a single fixed point.[2]

Instead of solving this hard problem, we will check how well we can do with the most naïve approach. Returning to the original objective function for matching the observables in Eq B8, we will minimize the norm distance numerically and check to see if the solution is close. The reason that this is not guaranteed to succeed is because it is possible to obtain a set of apparently inconsistent marginals from a finite, random sample where spins are hidden.

To make it clear what we mean by "inconsistent marginals," let us work through an example with the pairwise maxent model. Since only subsets of the pairwise correlations are observed, it is possible to find subsets that are described by no joint probability distribution. Take for example three voters from which we

[2] One can see this by setting up the Langrangian for entropy maximization and noting that the constraints now involve the reweighted probability distribution, which is nonlinear. Given that the Shannon entropy is a convex function of the probability p, linear constraints will preserve the convexity of the Lagrangian function (as we had used for fixing the correlations before), but nonlinear constraints will not in general.

happen to sample three configurations with one voter missing every time:

$$\begin{array}{ccc} s_1 & s_2 & s_3 \\ \hline 1 & 1 & \\ 1 & & 1 \\ & -1 & 1 \end{array} \tag{B9}$$

We calculate the pairwise correlations ignoring the missing entries where we have no data. Clearly, there is no triplet probability distribution that can accommodate this set of pairwise correlations because $\langle s_1 s_2 \rangle = \langle s_1 s_3 \rangle = 1$ implies that $\langle s_2 s_3 \rangle = 1$, whereas the latter is -1 in this sample. This example is a limiting case where the conflict between marginals is obvious. Such conflicts may be harder to identify for real data sets, but it is straightforward to describe the general case.

To state this problem more formally, we define

$$p_{\mathrm{ij\cdots k}} = p(s_\mathrm{i} = s_\mathrm{j} = \cdots = s_\mathrm{k} = 1) \tag{B10}$$

For any triplet, we can then write down the probability of every configuration as

$$\begin{array}{cc} \text{votes} & \text{probability} \\ + + + & p_{123} \\ + + - & p_{12} - p_{123} \\ + - + & p_{13} - p_{123} \\ + - - & p_1 - p_{12} - p_{13} + p_{123} \\ - + + & p_{23} - p_{123} \\ - + - & p_2 - p_{23} - p_{12} + p_{123} \\ - - + & p_3 - p_{23} - p_{13} + p_{123} \\ - - - & 1 - p_1 - p_2 - p_3 + p_{12} + p_{13} + p_{23} - p_{123} \end{array} \tag{B11}$$

Each probability must be confined within the bounds $0 \leq p_{\mathrm{ij\cdots k}} \leq 1$ and the sum over all states must be normalized. These constraints present a system of linear equations consisting of 2^{n+1} inequalities minus the measured pairwise correlations in the data

$$0 \leq A\mathbf{p} \leq 1 \quad \text{given } p_{\mathrm{ij\ldots k}} = p^{\mathrm{data}}_{\mathrm{ij\ldots k}} \tag{B12}$$

A is the matrix of coefficients for each possible configuration with vector of probabilities listed in $\mathbf{p}$ and the constraints are given by the data. We check whether or not the measured marginals are compatible by using the simplex algorithm that guarantees the discovery of a solution if it exists and the absence of a solution if it cannot find one [2]. Although it quickly becomes computationally expensive to perform this check for all combinations of higher orders, we can still check the consistency of many subgroups with large errors below a reasonable size.

Application to the Super Court

Looking at the Super Court from Chapter 1, we find that nearly all combinations of three voters satisfy these inequalities in Eq B12, but there are a few exceptions. Out of the 1,851 triplets where we observe at least two of the three pairwise correlations, 83 triplets violate the consistency requirements for a joint probability distribution. Although 34 of the 36 justices appear at least once in this set, most of the violations of the constraints are small. Even for triplets with the largest violations of the constraint, the errors in fitting the correlations of the subset pairs are small as show in Figure 1.9 even if beyond what would be expected from sample size errors.

We again check if the sets of marginals $\{p_1, ..., p_{12...k}\}$ for subsets of sizes $k = 4$, $k = 6$, and $k = 8$ are self-consistent. We find that for $k = 4$, 40% of the 1,554 quartets where all four voters voted together (13,174 quartets are possible if we count all quartets that compose a connected graph with edges $J_{ij} \neq 0$) are inconsistent and that larger errors are indicative of relatively more inconsistencies. This suggests that inconsistencies are to blame for poor fits. However, there is no such correlation for $k = 6$ and $k = 8$ because all sets of marginals are inconsistent. As an alternative check, we enumerate the triplets for each subset to calculate a fraction of inconsistent triplets. we find that the fraction is uncorrelated with the size of the error suggesting inconsistent constraints are not why larger subsets are poorly fit. Thus, inconsistencies in the set of marginals may explain some of the poor fits for subsets of size $k = 4$, but they do not explain the larger subsets of $k = 6$ and $k = 8$.

This result might be unsurprising because higher order correlations can only be measured over shorter periods of time with fewer natural courts as justices are replaced; lower-order correlations can be measured over many natural courts. If the distribution of cases or interactions fluctuate, we would not expect to extrapolate well to higher-order marginals after observing lower-order marginals because they come different distributions as we explain in the next section.

We show the fit to the pairwise correlations in Figure 1.9 and with the justices ordered by a measure of ideology in Figure B1. Though we found that the set of pairwise marginals was inconsistent, the found model comes close to fitting all the observed pairwise correlations and does much better than an example of standard political science model W-Nominate (Appendix B.3).

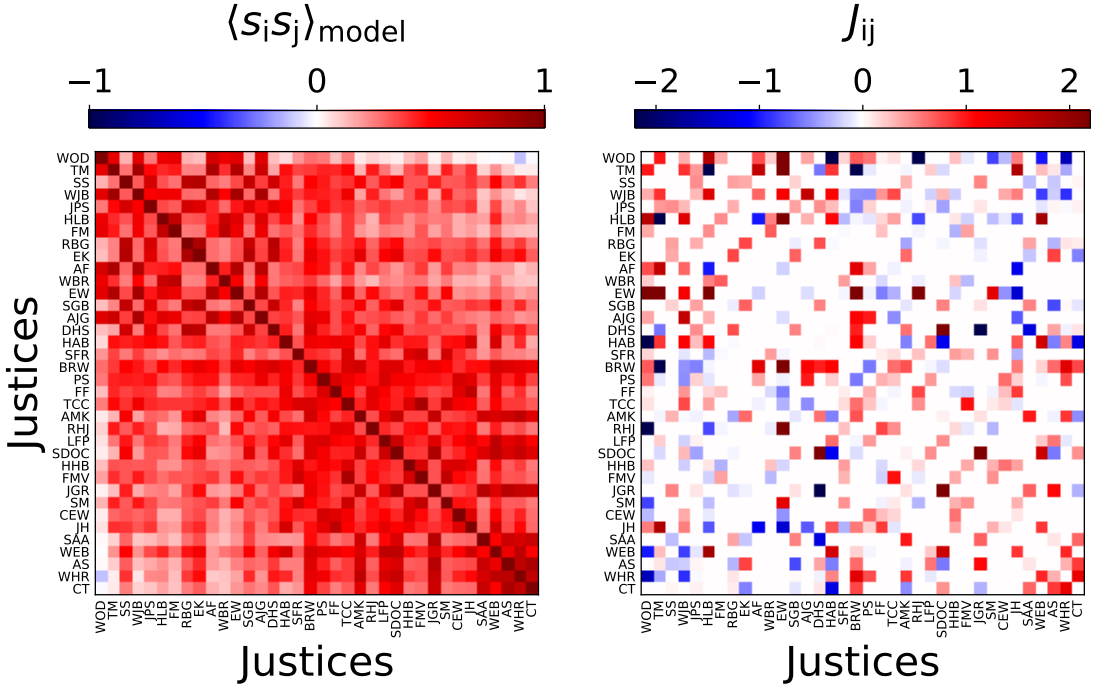


Figure B1: Predicted correlations $\langle s_i s_j \rangle$ and couplings J_{ij} where justices are ordered by mean Martin-Quinn ideological scores. See Figure 1.9 for the justices ordered by date of retirement.

B.2 Evaluating the fit of the maxent model

Since we only fit to pairwise statistics, we test the model against higher-order correlations. we show the fits to 4th, 6th, and 8th order correlations in Figure 1.9. The fits are overall good with correlation coefficients $\rho_4 = 0.81$, $\rho_6 = 0.56$, and $\rho_8 = 0.24$, but there are a small number of correlations that deviate strongly from the model. As another check, we compare the full probability distribution of votes for all nine-member natural courts in Figure B2. Again, the model mostly captures these distributions, but there are some natural courts with significantly larger deviations from the model than others. If the deviations were explained by statistical errors from small sample size, then the errors on the pairwise correlations would be given by the width of the binomial distribution $\delta_{ij} = \sqrt{(\langle s_i s_j \rangle + 1)(1 - \langle s_i s_j \rangle)/4K_{ij}}$ with K_{ij} votes. We find that errors on the pairwise correlations for some natural courts are larger than what would be expected from sample size fluctuations.

To summarize the fit to all the higher-order correlations that we present in Figure 1.9, we calculate the KL divergence between the data distribution of a subset of size k, $p(s_{i_1}, s_{i_2}, ..., s_{i_k})$, with the model $q(s_{i_1}, s_{i_2}, ..., s_{i_k})$ normalized by the

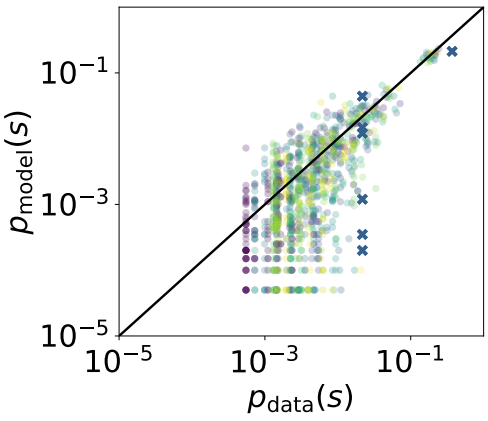


Figure B2: Probability distributions of each 24 nine-member natural court in the data as predicted by the pairwise maxent model. Each distribution indicated by color is calculated from a set of cases where the same nine members voted. One distribution corresponding to a short voting record and is particularly poorly fit is indicated with bold x's.

entropy of q,

$$F(p\|q) \equiv 1 + \frac{\sum_s p(s)\log_2\left(\frac{p(s)}{q(s)}\right)}{\sum_s q(s)\log_2 q(s)} \tag{B13}$$

When the pairwise model q is exactly the distribution over the subset of size k we have $F = 1$, but if the distributions are completely different such that $p(s) = 0$ wherever $q(s) \neq 0$ we have $F = 0$.[3]

For every subset of size k, we plot F as function of the number of votes in which all k members vote in Figure B3. F is smaller for smaller samples whereas large samples are fit well. If fluctuations in the sample size for estimating $p(s)$ were to blame for a poor fit, we would not expect to see a systematic decline as sample sizes got smaller. As we take larger k, the fit becomes poorer according to F. Since we are narrowing the measured distributions to fewer natural courts with larger k, the distribution of votes in some natural courts must be systematically different from that captured by the pairwise correlations.

Such deviations might be expected because we measure the pairwise correlations averaged over the entire period of overlap between two justices which

[3]Information quantities like these can be tricky to estimate correctly with a small sample size [7, 45]. In particular, the entropy of the data p may be systematically underestimated when calculated in the way presented in Eq B13, but such a bias would only be significant for larger subsets, would bias F to be larger, and thus would not change the following conclusions.

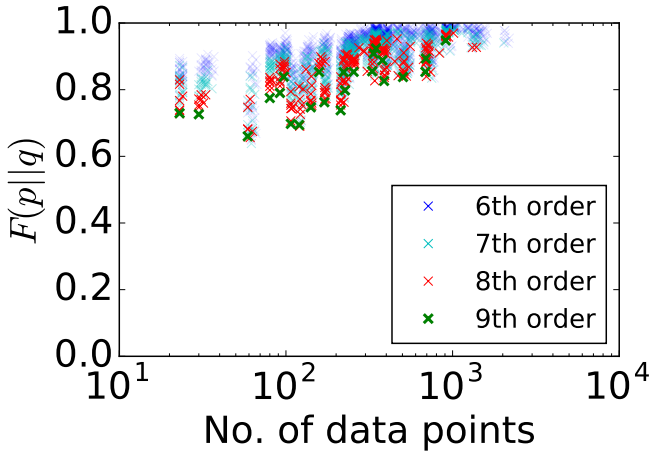


Figure B3: Normalized Kullback-Leibler divergence defined in Eq B13 for all measured distributions for subsets of size 6 (blue), 7 (cyan), 8 (red), and 9 (green). Pairwise maxent model does systematically worse when compared to higher-order marginals calculated from fewer data point.

can be over many annual terms. The higher the order of the observed correlations, however, the shorter the time period during which those k justices voted together, so a plausible explanation would be that correlations are subject to strong short-time temporal fluctuations that are not captured by the long-time averaged pairwise correlations.

In Figure B4, we show how strong these fluctuations are for pairwise correlations using two justices as an example. For the two ideologically median voters on the Second Rehnquist Court (1994–2005), we calculate the autocorrelation function (ACF) for the pairwise correlations as a function of the year $\langle s_i s_j \rangle_t$. There are strong fluctuations from year-to-year and even interesting patterns on longer time scales. To compare with the ACF for pairwise correlations, we also show the ACF for the probability that a particular legal topic is decided that year (as obtained from the Supreme Court Database Project). Again, we find strong fluctuations and observe that changes in the frequency of some of the topics follow patterns similar to that of the pairwise correlations. Inspecting Figure B4, we see that the change in the frequency of First Amendment cases (topic 1) is correlated with the linear decay in S.D. O'Connor's ACF with most other justices. The prominent peaks in A.M. Kennedy's ACF with A. Scalia and C. Thomas on even and odd years seem mostly to coincide with the prevalence of judicial power cases. Without knowing how such topics are chosen (such details are explicitly omitted from the database), it is impossible to know if they are responsible for the yearly variation in the pairwise correlations. However, the cases heard by the Supreme Court are not sampled randomly from a fixed distribution, so it seems likely that fluctuations in pairwise correlations would

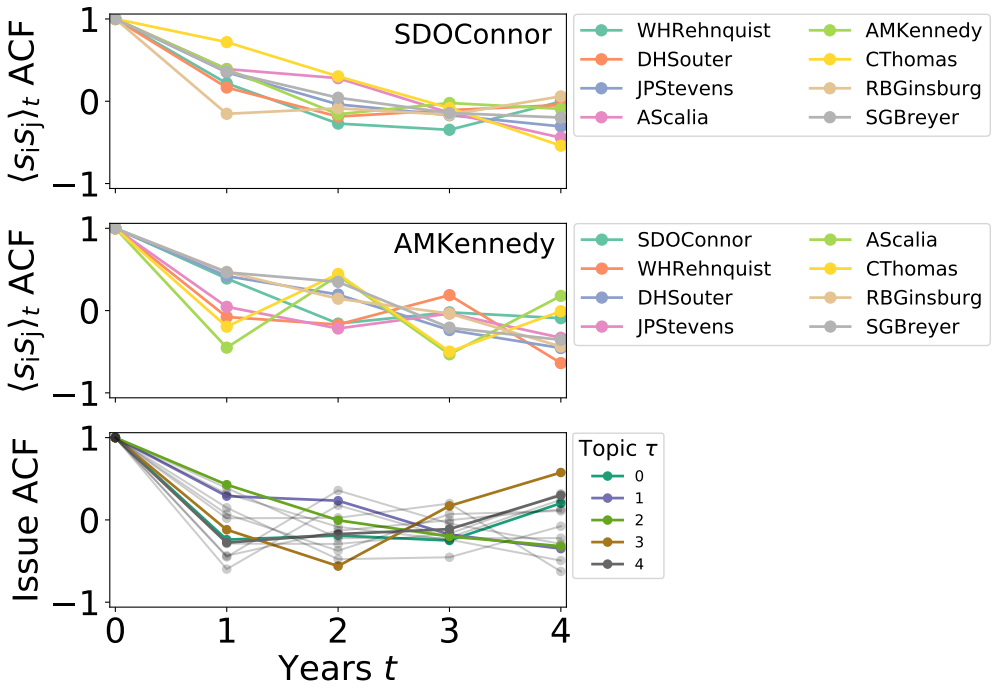


Figure B4: Pairwise correlation $\langle s_i s_j \rangle$ autocorrelation functions (ACF) for swing voters (top) S.D. O'Connor with the 8 other justices during the Second Rehnquist Court 1994–2005. (middle) Same graphs for A.M. Kennedy. (bottom) Autocorrelation function of the probabilities of the most frequently occurring five legal topics during this natural court where $\tau = 0$ refers to Civil Rights, $\tau = 1$ First Amendment, $\tau = 2$ Due Process, $\tau = 3$ Judicial Power, $\tau = 4$ Federalism, $\tau = 5$ Interstate Relations, $\tau = 6$ Federal Taxation. Remaining 7 issues are shown in light gray. Lines drawn to guide the eye.

be related to those in the cases heard.

We find by probing into correlations that are not well fit that higher-order subsets are more likely to have substantial prediction errors, but that these errors can be explained by short temporal fluctuations in the measured correlations between justices. Part of these fluctuations seem to derive from the mix of cases that justices review in a given year [5]. we show that one proxy, the legal topic, shows strong fluctuations in the distribution by term that seem correlated with the variation in pairwise correlations. Another possible explanation is that the underlying interactions between justices change in time. The Supreme Court literature points to some justices that show ideological drift like H.L. Blackmun who became more liberal over time [51], a pattern that also manifests in his pairwise correlation ACFs (not shown). Overall, the poor fit to some of the higher-order correlations can be explained by temporal fluctuations in the higher-order

interactions that are not reflected in the long-time-averaged pairwise correlations. Though we consider only a statistical model of these pairwise correlations, it may be worthwhile in later research to consider the insights that temporal fluctuations provide about voting behavior on the Court [32, 34, 41].

B.3 W-Nominate model

A standard approach from political science for modeling votes involves spatial voting models. In these models, voters are described by a point in a multidimensional preference space likewise occupied by the cases or legislative bills on which they vote. The distances in this space determine probabilistically the preferred vote for the voter. The goal is to uncover latent parameters for the preferences of voters and of the cases in order to best fit the voting record.

The original Nominate model proposed in reference [62] is still widely used today. We consider a variation thereof that called W-Nominate that includes a weight parameter for each preference dimension to account for variability in how strongly certain issues might influence the voter's decision.

The W-Nominate model is a classic spatial voting model that is used widely, with variations thereof, in political science [51, 62]. It is called a "spatial" voting model because it embeds voters and cases in a D-dimensional policy space where distances between voters' "ideal points," or preferences, and the location of the case determine how voters prefer to vote. The terminology used is that the voters prefer to maximize their "utility function." Utility is maximal when the vote in the case is precisely on top of the voters' ideal points and decreases with distance. In a minimal version of the model that ignores temporally fluctuating ideal points, different parametrizations of utility, etc., the utility function takes the form [62]

$$u_{\mathrm{ik,M}} = \beta \exp\left(- \sum_{\mathrm{d}=1}^{D} w_{\mathrm{d}}^2 M_{\mathrm{ikd}}^2 / 2 \right) \tag{B14}$$

$$u_{\mathrm{ik,m}} = \beta \exp\left(- \sum_{\mathrm{d}=1}^{D} w_{\mathrm{d}}^2 m_{\mathrm{ikd}}^2 / 2 \right) \tag{B15}$$

where M_{ikd} and m_{ikd} refer to the linear distance between the ith voter in the kth case for the dth policy dimension when voting with the majority (M) or minority (m). Every case has a policy midpoint z_{kd} and spread Δz_{kd} such that for a voter with ideal point θ_{id}

$$M_{\mathrm{ikd}} = \theta_{\mathrm{id}} - (z_{\mathrm{kd}} + \Delta z_{\mathrm{kd}}) \tag{B16}$$

$$m_{\mathrm{ikd}} = \theta_{\mathrm{id}} - (z_{\mathrm{kd}} - \Delta z_{\mathrm{kd}}) \tag{B17}$$

In more familiar terms, the utility function corresponds to the negative of the energy function where energy decreases exponentially with distance as we approach an individual's preferred orientation. Thus, the probability that a voter votes either with the majority or the minority is

$$p(\sigma_{\mathrm{ik}} = \mathrm{M}) = e^{u_{\mathrm{ik,M}}}/(e^{u_{\mathrm{ik,M}}} + e^{u_{\mathrm{ik,m}}}) \tag{B18}$$

$$p(\sigma_{\mathrm{ik}} = \mathrm{m}) = e^{u_{\mathrm{ik,m}}}/(e^{u_{\mathrm{ik,M}}} + e^{u_{\mathrm{ik,m}}}) \tag{B19}$$

Here, every voter votes independently, but correlations are induced by voters' relative distances in policy space.

In the language of machine learning and statistics, spatial voting is a kernel technique [9, 66, 72]. Eq 1.14 describes a form of Gaussian process regression with a radial basis kernel where the location of points in the space are parameters along with the weights w_{d} and inverse temperature β that are found by posterior maximization. This maximization procedure is implemented in an R-package as described in reference [63].

Spatial voting models are usually less parsimonious than maxent methods both in their derivation and parameterization. Although will not discuss this point in much detail here, minimal models are often favorable, especially from the Bayesian perspective, for generalization [50]. More specifically, the W-Nominate spatial voting model used here has number of parameters C_{nom} that scales with the number of dimensions D, the number of voters N, and the number of cases K as

$$C_{\mathrm{nom}} \sim N \times D + K \times D \times 2 \tag{B20}$$

$$= 36 \times 10 + 8738 \times 10 \times 2 \tag{B21}$$

$$\sim 10^5, \tag{B22}$$

whereas the pairwise maxent model has only

$$C_{\mathrm{maxent}} = 630. \tag{B23}$$

It is important to point out that the W-Nominate model fits each case individually whereas the maxent model presumes some average over all cases that have been observed. If were able to describe the distribution of cases $p(z_{\mathrm{kd}}, \Delta z_{\mathrm{kd}})$ in some systematic way, could compress the W-Nominate model drastically—at least in the number of parameters. For example, if $p(z_{\mathrm{kd}}, \Delta z_{\mathrm{kd}})$ could be described by a 10×10 covariance matrix (such that a multivariate normal distribution captures the distribution over policy dimensions), would reduce the number of parameters to roughly the same number as the pairwise maxent model. As show

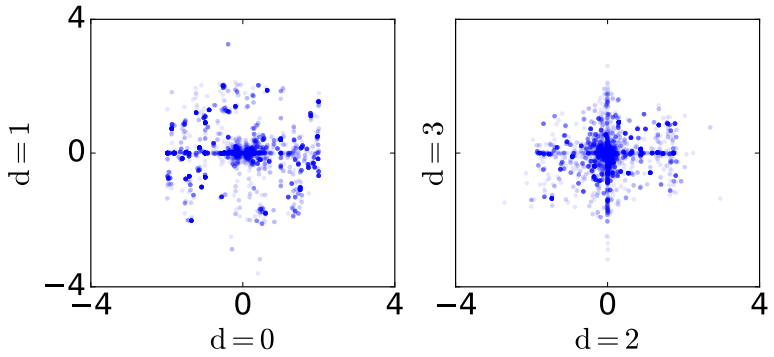


Figure B5: Policy dimensions z_{kd} for every case k in the data set as given by Eqs B16 and B17. (left) Principle policy dimensions are related in a complicated way showing that the W-Nominate model's policy dimensions cannot be compressed into a low-dimensional covariance matrix. Complicated features like seeming discontinuities at $z_{k0} = 0$ are apparent in the horizontal line of points cutting across the distribution. (right) The most correlated policy dimensions with correlation coefficient $\rho = 0.14$. Even this pair of policy dimensions does not follow some simple relationship.

in Figure B5, however, $p(z_{kd}, \Delta z_{kd})$ is complicated. When compare just two dimensions at a time, find that complicated, strongly non-Gaussian features characterize the distribution of cases. This means 10^3 parameters is an unrealistic lower bound for fully describing the W-Nominate model.

Although unanimous votes are a particularly notable feature in the pairwise maxent model, unanimous votes in the data are excluded from the training set for W-Nominate. Unanimous votes are degenerate in the sense that every voter has the same apparent policy preference. In this sense, they give no information about the relative positions of the voters in policy space (and can negatively impact accuracy) and so it is standard procedure to exclude them from the analysis. Yet, in what seems inconsistent, unanimous votes can be generated from sampling from the W-Nominate model, but are not explicitly ruled out from the sample space [63, 64]. Nevertheless, find that the unanimous vote is the most frequently occurring pattern for the W-Nominate model as well, but with probability of only $p_{unan} = 0.01$. For a fair comparison between the models, we explicitly include the probability of a unanimous vote since it is hidden from W-Nominate as detailed in Eqs 1.15 and 1.16.

C Appendix for Chapter 1.3

C.1 Median Voter Model (MVM)

The MVM, as described in the main text, consists of an odd number of N voters where a single voter, the Median, who with probability q always votes in the majority and $N-1$ Random voters who vote randomly. More generally, we can interpolate between a perfect median and set of N random voters by setting a probability γ that Median votes in the majority and $1-\gamma$ that Median votes randomly (either -1 or 1). The remaining $N-1$ voters are always random. This model presents a simple testing ground for exploring the information geometry of a system with a unique, statistically well-defined median voter that can be range from random ($\gamma = 0$) to perfect median ($\gamma = 1$).

The probability distribution defined by the MVM cannot be exactly captured by a pairwise maximum entropy (maxent) model. We recall that pairwise maxent models can be derived by maximizing the entropy of the model $S = -\sum_s p(s) \ln p(s)$ while constraining the single component $p(s_i)$ and pair component $p(s_i, s_j)$ distributions to match the data. From the maxent perspective, the MVM is equivalent to specifying that Median be perfectly correlated with the majority of Random voters, a correlation of the form,

$$\left\langle s_M \frac{\sum_{i=1}^{N-1} s_{R_i}}{\left|\sum_{i=1}^{N-1} s_{R_i}\right|} \right\rangle = 1. \tag{C24}$$

In general, this nonlinear correlation cannot be written as the linear combination of pairwise correlations. Thus, the MVM serves as a testing ground as a reduced statistical model of a median voter whose nontrivial correlations with the majority vote cannot be captured perfectly by a pairwise maxent model.

The pairwise maxent formulation of the MVM has the energy function

$$E(s) = -J_M \sum_{i=1}^{N-1} s_M s_{R_i} - \frac{J_R}{2} \sum_{i,j}^{N-1} s_{R_i} s_{R_j}, \tag{C25}$$

with couplings that connect M with every other R, J_M, and every pair of Rs with one another, J_R. The couplings are determined by the chosen value of q. As usual, the partition function is a sum over all configuration, but we can rewrite it as a sum over the number of voters in the majority k by using the fact that Rs

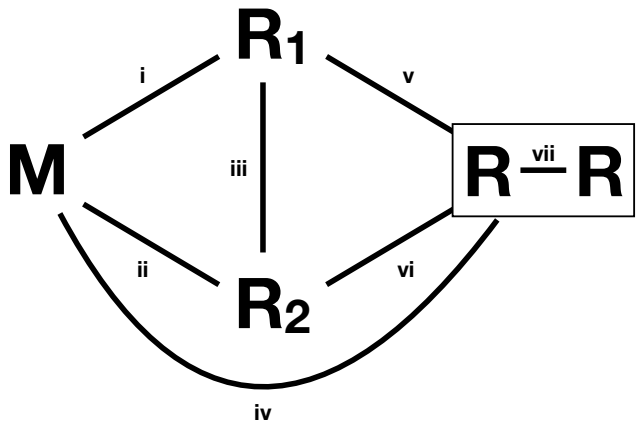


Figure C6: Minimal set of couplings required to calculate Fisher information matrix. Box on the right represents a fully-connected graph of Random voters all coupled to one another by J_{vii}.

are interchangeable:

$$Z = \sum_s e^{-E(s)} \tag{C26}$$

$$= \sum_k \binom{N}{k} Z_k = \sum_k \binom{N}{k} \exp(-E_k). \tag{C27}$$

The corresponding energy terms are then

$$\log E_k = \frac{k}{N} \exp\left(J_{\text{M}} \left[2k - N - 1 \right] + J_{\text{R}} \left[\binom{k-1}{2} + \binom{N-k}{2} - (k-1)(N-k) \right] \right) + \frac{N-k}{N} \exp\left(J_{\text{M}} \left[N - 2k - 1 \right] + J_{\text{R}} \left[\binom{N-k-1}{2} + \binom{k}{2} - (N-k-1)k \right] \right), \tag{C28}$$

accounting for the couplings between M and Rs voting with and against it, the couplings between Rs within the two opposing sides of Rs, and between the two opposed blocs of Rs.[4] This expression for the partition function can be evaluated in time linear with N, permitting us to solve the pairwise maxent model for the MVM efficiently.

Calculating the Fisher information matrix for $p(k)$ is more complicated because we must consider all the possible ways that the couplings split for each pair of voters we perturb. For example, if we were to make M like R′, then the coupling J_{M} connecting M with every other R must split into a coupling with R′, $J_{\text{MR}'}$ and a coupling with the remaining Rs, J_{MR}. Likewise, the coupling between R′ and

[4] As a simple check, one can see that the total number of couplings accounted for sums up to $\binom{N}{2}$.

other Rs will change to $J_{RR'}$ and the other Rs will be coupled to each other by J_{RR}. The other possible perturbation involves changing the relationship between a pair of Rs and not involving the special R′ we have picked out. To encompass all possible perturbations we can effect, we must expand the model to include seven couplings indexed by the roman numerals in Figure C6. Using this expanded model, we numerically solve the three different sets of changes to the couplings: $\Delta J_{M\to R}$, $\Delta J_{R\to M}$, and $\Delta J_{R\to R}$, where the first represents the changes in the couplings for the perturbation M→R as defined in Eqs C39 and the remaining ones are likewise defined. Then, we combine pairs of these perturbations to the couplings to calculate the entries of the Fisher information matrix as shown in Figure 1.16.

That there are only three possible perturbations that we can effect means that there are only six possible different values in the FIM (three along the diagonal and three off the diagonal). Once these are calculated, we can populate the entire FIM without any explicit calculation, using the combinations of indices involved in the rows and columns of the matrix. As a result, we can quickly construct the FIM for large systems and are then limited by numerical precision and not computation time.

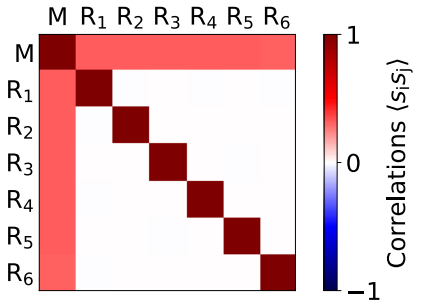
We calculate the eigenvalue spectrum for each voter subspace (the diagonal blocks of the FIM) as a function of N for a perfect median in Figure C8. The increasing importance of M is clear in the linear growth of its principal subspace eigenvalue $\lambda_M(N)$ compared with the increasingly slow growth of $\lambda_R(N)$. This increasing importance of M to the distribution of votes in the majority $q(k)$ mirrors the increasing divergence between the MVM distribution and a distribution of random voters. When the system size is small, the MVM distribution is similar to that of random voters, rendering M most similar to R voters. For larger systems, the unusual role of M becomes increasingly clear and it becomes the unique, pivotal voter in the set.

C.2 "Pivotal components" maxent examples

Across all the system that we study, we find that the pairwise maxent model captures well higher-order features of the data even when only fit to the pairwise correlations. To characterize this fit, we use a property of maxent models. The entropy of the maxent model always decreases with the inclusion of additional constraints such that entropy is largest when all N components are treated as independent $S_1 = \log_2 N$ bits and minimized when the entire probability distribution of the data is fit exactly $S_N = S_{\text{data}}$, where S_i is the entropy of the model

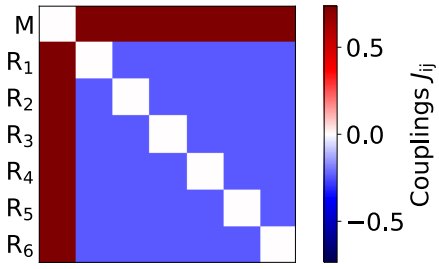
(A) Pairwise correlations

$$\langle s_{\rm i} s_{\rm j} \rangle = \sum_s p(s) s_{\rm i} s_{\rm j}$$



(B) Graph model

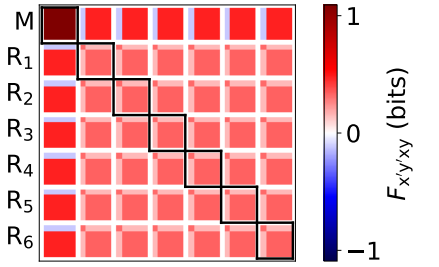
$$p(s; \{J_{\rm ij}\})$$



(C) Sensitivity matrix

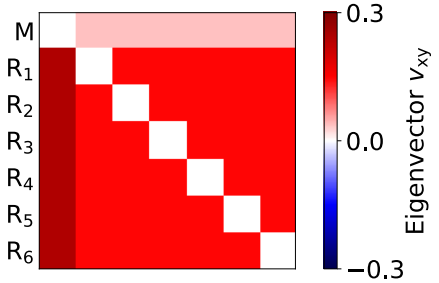
$$F_{\rm x'y'xy} = \lim_{\epsilon \to 0} \frac{2}{\epsilon^2} D_{\rm KL} \left[q_J || \tilde{q}_{\tilde{J}} \right]$$

$$\tilde{J} - J = \Delta_{\rm xy} J(\epsilon) + \Delta_{\rm x'y'} J(\epsilon)$$



(D) Eigenmatrix

$$F_{\rm x'y'xy} v_{\rm xy} = \Lambda v_{\rm xy}$$



(E) Pivotal measure & asymmetry

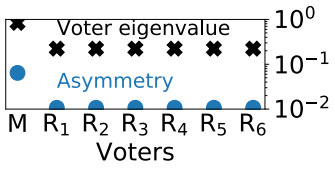


Figure C7: Overview of method for identifying pivotal voters for the $N = 7$ Median Voter Model to complement Figure 1.16. (A) Taking the matrix of pairwise correlations, (B) we solve a pairwise maxent model to learn the probability distribution $p(s; \{J_{ij}\})$ parameterized by the couplings J_{ij}. (C) We calculate the FIM for $q(k)$, the probability of k votes in the majority, measuring the sensitivity of $q(k)$ to changes in voter behavior. The perturbation to the vector of couplings determined by Eqs 1.22 and 1.23 are denoted as $\Delta_{xy}J$. The matrix is segmented into 6x6 blocks for readability. The variation in the entries of the FIM clearly indicates the unique role of the median. Each entry $F_{x'y'xy}$ of the FIM shows how quickly $q(k)$ changes when two pairs of voters (y becomes x and y' becomes x') are changed together. When at least one index is M, we find values different from when only R's are involved. (D) The principal eigenvector of the FIM v_{xy}, reshaped into an "eigenmatrix." (E) The asymmetry a_y measures the difference in perturbations localized to a specific voter vs. all its neighbors in turn. The principal subspace eigenvalues, computed from each outlined diagonal block in panel C, give our pivotal measure after normalization.

matching all correlations up to and including order i.[5] Thus, as we increase the number of parameters, we impose higher-order structure, and we monotonically approach the entropy of the data. This suggests as a measure for comparing maxent models, the total multi-information captured [45], which varies from 0 (no improvement beyond the independent model) to 1 (exact fit to the data). For the examples considered in the main text, the pairwise maxent model serves as an excellent fit, capturing over 94% of the multi-information in all cases. The pairwise maxent model captures over 98% of the multi-information for the $N = 7$ MVM. Overall, the pairwise maxent model is a minimal but convincing approximation of the ensemble statistics of the examples we consider [8, 52].

C.3 Specifying the Fisher information metric

If we have a statistical model described by a probability distribution $p(s; \theta)$ over a set of discrete states s and parameterized by parameters θ, how do we measure how different one model is from another? The Kullback-Leibler (KL) divergence is one such measure that tells us how much information is necessary to reach a

[5] For how to estimate the entropy of the data see references [45] or [7]. Calculating an unbiased estimate entropy of the entropy of the data can be an issue for sparse samples, but is straightforward for the relatively large number of samples we have.

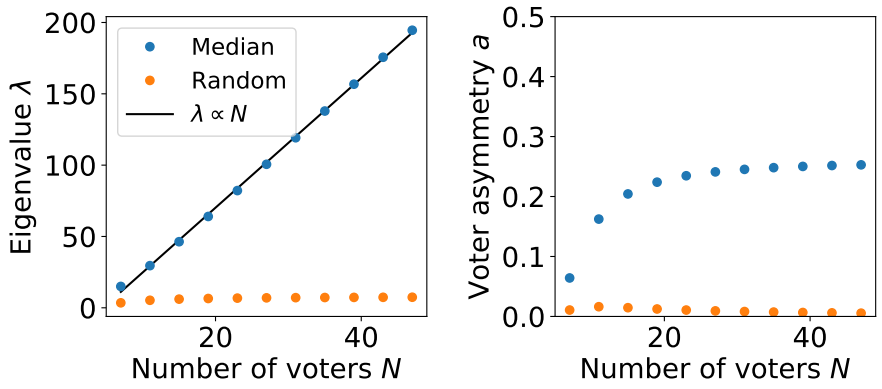


Figure C8: Voter eigenvalues and asymmetries as a function of system size for the MVM with $q = 1$. The subspace eigenvalue for Median grows linearly with system size while its asymmetry asymptotically dominates over each Random voters'.

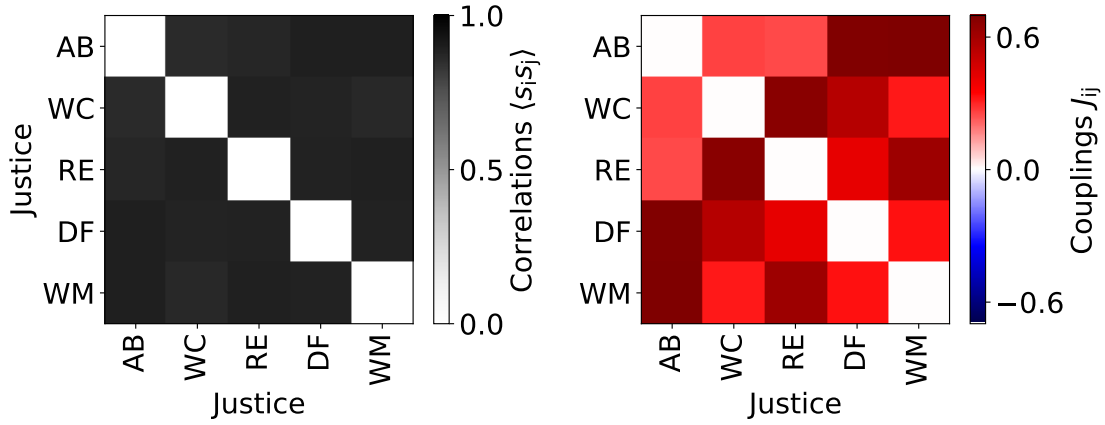


Figure C9: Pairwise correlations and couplings for AK Supreme Court.

distribution $p(s;\tilde{\theta})$ if we know $p(s;\theta)$ [20, 64],

$$D_{\mathrm{KL}}[p(s;\theta)\|p(s;\tilde{\theta})] = \sum_s p(s;\theta)\ln\left(\frac{p(s;\theta)}{p(s;\tilde{\theta})}\right). \tag{C29}$$

In the limit where the two distributions are infinitesimally close to one another, the KL divergence becomes a metric. The constant and linear terms go to zero, and the first nonzero term is the curvature of the divergence, the Hessian, which is also known as the Fisher information,

$$F_{ij} = \left.\frac{\partial^2 D_{\mathrm{KL}}[p(s;\theta)\|p(s;\tilde{\theta})]}{\partial\tilde{\theta}_i\partial\tilde{\theta}_j}\right|_{\tilde{\theta}_i=\theta_i,\tilde{\theta}_j=\theta_j}. \tag{C30}$$

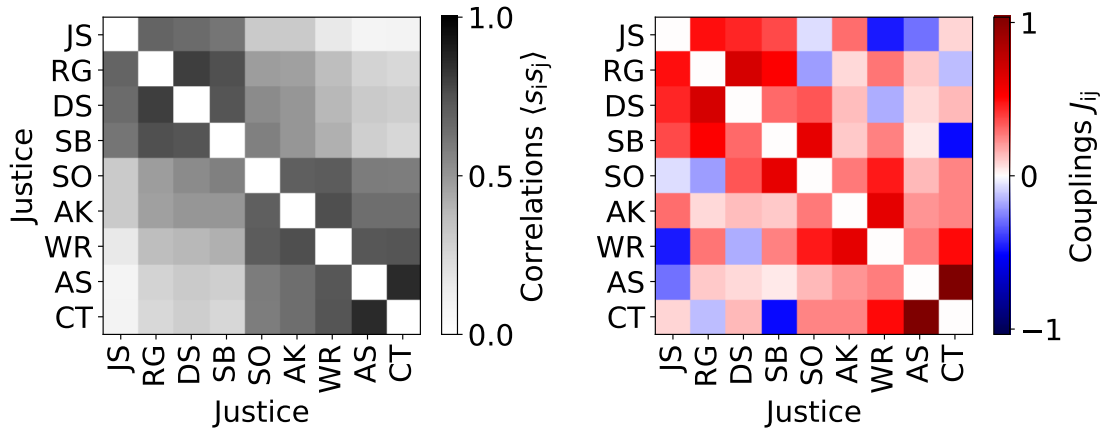


Figure C10: Pairwise correlations and couplings for US Supreme Court.

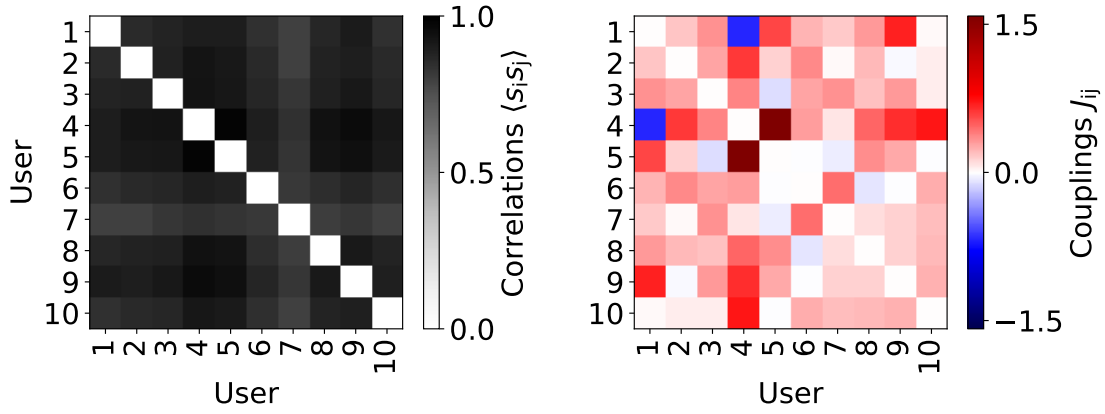


Figure C11: Pairwise correlations and couplings for Twitter K-pop community.

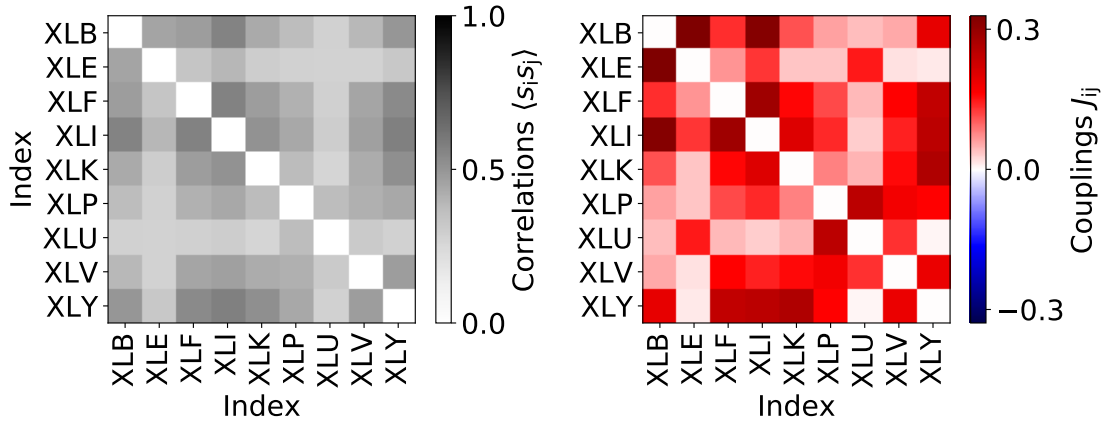


Figure C12: Pairwise correlations and couplings for SPDR economic sector indices.

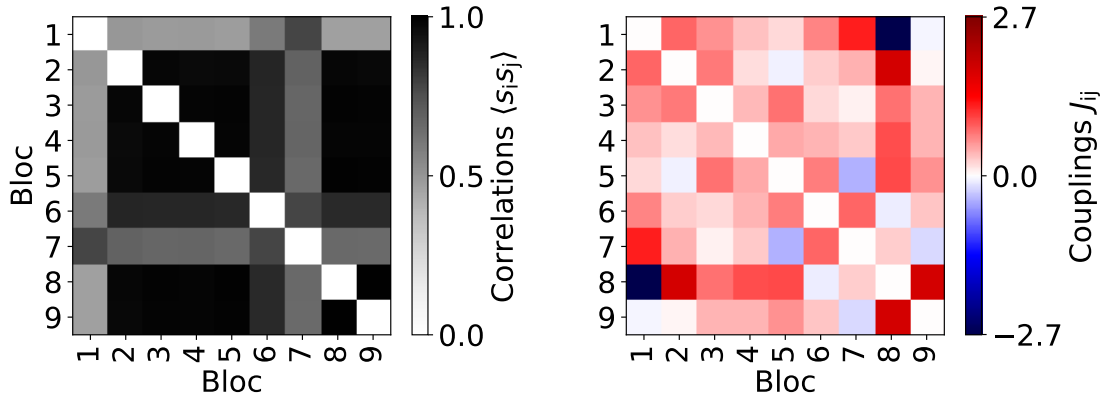


Figure C13: Pairwise correlations and couplings for CA Assembly 1999 session. Composition of blocs is given in Figure C21.

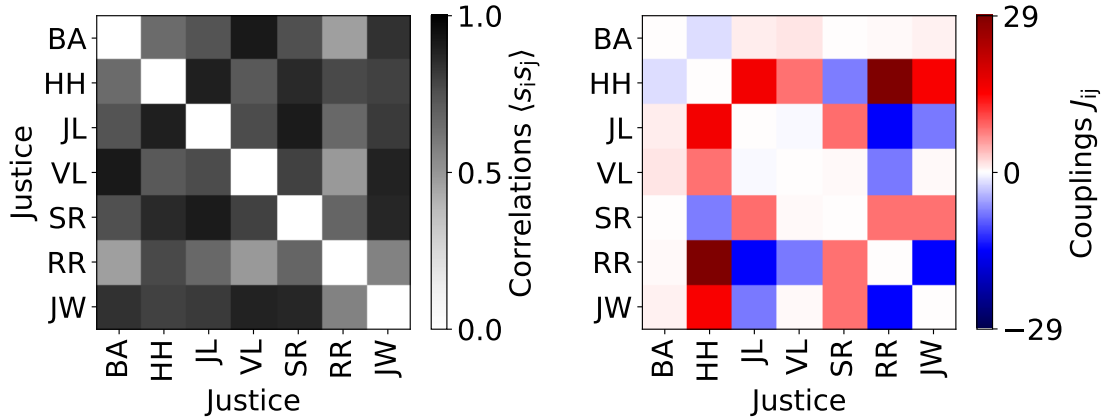


Figure C14: Pairwise correlations and couplings for NJ Supreme Court. The initials stand for Barry T. Albin, Helen E. Hoens, Jaynee, LaVecchia, Virginia Long, Stuart Rabner, Roberto A. Rivera-Soto, and John E. Wallace Jr.

Thus, the FI is a description of how quickly the probability distribution changes if we move along various directions in parameter space. Because it is a metric, the eigenvectors of the Hessian correspond to orthogonal directions in the tangent space of the model manifold, where the eigenvalues describe how quickly the manifold is varying along these directions.

The Fisher information also measures how much information about a parameter is in a random sample [20]. When a parameter is extremely sensitive to the distribution of data, then the information shared is high, whereas when it is fairly insensitive the information shared is low. This is described formally by the Cramér-Rao bound, which sets a lower bound on the precision of an unbiased estimator for the parameters [20]. This picture, more formally, has been used as technique for model reduction, removing degrees of freedom in parameter space to which the system is insensitive [80]. Here, we focus on the sensitive degrees of freedom, using them to identify components interesting because their behavior is precisely determined by the statistics of the data, or equivalently on whose behavior collective statistics depend the most sensitively.

We propose using as parameters aspects of the system that provide transparent insight into how perturbation of the parameters affects the system. In physical systems, it is natural to consider the couplings J_{ij}, or more generally the Lagrangian multipliers from the maximum entropy formulation, as the parameters by which to control the system because they are experimentally accessible (e.g., a magnetic system can be tuned by an applied field or by changing temperature that modulates all couplings by a factor). For a statistical model of a social system, however, the meaning of the terms in the energy functional are opaque and often nontrivial. In other words, we do not know how to access the coupling parameter J_{ij}. Certainly, we could be methodical about it, calculate the corresponding changes in the set of pairwise correlations for a perturbation in J_{ij}, but that would result in changes across all pairwise correlations in varying amounts. This change may be difficult to effect in a social system when opportunities for control are often limited.

This impracticality suggests a different approach, where we instead consider how the measured behavior of the system might be perturbed directly since these are straightforward to measure. This reasoning leads us to consider as parameters the observables. Formally, the observables for a maxent model are the conjugate variables to the Langrangian multipliers as given by the Legendre transform [85]. There is no difference in knowing one or the other: the transformation is a one-to-one mapping. For the pairwise maxent model, the "natural parameters" are the couplings and their conjugate the pairwise correla-

tions

$$J_{ij} \Leftrightarrow \left\langle s_i s_j \right\rangle \tag{C31}$$

$$-\ln Z(\{J_{ij}\}) + S(\{\left\langle s_i s_j \right\rangle\}) = \sum_{i=1}^{N-1} \sum_{j>i}^{N} J_{ij} \left\langle s_i s_j \right\rangle . \tag{C32}$$

Eq C32 states the well-known relation that the "free energy," $-\ln Z$, is the Legendre transform of the Shannon entropy, having set the units from Boltzmann's constant and temperature $k_B T = 1$. By working in the space of observables, we do not lose any information—indeed the model started with the observables in the first place—but find a more amenable representation.

In the main text, we take one further step by choosing to consider changes to the observables that are interesting as specified in Eq 1.22. For example, it is a common thought experiment in discussing Supreme Court voting to imagine how the system would change if the justices were different. The justices could be different in any which way, but we narrow the range of possible perturbations substantially by focusing on relative voting records, restricting ourselves to the range of behavior already observed in a system. It makes intuitive sense to ask how the Court would change if Justice Scalia were to vote more like Justice Thomas because their behaviors are specified by the voting record, but it requires much more work to determine what would happen if Scalia were to vote more like a judge picked from the appellate courts. There is no reason such a counterfactual could not be entertained in principle, but it would require modeling that judge's votes on the same set of cases that Scalia voted on. We restrict ourselves from considering such open-ended questions, leaving them as potential extensions of our work. Importantly, we choose perturbations that are localized to particular components, interpretable in their mapping to behavioral changes, and applicable across a wide range of systems.

Another advantage of treating observables as parameters is that it offers some independence from the choice of model. As a simple example of the distinction between treating an observable as the parameter or a term in the energy function, consider the biased coin. With probability p the coin flips heads and with probability $1-p$ it flips tails. If the parameter is the bare observable, the average coin flip, that is equivalent to changing p up to a constant factor.

$$\tilde{p} = p + \epsilon \tag{C33}$$

Taking the transformation in Eq C33, we calculate the Fisher information (FI) to find (Figure C15)

$$F = \frac{1}{\log 2}\left[\frac{1}{p} + \frac{1}{1-p}\right]. \tag{C34}$$

The Fisher information diverges at the boundaries of the parameter space $p = 0$ and $p = 1$ because that is where a finite change in p can lead to a diverging information distance. More closely with the perturbation considered in the main text, we could insist that the coin "mimic" a perfectly biased coin such that

$$\tilde{p} = \epsilon + (1 - \epsilon)p \tag{C35}$$

Under this scenario, the Jacobian captures the fact that the coin's bias makes no finite jump near $p = 1$, but changes ever more slowly when it is almost a perfectly biased coin. As a result,

$$F = \frac{1}{\ln 2}\left[\frac{1}{p} - 1\right], \tag{C36}$$

which goes to zero at $p = 1$.

Now, consider a maxent version of this problem which is the nonlinear transformation

$$p = [\tanh(x) + 1]/2 \tag{C37}$$

where the field determining the bias is x and

$$F = \frac{\operatorname{sech}\left(x^2\right)}{2 \ln 2}. \tag{C38}$$

In contrast with using p as the parameter, the quantity in Eq C38 peaks at $p = 1/2$ and decays to 0 at the boundaries, and an infinite change in x is necessary to reach $p = 0$ and $p = 1$. Of course, we could have chosen any possible model, choosing instead of the maxent transformation in Eq C37, our favorite nonlinear transformation. Thus, by choosing our favorite model, we would end up effectively specifying the FI. This is generally not an issue if one cares about measuring the relationship between a particular model and the data, but it does become an issue if one cares more about the statistics of the data rather than of the particular model specified.

If we restrict ourselves to perturbing p, we ensure that the choice of perturbation does not depend on the choice of model. Additionally when the probability distribution is matched exactly, there is no dependence on the model class. In the special case of the biased coin, the probability distribution is specified exactly by a single parameter. Assuming we can measure it with infinite precision and we have a model that can fit the measured p exactly (e.g., a model limited to $0 \leq p \leq 1/2$ does not count), the calculation of FI—whether Eq C34, C36, or some other choice—is concretely defined. More generally, data resolution is

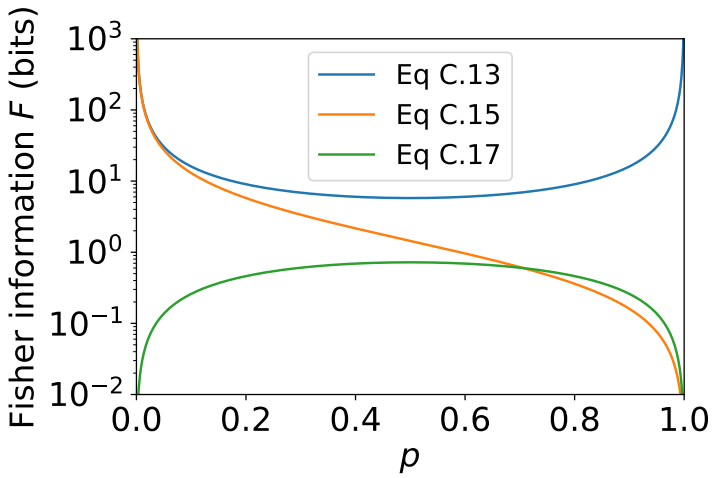


Figure C15: Fisher information for biased coin according to different choices of perturbation: changing p directly (Eq C34), substitution with a biased coin (Eq C36), and changing the "field" (Eq C38).

not perfect and so we must infer probabilities for configurations that we have not observed. As an example, the pairwise maxent approach assumes that the pairwise marginal distributions are known exactly, but the higher-order joint probabilities are assumed to conform to the maxent principle. When any such model feature is used in the calculation of the FI, clearly the FI will depend on the assumptions of the model. For the pairwise maxent model, this means that the calculation of FI on features of the distribution that are constructed explicitly from pairwise marginals matches the data exactly, but in general the distribution of majority-minority divisions does depend on the maxent assumptions. This is not always unfavorable: if the higher-order terms decrease in importance such that they behave as small perturbations on top of the pairwise model, the FI will show weaker dependence on these corrections [52]. Thus, we propose a generalizable approach that links minimal, maxent models with a simple class of perturbations defined on the range of relative component behavior.

C.4 Calculation of the Fisher information matrix (FIM)

Here, we calculate the FIM for the transformation described in Eqs 1.22 and 1.23 and go through some examples to show how to calculate the FIM. We go into some detail with the derivation to make clear how to perform such a calculation for those less familiar with maxent models and information geometry.

In Eqs 1.22 and 1.23, we consider how the correlations between component y and all other components x′ change when component y appears to vote more like component x. To effect this perturbation, we use a parameter $\epsilon \to 0$ that

leads to a linear change in the couplings $J_{x'y'}$ as described by the rate of change $dJ_{x'y'}$, where we are taking a total derivative with respect to the change in the pairwise probabilities described by the vector $r(s_x = s_y)$. To obtain this derivative, we perturb to first order in ϵ the expression for the pairwise correlations to obtain the self-consistent equation

$$\left\langle s_x s_y \right\rangle_{\text{pert}} - \left\langle s_x s_y \right\rangle = \frac{1}{2} \sum_{x',y'}^{N} \Delta_{xy} J_{x'y'} \left(\left\langle s_x s_y s_{x'} s_{y'} \right\rangle - \left\langle s_x s_y \right\rangle_{\text{pert}} \left\langle s_{x'} s_{y'} \right\rangle \right). \tag{C39}$$

By self-consistent, we are referring to the fact that the new pairwise correlations after perturbation $\left\langle s_i s_j \right\rangle_{\text{pert}}$ depend on the change in the couplings $\Delta_{xy} J_{xy}$, so the perturbation to the couplings determine quantities on both sides of Eq C39.[6] The coupling perturbations, $\Delta_{xy} J_{x'y'}$, are related to the linear response of the couplings to change in the collective statistics induced by perturbing the pair of components x and y,

$$dJ_{x'y'} = \lim_{\epsilon \to 0} \Delta_{xy} J_{x'y'} / \epsilon. \tag{C40}$$

The resulting matrix of new couplings is

$$\tilde{J}_{x'y'} = J_{x'y'} + \Delta_{xy} J_{x'y'} \tag{C41}$$

$$\tilde{p}(s; \{\tilde{J}_{x'y'}\}) = p(s; \{J_{x'y'} + \Delta_{xy} J_{x'y'}\}). \tag{C42}$$

The set of the perturbations $\Delta_{xy} J \equiv \{\Delta_{xy} J_{x'y'}\}$ appear in Figure C7C.

Eqs C39–C42 describe the numerical algorithm for calculating the changes in the statistics of the system under the perturbation described in Eqs 1.22 and 1.23. Note that the algorithm implicitly depends on ϵ, which must be taken to zero. The remaining calculations are to coarsen the full distribution $p(s)$ to $q(k)$, the distribution of k votes in the majority and to calculate the FIM on q. For pedagogical clarity, we will first show how to calculate the FIM without coarse-graining.

There is a simple, intuitive form for the FIM for maxent models. Under an infinitesimal change in the parameters such that the energy of each voting con-

[6] Another way to describe Eq C39 is as the linear combination of the linear response functions of every pairwise correlation to a change in the corresponding coupling, also known as the susceptibility.

figuration $E(s) \to E(s) + \Delta E(s)$, we can expand

$$\begin{aligned}
\tilde{p}(s) &= \frac{e^{-E(s)-\Delta E(s)}}{\sum_{s'} e^{-E(s')-\Delta E(s')}} \\
&\approx \frac{e^{-E(s)}\,(1-\Delta E(s))}{\sum_{s'} e^{-E(s')}\,(1-\Delta E(s'))} \\
&\approx p(s)[1-\Delta E(s)][1+\langle \Delta E\rangle] \\
&= p(s)[1+\langle \Delta E\rangle - \Delta E(s)] + O(\Delta E^2)
\end{aligned} \tag{C43}$$

Now calculating the Kullback-Leibler divergence to second order,

$$\begin{aligned}
D_{\mathrm{KL}}[p\|\tilde{p}] &= \sum_s p(s)\left[\ln p(s) - \ln \tilde{p}(s)\right] \\
&= \frac{1}{2}\left\langle (\Delta E - \langle \Delta E\rangle)^2 \right\rangle + O(\Delta E^3).
\end{aligned} \tag{C44}$$

The constant term is zero because the $D_{\mathrm{KL}}[p\|p] = 0$ and the linear term is zero because the changes to the probability function under perturbation must sum to zero to preserve normalization. The FI is the second order term, or the curvature, so we must send the norm change in the energy $|\Delta E| \to 0$,

$$F = \lim_{\epsilon\to 0} \epsilon^{-2}\left(\langle \Delta E\rangle^2 - \langle \Delta E\rangle^2\right), \tag{C45}$$

where ϵ is defined in Eq C40. For the symmetrized pairwise maxent model class, ΔE is the sum over all couplings. Thus, the FI of the distribution over the full state space $p(s)$ has a simple form in terms of the change in energy for maxent models.

Alternatively, the result from Eq C45 can be expressed as a matrix of correlation functions. In other words, the correlation functions are the linear response functions for perturbations to the natural parameters, here the couplings J_{xy}. As the simplest example, consider an Ising model under perturbation to a particular coupling J_{xy}. Using our form Eq C45 for the FI and $\Delta J_{\mathrm{xy}} \equiv \tilde{J}_{\mathrm{xy}} - J_{\mathrm{xy}}$,

$$F_{\mathrm{xyxy}} = \lim_{\Delta J_{\mathrm{xy}}\to 0} \Delta J_{\mathrm{xy}}^{-2}\left\langle (\Delta J_{\mathrm{xy}} s_{\mathrm{x}} s_{\mathrm{y}} - \Delta J_{\mathrm{xy}}\left\langle s_{\mathrm{x}} s_{\mathrm{y}}\right\rangle)^2\right\rangle. \tag{C46}$$

The perturbations to the couplings ΔJ_{xy} do not depend on the state s, so they can be pulled out of the averages to obtain

$$= \left\langle s_{\mathrm{x}}^2 s_{\mathrm{y}}^2\right\rangle - \left\langle s_{\mathrm{x}} s_{\mathrm{y}}\right\rangle^2 \tag{C47}$$

$$= 1 - \left\langle s_{\mathrm{x}} s_{\mathrm{y}}\right\rangle^2. \tag{C48}$$

The diagonal entries of the FIM are the variance of the pairwise correlation, which is a well-known result. It is straightforward to see that the off-diagonal elements of the FIM are the covariance $\left\langle s_x s_y s_{x'} s_{y'} \right\rangle - \left\langle s_x s_y \right\rangle \left\langle s_{x'} s_{y'} \right\rangle$. Thus, the FI for maxent models reduces to the covariance of the set of observables chosen as constraints, when we are dealing with natural parameters.

As a more general formulation, consider the set of Lagrangian multipliers θ_n and their corresponding bare observables $f_{\theta_n}(s)$ ("bare" referring to the fact that we have yet to dress them with brackets by averaging over the ensemble). For the pairwise maxent model, the Lagrangian multipliers are the couplings and the bare observables are the pairwise products $s_x s_y$. Working through the same calculation as before but with this general formulation of a maxent model, we find for the FI,

$$F = \lim_{\epsilon \to 0} \epsilon^{-2} \left\{ \sum_n \Delta\theta_n^2 \left[\left\langle f_{\theta_n}^2 \right\rangle - \langle f_{\theta_n} \rangle^2 \right] + \sum_{n,m} \Delta\theta_n \Delta\theta_m \left[\langle f_{\theta_n} f_{\theta_m} \rangle - \langle f_{\theta_n} \rangle \langle f_{\theta_m} \rangle \right] \right\}, \tag{C49}$$

where the $\Delta\theta_n$ depend implicitly on ϵ. As noted earlier, the perturbation in the pairwise agreement probabilities leads to a nontrivial combination of changes to the entire vector of couplings. As a result, the FI in Eq C49 contains cross terms between all pairwise correlations and the change in the Langrangian multipliers $\Delta\theta_n$ each come with a factor of the Jacobian relating changes in the pairwise marginals to the couplings as described by Eq C39.

For the analysis in the main text, however, there is an additional step. We do not consider the full state space, but coarse-grain each $p(s)$ to the distribution of k votes in the majority $q(k)$. As a result, we are not calculating the variance in the energies for the pairwise maxent model as described in Eq C45, but the variance in the logarithm of the sum of all terms in the partition function with k voters in the majority. We label the set of all states with k voters in the majority S_k to write

$$q(k) = \sum_{s \in S_k} p(s) = \frac{1}{Z} \sum_{s \in S_k} e^{-E(s)} = \frac{1}{Z} e^{-E_{\text{maj}}(k)} \tag{C50}$$

$$E_{\text{maj}}(k) \equiv -\ln \left(\sum_{s \in S_k} e^{-E(s)} \right). \tag{C51}$$

Eq C51 defines an effective "k majority" energy such that under perturbation to

the pair of components x and y as indicated by Δ_{xy}

$$F_{xy} = \lim_{\epsilon \to 0} \epsilon^{-2} \left(\left\langle \Delta_{xy} E_{maj}^2 \right\rangle - \left\langle \Delta_{xy} E_{maj} \right\rangle^2 \right). \tag{C52}$$

Eq C52 is the form that the limit in Figure 1.16C takes.

To summarize the algorithm, we first find the total derivative of each coupling $dJ_{x'y'}$ with respect to the change in the pairwise marginals as explained in Eq C40. Then, we calculate the change in the distribution of k votes in the majority for both the model without the perturbation $q(k)$ and with $\tilde{q}(k)$ for a range of small values ϵ as in Eq C44. By comparing these two distributions for increasingly smaller ϵ, we estimate numerically the FIM, relying on the definition of an "effective" energy E_{maj} as in Eq C51 to deal with issues in numerical precision that may arise when comparing ratios of floating point numbers. These steps generate the FIM as shown in Figure 1.16C with which we calculate the eigenvalue spectrum to measure our pivotal voters.

C.5 Dissenting coalitions

In Figure 1.17, we project the eigenvectors onto the probabilities of dissenting coalitions to obtain a detailed picture of how the parameter directions obtained from the FIM affect dissenting coalitions. Such a projection involves taking the sum over all the probabilities of the states with the particular dissenting bloc and calculating the effective energy. Expanding the log-likelihood to first order, we calculate the rate at which this probability changes to be

$$d \ln q(k) = \lim_{\epsilon \to 0} \frac{1}{\epsilon} \left(\Delta E_{maj}(k) - \left\langle \Delta E_{maj}(k) \right\rangle \right). \tag{C53}$$

The limit $\epsilon \to 0$ refers to an infinitesimal perturbation of $q(k)$ along the first eigenvector of the FIM. Then, the rate of change in the log-likelihood simplifies to comparing the change in the effective majority energy $E_{maj}(k)$ with the average change across $p(s)$.

C.6 Measure of asymmetry

As a way of measuring the heterogeneity between the components of a system, we calculate the asymmetry of the "eigenmatrices" of the Fisher information matrix as defined in Figure 1.16D. Given that the matrices are normalized $\sum_{xy} v_{xy}^2 = 1$, the total asymmetry can be written

$$A = \frac{1}{2} \left(1 - \sum_{xy} v_{xy} v_{yx} \right). \tag{C54}$$

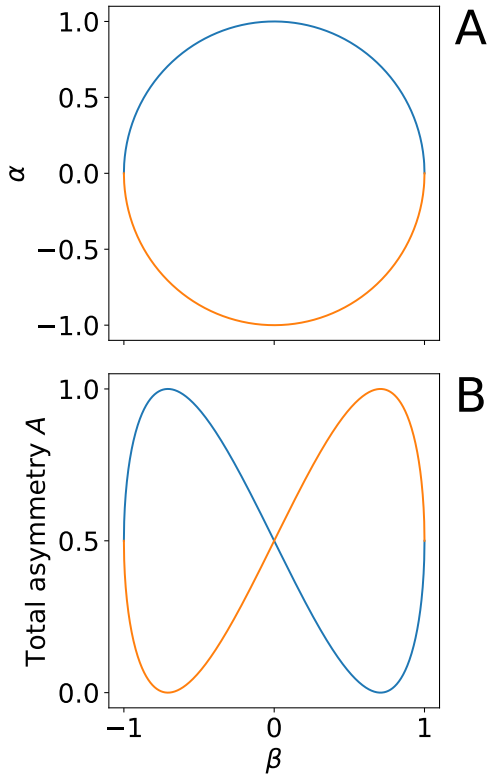


Figure C16: Total asymmetry for the binary system specified by the 2x2 matrix in Eq C55.

Thus, we might think of the asymmetry as a measure of correlation between the entries in the upper triangle and lower triangle of the eigenmatrix. When they are perfectly correlated such that the matrix is symmetric, $A = 0$. If the entries v_{xy} are completely uncorrelated with their partners in the transpose v_{yx}, then $A = 1/2$. When they are anti-correlated, the summation in Eq C54 can become negative and $A > 1/2$.

As an example, consider the asymmetry for a 2x2 matrix

$$\begin{pmatrix} 0 & \alpha \\ \beta & 0 \end{pmatrix} \tag{C55}$$

The normalization constrains α and β to the unit circle,

$$\alpha = \pm\sqrt{1-\beta^2}. \tag{C56}$$

In Figure C16A, we have colored the upper and lower halves of this circle (for positive and negative values of α) by different colors. Now calculating the total

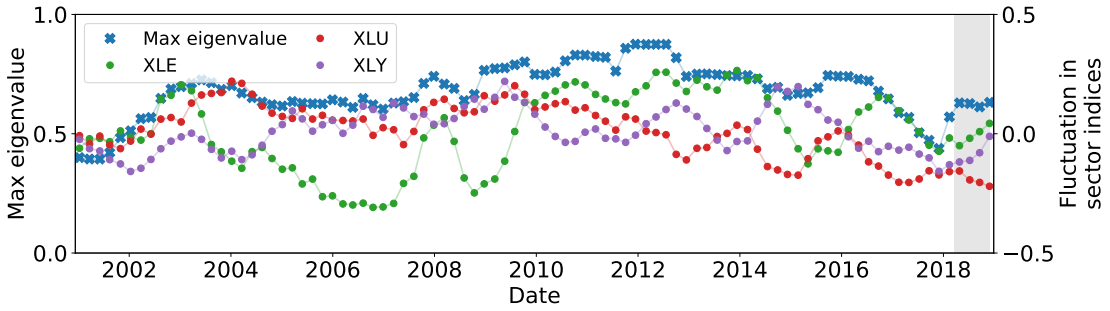


Figure C17: Retrospective time series analysis of the SPDR for the two most pivotal indices XLE and XLU and least pivotal XLY. We use a moving windowed window of duration $t = 256$ days and a shift of $\Delta t = 50$ days. The width of the moving window is delimited by the gray box. We compare the maximum of the normalized eigenvalues of the covariance matrix with η^*, the projection of the windowed time series fluctuation onto the stock index principal subspace eigenvector (Eq C58). Lines are drawn for readability.

asymmetry,

$$A = \frac{1}{2} \mp \frac{1}{2}\beta\sqrt{1-\beta^2}. \tag{C57}$$

We plot Eq C57 in Figure C16B and again color the curves differently depending on the half of the unit circle that we are tracing out. When $\beta = -1$, normalization asserts that $\alpha = 0$ and the total asymmetry $A = 1/2$. As we increase β, we can follow α along the positive (negative) route which leads to maximization (minimization) of A at $\beta = -1/\sqrt{2}$. As we keep increasing to $\beta = 0$, we return to $A = 1/2$ and have effectively swapped the roles of $\alpha \to -\beta$ and $\beta \to \alpha$ (a rotation of the matrix in Eq C55 by $-\pi/2$).

C.7 Time series analysis of SPDR

The temporal fluctuations of the sector indices in the SPDR represent potentially useful information about changing economic conditions. As a preliminary demonstration of the type of analysis that may be interesting in this context, we consider a retrospective analysis of how local temporal fluctuations in the market are reflected in the subspace eigenvectors of the FIM.[7]

To do this, we take a long temporal window $t = 256$ days that allows us to obtain a precise estimate of the distribution of configurations in a time window $p_{\text{win}}(s)$.

[7]Note that the subspace eigenvectors are calculated from the entire time series available, whereas realtime analysis would rely only the statistics available up to the current time. This is why we call this example "retrospective."

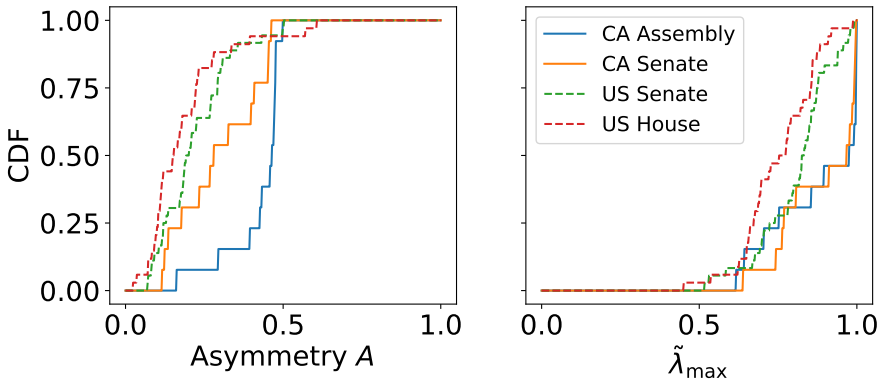


Figure C18: Comparison of the cumulative distribution functions (CDF) of (left) the total asymmetry A and (right) the dominant pivotal measure across CA state legislatures and the US House of Representatives and Senate. We compare the distributions of $\tilde{\lambda}_{\max}$ in Figure C19.

Then, we minimize the KL divergence between p_{win} and the pairwise maxent model solved on the entire data set with change in the couplings $\tilde{J}_{ij} = J_{ij} + \eta \Delta J_{ij}$ constrained to be along the principal stock index subspace eigenvector ΔJ_{ij} by adjusting the coefficient η,

$$\eta^* = \arg\min_{\eta} \sum_s p_{\text{win}}(s) \ln\left(\frac{p_{\text{win}}(s)}{p(s;\{J_{ij} + \eta \Delta J_{ij}\})}\right). \tag{C58}$$

The magnitude of this coefficient $|\eta^*|$ is a measure of how strongly the fluctuations in the windowed time series are reflected in the direction of parameter space specified by the subspace eigenvector. As we show in Figure C17, the fluctuations show patterns that diverge at many points from the maximum of the normalized eigenvalue of the windowed covariance matrix, a measure used to determine when economic conditions are changing [10]. In particular, we note that periods of time where the best fit value η^* between the various stock indices are correlated or anti-correlated may be useful indicators. Although it remains to relate these patterns to recognized features of the time series, this presents a potentially useful complement to existing tools for analyzing market data.[8]

C.8 Comparison of CA state and federal legislatures

Are institutional differences captured in our measures of the eigenvectors of the FIM? As an example, we compare the CA Assembly and Senate with the

[8]We do not discuss in detail here the difficulty of estimating information quantities in the limit of small data, an important issue for realtime forecasting of changing economic conditions. Entropy estimation for small samples remains an active research problem [7, 55], and we avoid this issue by taking long windows.

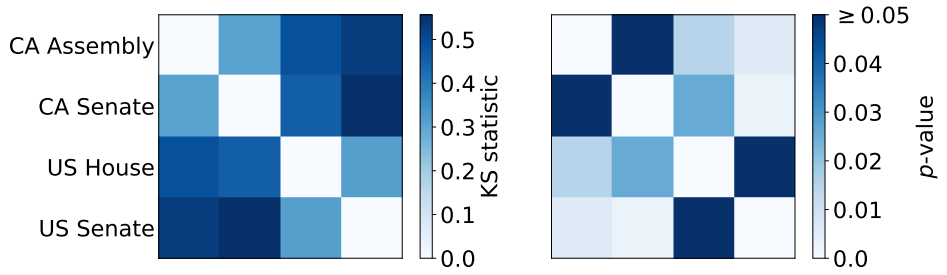


Figure C19: Results of Kolmogorov-Smirnov test on the dominant pivotal measure across CA state legislatures and the US House of Representatives and Senate.

US House of Representatives and Senate using our measures of the dominant pivotal measure and total asymmetry defined in Figure C18.

When we inspect the distribution of the principal pivotal component $\tilde{\lambda}_{\max}$, we find significant distinction between the state vs. federal levels. With the Kolmogorov-Smirnov (KS) test—testing whether or not the largest difference between the two sample cumulative distribution functions (CDF) rules out a coincident underlying distribution—summarized in Figure C19, we show the KS statistics to be larger and the p-values smaller between state and federal levels. Thus, the distribution of the principal pivotal measure is one way of distinguishing between the different levels of legislature. Notwithstanding further questions about the choice of coarse-graining—which was implemented following the same procedure outlined for the CA legislatures in Appendix E.1—our findings suggest that the structure of CA state legislatures makes them more conducive to the presence of pivotal voting blocs that precisely determine the distribution of majority-minority coalitions.

After measuring the total asymmetry for all the sessions we solved, we find that the distributions of total asymmetry for the House and the Senate are strongly similar, but those of the CA Assembly and CA Senate are not. We find that the large difference between the House and the CA Senate and between the Senate and the CA Senate is not statistically significant, given the number of data points. However, all three are significantly different from the CA Assembly with $p < 0.05$, suggesting that some systematic, institutional pattern might be gleaned from inspecting total asymmetry.

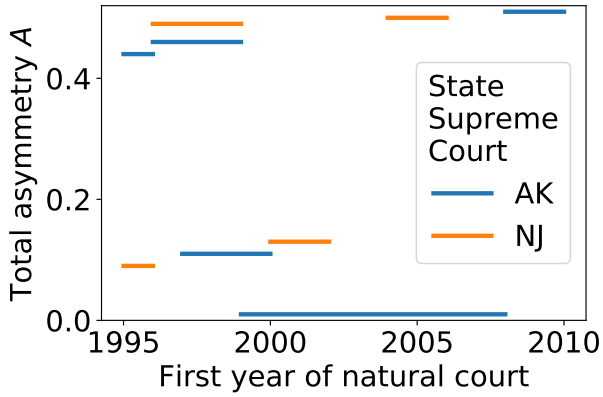


Figure C20: Total asymmetry for the natural courts in the AK and NJ Supreme Courts. We show the calculated asymmetry with lines spanning the first to last years on record for a full vote (including every sitting justice). There is overlap between natural court years because some of the data are mislabeled and show justices participating in votes after their official date of retirement.

C.9 Additional notes on data sets

US state supreme courts

We obtained the latest data set from the State Supreme Court Data Project (SSCDP) and used their binary coding of justice votes [49].

We show the total asymmetry for all the natural courts on the Alaska and New Jersey Supreme Courts we considered in Figure C20. We only considered natural courts with at least 100 where the full complement of justices were voting. As we mention in the main text, there is variation in the measured value of asymmetry that makes it unclear whether or not there is relationship between the total asymmetry and the codified institutional rules of voting.

SCOTUS

We use data from the Supreme Court Database Version 2016 Release 1, taking their binary coding of majority-minority votes [75]. This same data set and version has been analyzed previously. See references [44, 45].

C.10 SPDR

The SPDR Select Sector indices track the Standard & Poor's (S&P) and Morgan Stanley Capital International (MSCI) Global Industry Classification Standard (GICS) sectors. As described on Wikipedia, GICS "is an industry taxonomy developed in 1999 by MSCI and S&P for use by the global financial community.

Bloc 1	Bloc 2	Bloc 3	Bloc 4	Bloc 5	Bloc 6	Bloc 7	Bloc 8	Bloc 9
Migden	Corbett	Knox	Cardenas	Dutra	Reyes	Robert Pacheco	Olberg	Brewer
Aroner	Cedillo	Hertzberg	Lowenthal	Mazzoni	Havice	Leach	Strickland	Ashburn
Bock	Keeley	Torlakson	Ducheny	Wright	Florez	Dickerson	Campbell	Leonard
Romero	Firebaugh	Strom-Martin	Vincent	Nakano	Cunneen	Cox	Briggs	Baugh
Kuehl	Villaraigosa	Alquist	Thomson	Wayne	Maldonado	Maddox	Runner	Ackerman
Longville	Washington	Gallegos	Davis	Papan	Pescetti	Rod Pacheco	Aanestad	Thompson
Wildman	Wiggins	Calderon	Lempert	Machado	Granlund	Battin	Oller	Baldwin
Shelley	Honda	Wesson	Scott	Cardoza	Zettel	Margett	House	Kaloogian
Steinberg		Jackson		Correa		Bates		McClintock

Figure C21: Names of congressmen and congresswomen in 1999 CA Assembly session by voting bloc as determined by ranking on first W-Nominate dimension. Though all members were included for the W-Nominate analysis, only members who voted in more than 20% of the recorded votes were included for the coarse-graining and maxent solution.

The GICS structure consists of 11 sectors, 24 industry groups, 69 industries and 158 sub-industries into which S&P has categorized all major public companies. The system is similar to ICB (Industry Classification Benchmark), a classification structure maintained by FTSE [Financial Times Stock Exchange] Group."

We focus on these assets and their adjusted price action because (1) they are the most heavily-traded and representative sector assets in the world, so their prices and volumes reflect actual interest in exposure to the sectors, (2) they have been traded daily without exception for over 20 years, and (3) unlike the Dow indices, the S&P indices are not subject to effects of price-weighting such as reverse-split over-weighting. The historical price data is available online on Yahoo! Finance.

C.11 Twitter

We analyze one of the communities from the data considered in reference [33]. In this work, the authors divide the Twitter community into smaller subcommunities using the CNM algorithm [16]. We take one example from their K-pop community with 10 individuals.

C.12 CA Assembly and Senate

Session records were obtained from Prof. Jeff Lewis' scrape of the CA legislature's public data API [47]. For all sessions from 1993–2017, we solved the W-Nominate model using the code provided in reference [63]. We then removed any voter who did not participate in more than 20% of the votes, rank-ordered the voters by the first W-Nominate dimension, and divided them as equally as possible into 9 groups as shown for the 1999–2000 session in Figure C21.

For the results of bootstrap sampling to calculate error bars, we found that 3% of the Assembly samples showed significant error from the fit correlations because of numerical precision issues. This is generally an issue for systems that are poised near the boundaries of the model manifold where the couplings become large. For the error bars on the normalized subspace eigenvalues, however, the contribution from these three missing samples is negligible.

In Figure 1.19, the most pivotal bloc that we observe, Bloc 8, is constituted of Republicans whilst the State Assembly's majority is held by the Democratic party. Indeed, we find that the mutual information between the vote of Bloc 8 with the majority vote across the blocs is $I_8 = 0.96$ bits versus that of a Democratic Bloc 1, $I_1 = 0.17$ bits. This measure of correlation indicates that this Republican bloc is, like a median, highly predictive of the majority outcome across all of these blocs and, additionally, is pivotal.[9]

C.13 US House of Representatives and Senate

Data was obtained from Voteview [48]. We analyze the 80–113th Senate sessions and the 80–115th House sessions. The 80th congressional session started in 1947 and each session lasts two years.

For filtering voters and coarse-graining, we used the same procedure as specified by the CA Assembly above.

[9] Examples of blocs that are predictive of outcome but not pivotal include Bloc 3 ($I_3 = 0.96$ bits but $\tilde{\lambda}_3 = 0.02$) and A.K. and S.O. on SCOTUS [45].

D Appendix for Chapter 2.1

D.1 Empirical Methods & Study System Description

Study system

In this section, we provide details on our empirical study system. The data were collected by JCF in 1998 from a large group of captive pigtailed macaques (*Macaca nemestrina*) socially-housed at the Yerkes National Primate Center in Lawrenceville, Georgia. Pigtailed macaques [12] are indigenous to south East Asia and live in multi-male, multi-female societies characterized by female matrilines and male group transfer upon onset of puberty. Pigtailed macaques breed all year. Females develop swellings when in Œstrus. Macaque societies more generally are characterized by social learning at the individual level, social structures that arise from nonlinear processes and feed back to influence individual behavior, frequent non-kin interactions and multiplayer conflict interactions (reviewed in reference [26]).

The study group contained $n = 64$ non-infant individiuals (adults, subadults and juveniles) and 84 individuals in total. The study group had a demographic structure approximating wild populations and subadult and adult males were regularly removed to mimic emigration occurring in wild populations. All individuals, except 8 (4 males, 4 females), were either natal to the group or had been in the group since formation. The group was housed in an indoor-outdoor facility, the outdoor compound of which was 125 x 65 ft.

Data on social dynamics and conflict were collected from this group over a stable, four month period. Operational definitions are provided below in the next section.

Operational definition of a fight

Fight: includes any interaction in which one individual threatens or aggresses a second individual. A conflict was considered terminated if no aggression or withdrawal response (fleeing, crouching, screaming, running away, submission signals) was exhibited by any of the conflict participants for *two minutes* from the last such event. A fight can involve multiple individuals. Third parties can become involved in pairwise conflict through intervention or redirection, or when a family member of a conflict participant attacks a fourth-party. Fights in the data set analyzed here ranged in size from 2 to 35 individuals, counting only the socially-mature animals. Fights can be represented as small networks that grow and shrink as pairwise and triadic interactions become active or terminate until there are no more individuals fighting under the above described two minute

criterion. In addition to aggressors, a conflict can include individuals who show no aggression or submission (e.g., third-parties who simply approach the conflict or show affiliative/submissive behavior upon approaching, and recipients of aggression who show no response to aggression (typically, threats) by another individual). Because conflicts involve multiple actors, two or more individuals can participate in the same conflict but not interact directly.

In this study only information about the number of participants in a fight and duration are used. Only fights that included two or more socially-mature individuals were used in the analysis; the data includes $N = 1{,}086$ such fights. We do not consider internal aspects of the fight, such as who does what to whom, except for the order of each individual's first involvement in the fight (used to estimate time-ordered conditional probabilities for use in the dynamical branching process model). Time data were collected on fight onset and termination but only total duration data are used in these analyses.

Data collection protocol

The data were collected by Jessica C. Flack using all-occurrence sampling using a voice recorder and digital stop watch and working from an observation tower. All individuals were visible during data collection. Observations were uniformly distributed over the hours of 1100 to 2030.

D.2 Peace is different from conflict

Pigtailed macaque conflict is characterized by long peaceful periods, with mean duration ~400 s, that are interrupted by brief periods of conflict, with mean duration ~40 s. We compare the distributions of peace and fight durations in Figure D1. Both span multiple orders of magnitude, but are distinct in shape. We emphasize this difference in Figure D1 by rescaling each distribution by its standard deviation. Fight durations have a fatter tail quantified by a skewness of 5.1 compared to 3.3 for peace (KS statistic of 0.40, $p < 10^{-9}$).

Similar scale free forms for growth processes have been found in distributions of votes in elections [27], size distributions of unicellular eukaryotes [29], bacterial cell sizes in varying conditions [77], and more generally under the name of fluctuation scaling in a variety of systems [24]. Additionally, growth processes produce distributions well-approximated by lognormals when they are the result of many independent multiplicative factors (by the analogue of the Central Limit Theorem in log-space). Below, we argue for a different mechanism in our system: our scaling collapse results from pairwise fight durations having a

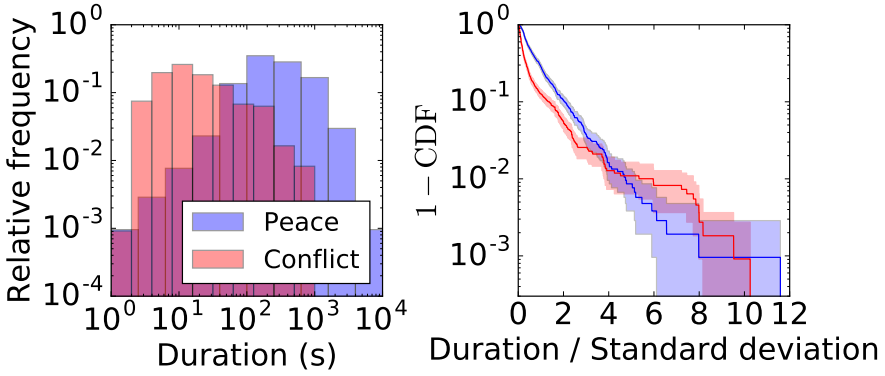


Figure D1: Comparison of peace and conflict durations. (left) Distributions of peace and conflict durations. (right) Complementary cumulative distribution function of peace and conflict durations rescaled by standard deviation for comparison. The distributions are different as captured by the KS statistic of 0.40 ($p < 10^{-9}$). The difference in the decay of the tails is reflected in the skewness—defined as the normalized third moment $\left\langle((t-\mu)/\sigma)^3\right\rangle$—of 3.3 and 5.1 for peace and fights, respectively.

nearly lognormal distribution and larger conflicts consisting of correlated sums of pairwise durations.

D.3 Diffusion models

The scaling collapse with a universal lognormal curve suggests that the standard deviation is proportional to the mean: $\sigma \propto \tilde{\mu}$. As we show in Figure D4, however, many of the measured variances fall below the line and weighted least squares returns a sublinear relationship, $\sigma \propto \tilde{\mu}^{0.95}$. If conflict duration is explained as the sum of $\alpha\binom{n}{2}$ random variables representing the duration of pairwise interactions then this scaling implies that the interactions are correlated in duration, but the correlation is not perfect (Table 2.1). If they were uncorrelated, we would observe $\sigma \propto \tilde{\mu}^{1/2}$. Here, we describe the two different diffusion-drift models we use to measure the deviation from perfect correlation.

We propose a random walk model where a single conflict consists of a sequence of pairwise interactions each with some duration t_k at the kth interaction. In the limit that the duration of all interactions in a sequence are perfectly correlated, $t_k = t_j$ for all pairs (j,k), the correlation time is infinite and the probability distribution of the first interaction $p(t_0)$ should be consistent with the observed distribution $p_2(t)$ so that we recover the observed lognormal distribution (Eq 2.5). In the limit where each sequential interaction is completely independent of the past, the distribution at each time step should be consistent with $p_2(t)$ and con-

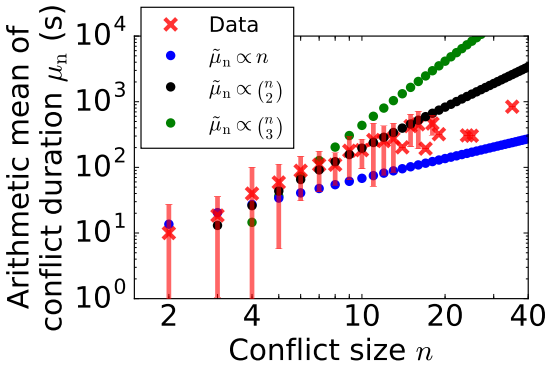


Figure D2: Scaling of the arithmetic means of conflict duration with fight size with error bars spanning two standard deviations. See Figure 2.4 for analogous plot with geometric means.

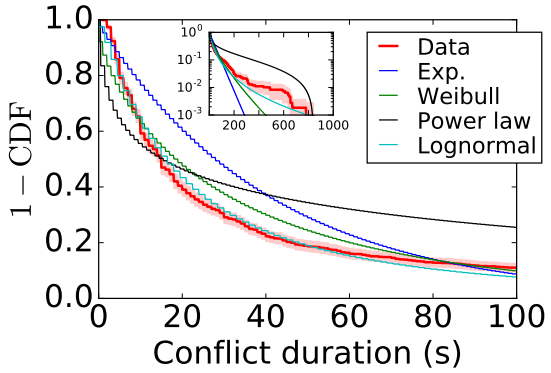


Figure D3: Fits to aggregated distribution of conflict duration. Maximum likelihood fits of several common families of distributions to observed conflict durations. Exponentially decaying distributions like the Weibull cannot fit the heavy tail. Note that on a logarithmic plot, deviations in the tail are highly conspicuous [17, 57].

sequently the equilibrium distribution $p(t_{k\to\infty}) = p_2(t)$, but this leaves open the question of the temporal dynamics that lead to it.

A simple and solved set of dynamics that converges to the observed log-normal distribution corresponds to diffusion in logarithmic space, or the Ornstein-Uhlenbeck process in statistics and more commonly known as the Fokker-Planck equation in statistical physics [68, 81]. Besides having the proper stationary distribution, it has a single parameter D that controls how quickly correlations disappear in sequential interactions. Generally speaking, different dynamics will lead to differently shaped curves for decorrelation, but we are primarily interested in the characteristic decay time for correlations which

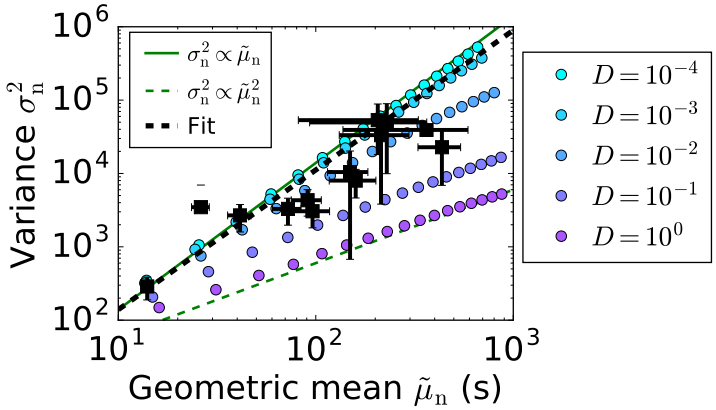


Figure D4: PDM scaling of variance with geometric mean as a function of diffusion constant D. Data follows $\sigma_n^2 \propto \tilde{\mu}_n^{1.9}$ (dashed black), significantly faster scaling than the completely uncorrelated case where $\sigma_n^2 \propto \tilde{\mu}_n$ (dashed green), but not quite the perfectly correlated case $\sigma_n^2 = \tilde{\mu}_n^2 e^{\tilde{\sigma}^2}(e^{\tilde{\sigma}^2} - 1)$ (solid green). Weighted linear fit has slope between with $D = 10^{-3}$ and $D = 10^{-4}$ but data seems to also agree with $\tau \gtrsim 10^2$ s, consistent with the results from the FPM suggesting $\tau > 270$ s (Figure 2.7).

determines the scaling between the variance and the means. Moreover, we show that we find similar results for another model with different dynamics later.

In the Fokker-Planck model (FPM), the initial interaction takes some duration t_0 with distribution $p_2(t_0)$ and the next duration is given by a random multiplicative factor ξ_1 that multiplies the duration of the previous interaction $t_1 = \xi_1 t_0$. This corresponds to a random walk in logarithmic space. The variance of the multiplicative factor is controlled by a diffusion constant D that determines how quickly sequential pairwise interactions decorrelate. The corresponding differential equation for the movement of the random walk (during a *single* conflict) in log-duration space ($\eta = \ln t/\tilde{\mu}_2$) is

$$\partial_k p(\eta, k) = \gamma \partial_\eta [\eta\, p(\eta, k)] + D\partial_\eta^2 p(\eta, k) \tag{D1}$$

where k corresponds to the kth pairwise interaction in the sequence. The ratio D/γ determines how quickly the random walker returns to the equilibrium distribution.

The stationary distribution is given by

$$p(\eta, k \to \infty) = \sqrt{\frac{\gamma}{2\pi D}} e^{-\gamma\eta^2/2D}. \tag{D2}$$

Thus, the parameter γ is fixed by the observed geometric variance $D/\gamma = \tilde{\sigma}^2$, and there is only one parameter that determines how quickly correlations decay.

The correlation of the duration of sequential pairwise interactions is the average over all initial starting conditions and all possible trajectories given an initial starting duration t_0

$$\langle t_0 t_k \rangle - \langle t_0 \rangle^2 = \int_0^\infty dt_0 \, t_0 \, p_\infty(t_0) \int_0^\infty dt_k \, t_k \, p(t_k|t_0) - e^{2\ln\tilde{\mu}_2 + \tilde{\sigma}^2} \tag{D3}$$

$$= e^{2\ln\tilde{\mu}_2 + \tilde{\sigma}^2} \left(e^{\tilde{\sigma}^2 e^{-kD/\tilde{\sigma}^2}} - 1 \right), \tag{D4}$$

where the transition probability $p(t_k|t_0)$ is the Greens function for the Orstein-Uhlenbeck process. Normalizing Eq D4 we find that the correlation in duration of sequential pairwise interactions decays with k as

$$\chi(k) = \left(e^{\tilde{\sigma}^2 e^{-kD/\tilde{\sigma}^2}} - 1 \right) \Big/ \left(e^{\tilde{\sigma}^2} - 1 \right). \tag{D5}$$

We take the unitless characteristic decay time k^* to be the number of interactions it takes for χ to decay to $1/e$:

$$k^* = \frac{\tilde{\sigma}^2}{D} \left[\ln\tilde{\sigma}^2 - \ln\left(\ln\left(\frac{e^{\tilde{\sigma}^2} - 1}{e} + 1 \right) \right) \right]. \tag{D6}$$

To get some intuition for this expression, we expand around small $\tilde{\sigma}^2$, finding

$$\chi(k) \approx e^{-kD/\tilde{\sigma}^2} \tag{D7}$$

in which case

$$k^* \approx \tilde{\sigma}^2/D \tag{D8}$$

As we expect, a larger diffusion coefficient means that correlations decay faster.

Given that the (arithmetic) average pairwise fight has duration $\mu_2 = 10.0\,\mathrm{s}$, we define a decorrelation time in units of seconds as

$$\tau := \mu_2 k^*. \tag{D9}$$

We find that the data is consistent with $D < 0.0175$, or $\tau > 270\,\mathrm{s}$, well beyond the typical duration of a conflict.

The distribution of conflict durations at the kth interaction is the convolution of all distributions up to $k-1$.

In an alternative set of dynamics that we consider is the Probability Density Model (PDM). We imagine that we sample interactions from the data, where we choose we choose data points more similar to previous ones in low diffusion and more randomly in high diffusion. More formally, let x be a unitless "aggression" variable that varies from 0 to 1, and corresponds to the cumulative probability that the duration of a pairwise interaction is less than x: as a function of interaction duration t_k,

$$x_k = x(t_k) = \int_0^{t_k} p_2(t)\,dt. \tag{D10}$$

The diffusion model does a random walk in x-space. At each step, the probability of moving a distance δx in x-space is a Gaussian with variance D, with reflecting boundaries so that the step does not take x outside $[0, 1]$.

This model is expensive to simulate and difficult to analyze analytically. Instead of computing the shape of the full probability distribution like we do with the FPM, we approximate the distributions by simulating a set of trajectories. We compare the samples with the Kolmogorov-Smirnov test in Figure D5 over and plot the cumulative error between the data and approximate CDFs in Figure D6. We find similarly as with the FPM that characteristic decay times $\tau \gtrsim 10^2$ s return similar fits to the data. The results of the tests are summarized in Figure D5, and the corresponding decorrelation times are shown in Figure D7.

D.4 Predicting conflict growth

A crucial question for managing conflict is estimating the size and duration to which an ongoing conflict might grow. Accurate predictions of these quantities could help decide when to allocate limited resources to limit or even promote conflict. Using our model, we can make predictions of how many more individuals will join an ongoing fight and how long we can expect the conflict to last.

In the previous section, we have shown how our model captures the distribution of conflict duration conditional on the conflict size $p(t|n)$, and we combine

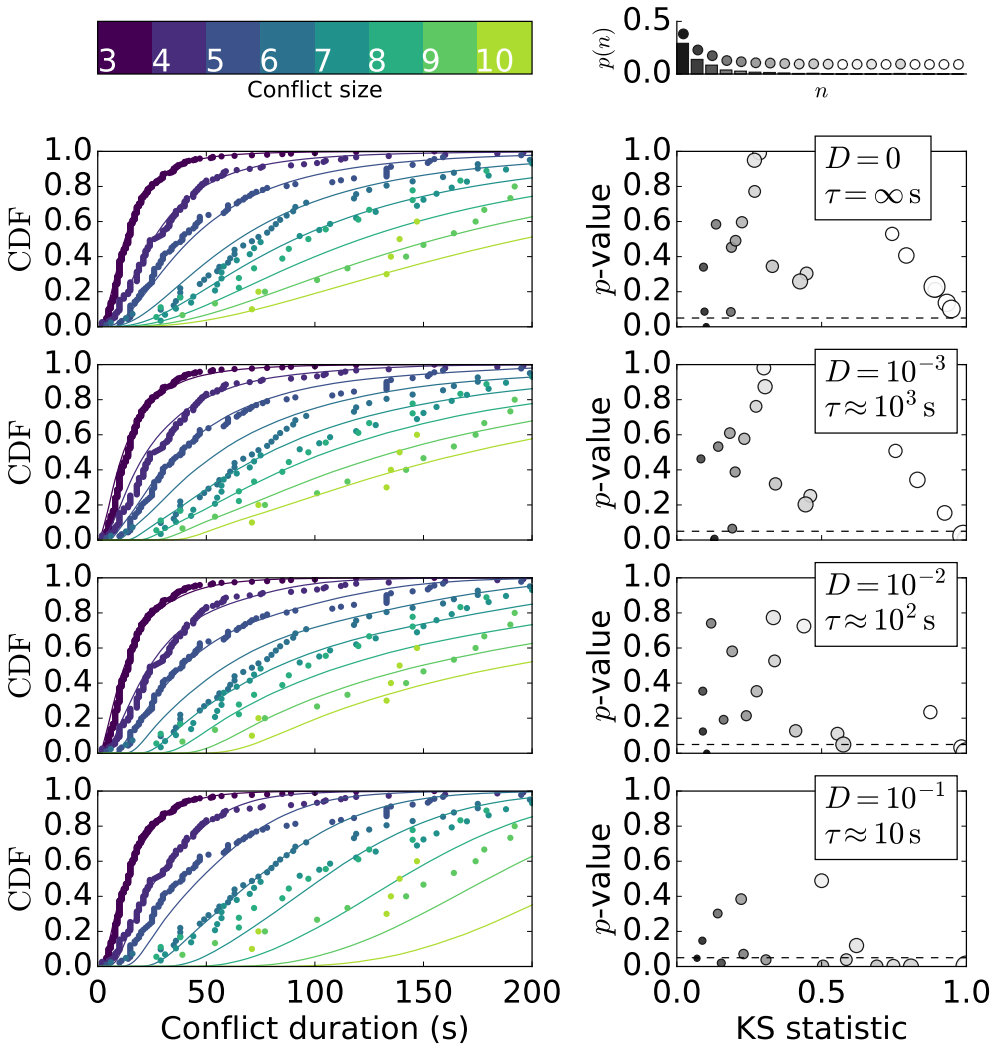


Figure D5: PDM fit to data. (left) PDM prediction of distribution of conflict duration. Conflict sizes go from 3–10 from black to green. Data as circles and model predictions are connected by a line. (right) KS test statistics and p-values, where the size of the circle is proportional to the conflict size and the color represents relative frequency of the conflict size (top). Dashed line demarcates $p \leq 0.05$. Fits are within expected fluctuations for $D = 0$ and continue to be reasonable for $D \leq 10^{-3}$. See Figure D6.

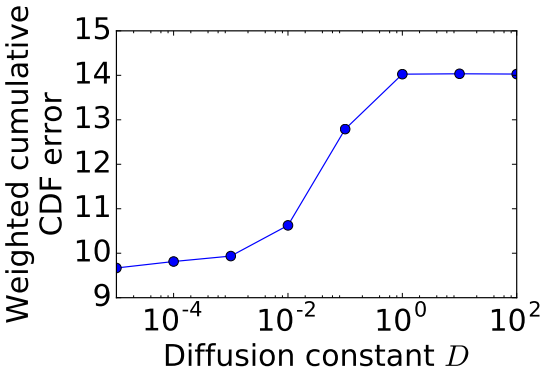


Figure D6: Cumulative CDF error between data and PDM. The cumulative error is the integral of the absolute difference in logarithmic space weighted by the fraction of data for a conflict of size n: $\sum_n p_n \int_{-\infty}^{\infty} |\mathrm{CDF}_{\mathrm{data}}(t) - \mathrm{CDF}_{\mathrm{model}}(t)|\, d \ln t$. Cumulative error starts to plateau between $D = 10^{-2}$ and $D = 10^{-3}$ corresponding to a decorrelation time of $100\,\mathrm{s} < \tau < 1000\,\mathrm{s}$ (Figure D7), in agreement with the more precise measurements from the FPM in Figure 2.7.

this with the observed $p(n)$ to model the full distribution $p(t, n) = p(t|n)p(n)$. The probability that a fight will extend by time Δt with Δn additional members given that we have observed n_0 participants at time t_0 is an application of Bayes' theorem:

$$p(\Delta t, \Delta n | t_0, n_0) = \frac{p(t_0, \Delta t, n_0, \Delta n)}{p(t_0, n_0)} \tag{D11}$$

$$= \frac{p(t = t_0 + \Delta t, n = n_0 + \Delta n)}{p(t_0, n_0)} \tag{D12}$$

$$= \frac{p(t|n)p(n)}{\sum_{\Delta n} \int p(t_0 + \Delta t | n_0 + \Delta n) p(n_0 + \Delta n)\, d\Delta t} \tag{D13}$$

where we have used the fact that the full duration of the conflict $t = t_0 + \Delta t$ and the final size of the conflict $n = n_0 + \Delta n$.

By marginalizing out either Δt or Δn, we can focus on how large a conflict might become or how long it might last:

$$p(\Delta n | n_0, t_0) = \int p(\Delta t, \Delta n | t_0, n_0)\, d\Delta t \tag{D14}$$

$$p(\Delta t | n_0, t_0) = \sum_{\Delta n} p(\Delta t, \Delta n | t_0, n_0) \tag{D15}$$

We show an example of a dyadic $n_0 = 2$ conflict that has been ongoing for $t_0 = 10\,\mathrm{s}$ in Figure D8. In this particular example, the conflict is more likely to grow to size 3 or 4 before resolving, reflecting the fact that most conflicts of

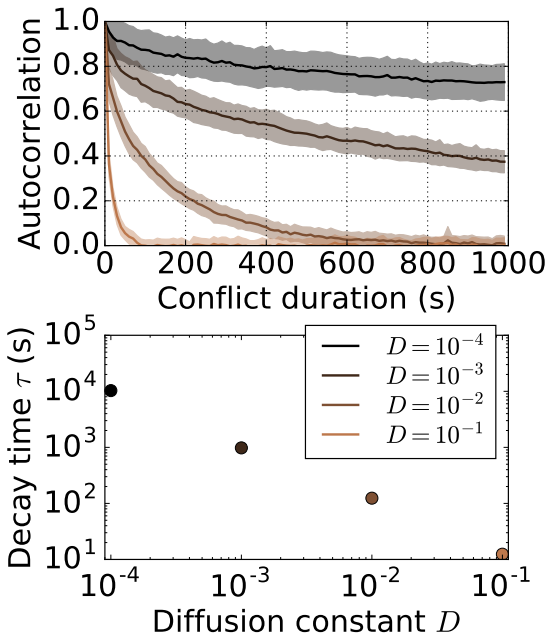


Figure D7: PDM decorrelation time as a function of diffusion constant D using the average pairwise interaction duration $\mu_2 = 10.0$ s as the scale.

size 2 are less than 6.7 s. On average, this fight will grow by 2.5 individuals but with rather large standard deviation of 2.8. The average time extended is (63 s) with standard deviation (177 s), the large standard deviation reflecting skewed statistics from a heavy-tailed distribution.

We can compare potential consequences of stopping the conflict at particular times. In Figures D9 and D10, we consider interventions that stop an ongoing fight between two individuals that has progressed for 2 s, 10 s, and 20 s.

One possible goal is to minimize total conflict size. Here, an early intervention at $t_0 = 2$ s does not prevent as many individuals from joining as an intervention at $t_0 = 10$ s when the probability of another individual joining grows to a factor of 2 above the probability of no other individuals joining. At $t_0 = 20$ s it is 3x more likely that another individual will eventually join.

Another objective could be to control the total duration at the time of observation t_0, but this scenario again depends on the particular risk we wish to minimize. If maximizing the probability that the fight ends immediately, we note fights are relatively most likely to end immediately sometime between $t_0 = 2$ s and $t_0 = 20$ s when it might make more sense to wait for the conflict to end rather

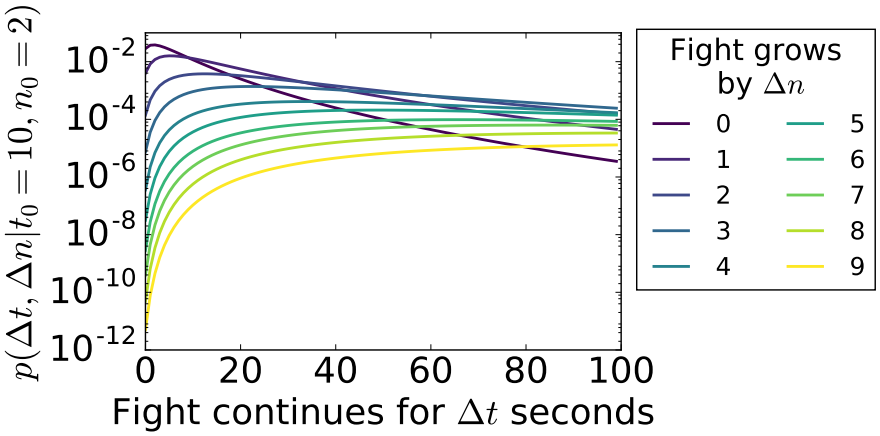


Figure D8: Probabilities $p(\Delta t, \Delta n | t_0 = 10, n_0 = 2)$ of an ongoing fight with 2 individuals after 10 s continuing for Δt seconds and Δn more participants. We use the lognormal model of fight duration instead of the empirical distributions and use the empirical distribution of fight sizes up to 18 (although we only show $\Delta n < 10$).

than intervening. On the other hand, given a single opportunity to intervene, the most consequential intervention on average is at 20 s because the average duration extended is 47 ± 145 s, 63 ± 177 s, and 87 ± 215 s, respectively; thus, a late intervention causes the most change in the expected duration of conflict. Averages, however, are skewed for heavy-tailed distributions. Indeed, this dominance of the tail beyond 20 s suggests that a risk-averse objective—minimizing the possibility of long conflicts—would involve stopping conflict early.

If only a limited number of interventions are available for a large number of conflicts, mitigating total time spent in conflicts or the number of participants would involve choosing intervention times order for maximum effect. In this particular scenario, efficient use of intervention opportunities may be an identifiable strategy applied by conflict managers in social systems.

D.5 Human conflict

We look at how the geometric mean of human armed conflict durations scales with the number of participating entities using two openly available datasets: the Correlates of War (COW) and Uppsala Conflict Data Program (UCDP) databases [30, 43, 70].

The COW database only includes wars that "must involve sustained combat, involving organized armed forces, resulting in a minimum of 1,000 battle-related

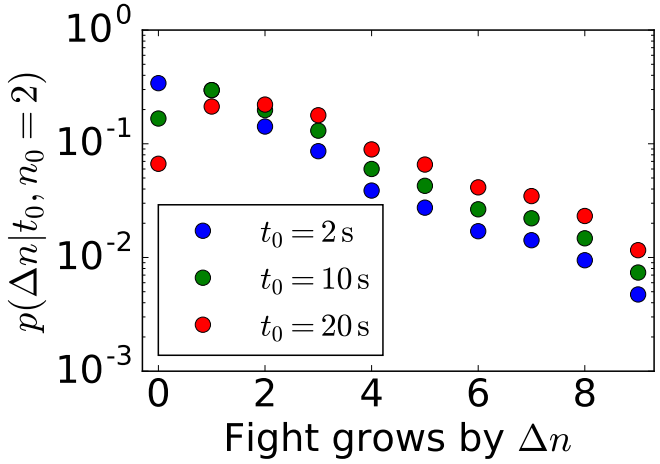


Figure D9: Probabilities $p(\Delta n|t_0, n_0 = 2)$ that fights with 2 individuals still ongoing at 2 s, 10 s, and 20 s grow by Δn more participants as from Eq D14.

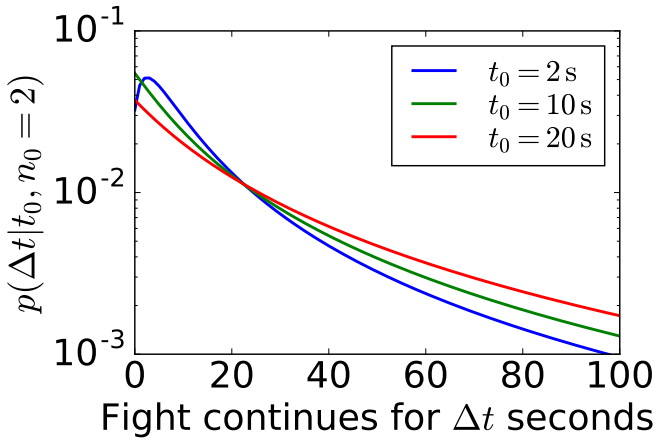


Figure D10: Probabilities $p(\Delta t|t_0, n_0 = 2)$ that fights with 2 individuals still ongoing at 2 s, 10 s, and 20 s last for Δt more seconds as from Eq D15.

combatant fatalities within a twelve month period" as well as be between entities that can provide "effective resistance" [70]. As we show in Figure D11, the geometric means for conflicts with $n \geq 3$ follow a binomial scaling better than a simple linear scaling as with the macaque conflict data. We compare the linear and binomial models by assuming a lognormal distribution for each conflict size with a mean given by the model with some unknown universal variance $\tilde{\sigma}^2$. When we take the log-likelihood ratio R of the binomial model to the linear model, the variance cancels out, and $R = 9.0$ for $N = 29$ conflicts of size $n \geq 3$.

One clear anomaly that we do not consider for our fit is the distribution of wars

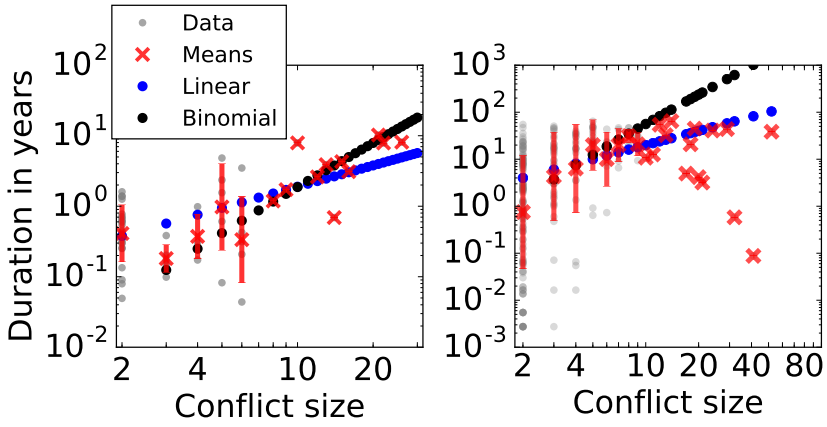


Figure D11: Scaling of geometric mean durations of interstate wars. (left) COW database and (right) UCDP. The COW means scale as $\binom{n}{2}$ like with the macaque conflict data for $n \geq 3$, and the scaling of the UCDP means is compatible with $\binom{n}{2}$ growth for small conflicts.

between only two states. Here, the distribution is much wider compared to larger wars, and the geometric mean is higher than for conflicts with 3, 4, or 6 states. It is not clear why conflicts for $n = 2$ do not follow the trend established by larger conflicts. One possibility is that pairwise interstate conflicts are different. Other possibilities may have to do with the classification of pairwise conflicts or how the duration is determined.

The UCDP database counts conflict episodes "defined as years of continuous use of armed force in a conflict" [43] where an armed conflict is "a contested incompatibility that concerns government and/or territory where the use of armed force between two parties, of which at least one is the government of a state, results in at least 25 battle-related deaths in a calender-year" [30]. Here, it is clear that there is a discrepancy in the growth of the mean after $n \gtrsim 10$. Fitting to conflicts $3 \leq n \leq 10$, we find that the log-likelihood ratio $R = 1.0$. This means that for small conflicts there is insufficient evidence to show that one scaling is a better explanation than the other.

That the scaling of means with size follows a similar pattern at all is highly intriguing because the UCDP dataset includes a variety of conflicts between states and between states and non-states within shared territory or even on outside territories. Furthermore, the definition of conflict and duration differ between the COW and UCDP databases. Evidently, other details are important for comparing analogous human conflicts with the macaque system on which we focus, but it seems that similar scaling ideas may provide a place to start for compari-

son across these systems.

E Appendix for Chapter 2.2

E.1 Armed Conflict Location & Event Data (ACLED) Project

We use the data set provided online as ACLED Version 7 [65]. This project measures political violence around the world with a focus on African states for 20 years (Jan. 1, 1997 through Dec. 31, 2016). The data set is organized around events, which have a specific date and time. We analyze three types of events included in the data set: Battles between armed groups ($K = 42,738$), Violence Against Civilians ($K = 39,127$), and Riots/Protests ($K = 37,582$).

According to the codebook, there are three different kinds of battles that we include in our Battles analysis. As quoted from the codebook, these are defined as

1. Battles - No change of territory: "A battle between two violent armed groups where control of the contested location does not change. This is the correct event type if the government controls an area, fights with rebels and wins; if rebels control a location and maintain control after fighting with government forces; or if two militia groups are fighting. Battles take place between a range of actors."
2. Battle - Non-state actor overtakes territory: "A battle between two violent armed groups where non-state actors win control of a location. If, after fighting with another force, a non-state group acquires control, or if two non-state groups fight and the group that did not begin with control acquires it, this is the correct event. There are few cases where opposition groups other than rebels acquire territory."
3. Battle - Government regains territory: "A battle between two violent armed groups where the government (or its affiliates) regains control of a location. This event type is used solely for government re-acquisition of control. A small number of events of this type include militias operating on behalf of the government to regain territory outside of areas of a government's direct control (for example, proxy militias in Somalia which hold territory independently but are allied with the Federal Government)."

We also investigate Violence Against Civilians (VAC):

> Violence against civilians is a violent act upon civilians by an

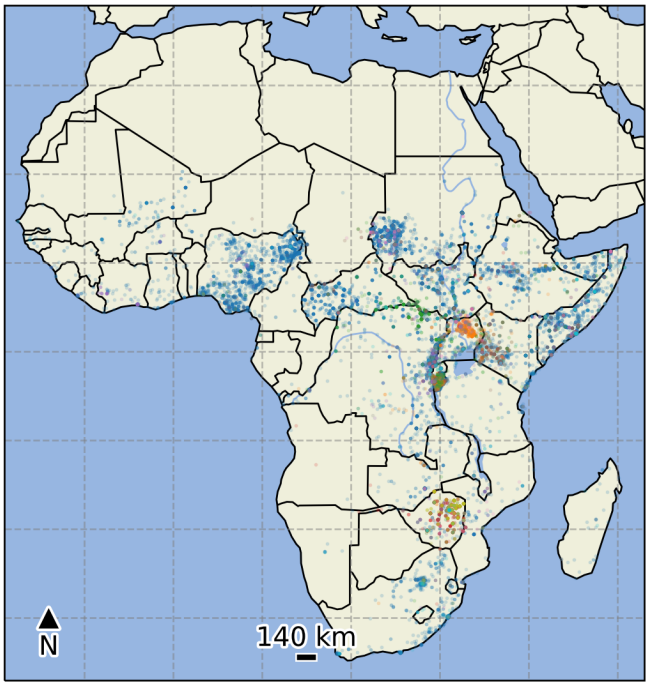


Figure E1: Spatial distribution of 10 largest conflicts involving Violence Against Civilians (VAC) given $b = 140$ km and $a = 128$ days. Map made with Natural Earth.

> armed, organized, and violent group. By definition, civilians are unarmed and not engaged in political violence. Rebels, governments, militias, external forces, and rioters can all commit violence against civilians. Protesters are also civilians, and significant violence against protesters falls under this category.

Finally, there are Riots/Protests:

> A protest is a public demonstration in which the participants do not engage in violence, though violence may be used against them. Often—though not always—protests are against a government institution. Rioting is a violent form of demonstration where the participants engage in violent acts, including but not limited to rock throwing, property destruction, etc. Both of these can be coded as one-sided events. All rioters and protesters are noted by generic terms (e.g. 'Rioters (Country)' or 'Protesters (Country)'); if representing a group, the name of that group is recorded in the respective 'associated actor' column.

Figure E2: Spatial distribution of 10 largest conflicts involving Riots/Protests given $b = 140$ km and $a = 128$ days.

In the analysis, we only consider statistics of the conflict avalanches where $T > 1$, $S > 1$, $F > 1$, and $L > 0$ although an event does not have to satisfy cutoffs simultaneously, i.e., we may use it for $P(S)$ but not $P(F)$. We find that the statistics of events below these lower cutoffs generally deviate from the observed power law statistics in the rest of the distribution, and such deviations are likely attributable to data problems. For some events, ACLED sets the estimate of fatalities $F = 0$ unless they have confirmed with a "reputable source," so some of these cases are simply missing statistics (there is no way to distinguish between missing data or no fatalities). As for time scales, the highest precision available in the data set is to the day which defines a lattice scale below which we cannot probe. As for length scales, we find many events occur exactly at the same geographic coordinates which presumably also involve some lattice scale below which the data aggregators either could not access or did not find a pressing need to do so.[10] Such resolution effects are akin to "lattice" artifacts common in human reported data. Such artifacts occur in other data sets like the Iraq War Logs where soldiers round the time at which events happened to 10 or 30 minute intervals. Importantly, these anomalies matter little at large length scales where such effects are dominated by the large scale regularities of the system.

[10] Accurate data on conflict is difficult and even dangerous to collect and necessarily this data set does not sample all events with equal accuracy or detail. Nevertheless, a conflict data project of this scale is unprecedented.

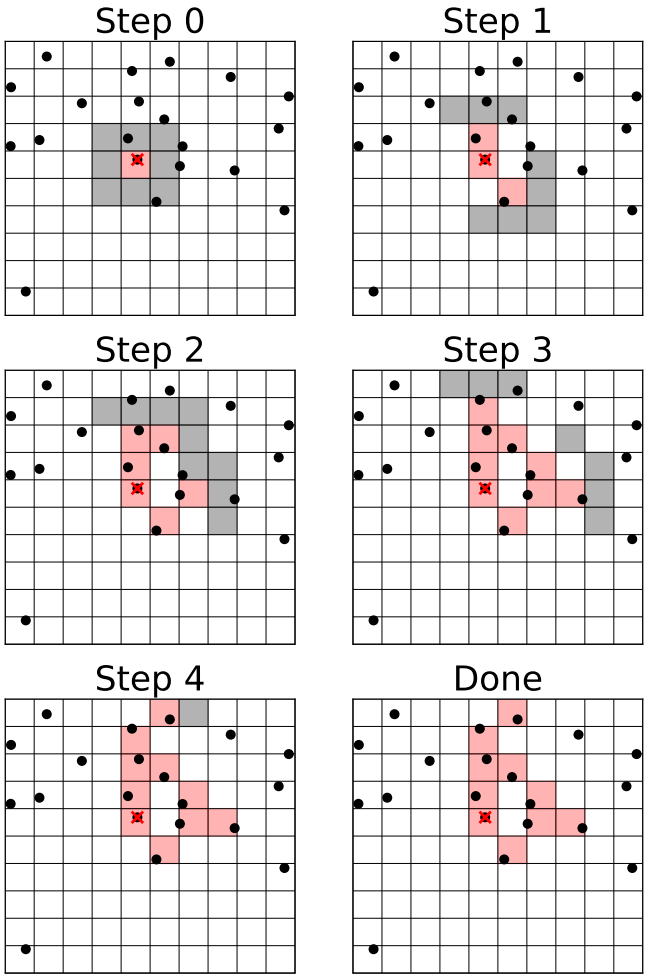


Figure E3: Schematic of clustering algorithm for building a conflict avalanche. At Step 0, the algorithm picks a random event and begins building a cluster there. Then, all new neighbors (gray) of new tiles added to the cluster (red) are evaluated for events (black circles). When no more tiles can be added, the algorithm stops.

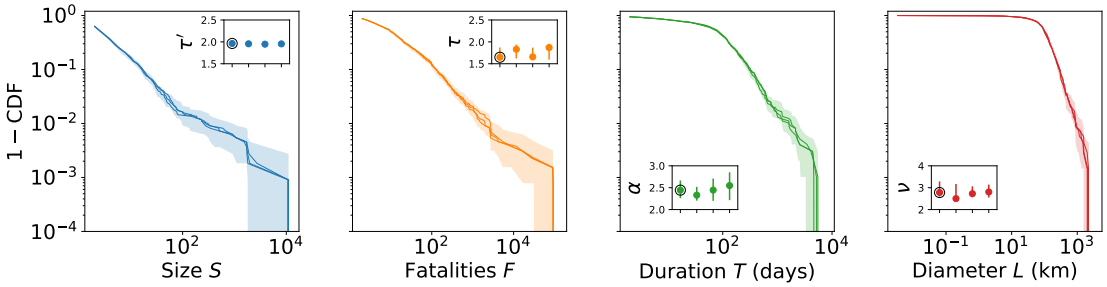


Figure E4: Distributions of scaling variables for several random Voronoi tilings. Bootstrapped confidence intervals of 90% for the data in Figure 2.11B are shown behind the distributions for three other Voronoi tilings (points connected by lines for visibility). These all fall well within expected statistical variation. (inset) Measured exponents for the four distributions where the circled exponent corresponds to data used in the main text. The choice of tiling does not substantially alter the exponent. The fluctuations in the means visible, for example for τ', reflect variation in the lower bound found for the data, variability that is inherent in the fitting procedure when a single lower bound is chosen [17]. Importantly, this fluctuation is captured by the bootstrapped confidence intervals.

E.2 Clustering algorithm

To generate our conflict avalanches, we choose a separation length scale b and separation time scale a that correspond to the minimum separation between sequential pairs of events in a single avalanche. To do this, we first bin the time points into bins of width a and consider any contiguous sequence of bins with at least one event to be potentially (we must account for geographic distance too) part of the same conflict avalanche. In contrast to how avalanches are constructed for neural systems [78], we do not discretize the day on which avalanches occurred to the scale a after constructing the avalanche, but preserve the precise time at which events were reported (except for rate profiles which are shown here in the SI). Such discretization to a lattice scale is unnecessary for exploring scaling relations. As a result, the temporal clustering procedure constructs sequences of contiguous events where breaks are inserted between any pair of events with at least separation of a days.

An exact analog of this unidimensional procedure to the surface of the Earth is impossible because no regular tiling of the surface of a sphere exists. Surely, one approach without bins would be to measure directly the pairwise distance between every pair of events, but this approach scales as $O(N^2)$ and is slow because geodesic distance calculations are expensive. With our data set of 10^4–10^5 events, such a procedure would take inordinately long on a desktop computer.

Instead, we generate a Voronoi tiling of the Earth using a Poisson disc sampling algorithm to generate a regularly-spaced set of tiles with average spacing of $b/2$ [22]. Neighboring "bins" correspond to Voronoi tiles whose centers are within a fixed distance b, and we can search for contiguous sets of tiles that have at least one event.[11] Importantly, this Voronoi algorithm only involves distance calculations that scale as the square of the number of *tiles* regardless of the density of events.

As a simple demonstration of the algorithm, we provide a schematic in Figure E3 that iterates through the construction of a single conflict avalanche in a 2-dimensional space (or one dimension of space and one of time). In this particular example, each tile has exactly 8 neighbors, whereas the actual number of neighbors will vary randomly in the Voronoi tiling. At each step, all *new* closest neighbors (gray) of the cluster (red) are evaluated and appended onto the existing cluster if they contain an event (black point). Once the cluster can no longer grow because there are no neighboring tiles with events, the algorithm stops. This procedure defines a systematic way of constructing sequences of related events given spatial and temporal scales.

Although different random Voronoi tilings will cluster events in a slightly different way, we find that the variation from such randomness is small compared to the statistical variation estimated from bootstrapped confidence intervals for a single Voronoi tiling. As we show in Figure E4, the distributions of conflict statistics across several random Voronoi tilings are all very similar. The measured exponents likewise agree within the bootstrapped confidence intervals. Thus, the Voronoi clustering procedure serves as a computationally efficient way of generalizing the temporal discretization procedure used to identify contiguous events in one dimension to curved surfaces.

As we mention in the main text, we focus on $b = 140\,\text{km}$ because it presents a "Goldilocks" zone where avalanches occur over a wide range of sizes. In Figure E5, we show the spatial distribution of the largest 10 Battle conflict avalanches as we vary a with $b = 140\,\text{km}$ fixed as in the main text. In Figure E6, we present an overview of avalanche statistics across a much broader range of spatiotemporal scales (b, a) ranging from $35 \leq b \leq 2{,}200\,\text{km}$ and $1 \leq a \leq 512\,\text{days}$. When the temporal scales are short, avalanches do not percolate far and we are limited to very small, short, and spatially localized conflicts (pink box repre-

[11] More generally, a tiling with spacing b/k has resolution (and computational cost) that increases with k, reducing variability with different random Voronoi grids such that when $k \to \infty$, there is a unique clustering. For our data, we find that $k = 2$ is sufficient to return similar statistics between different Voronoi grids.

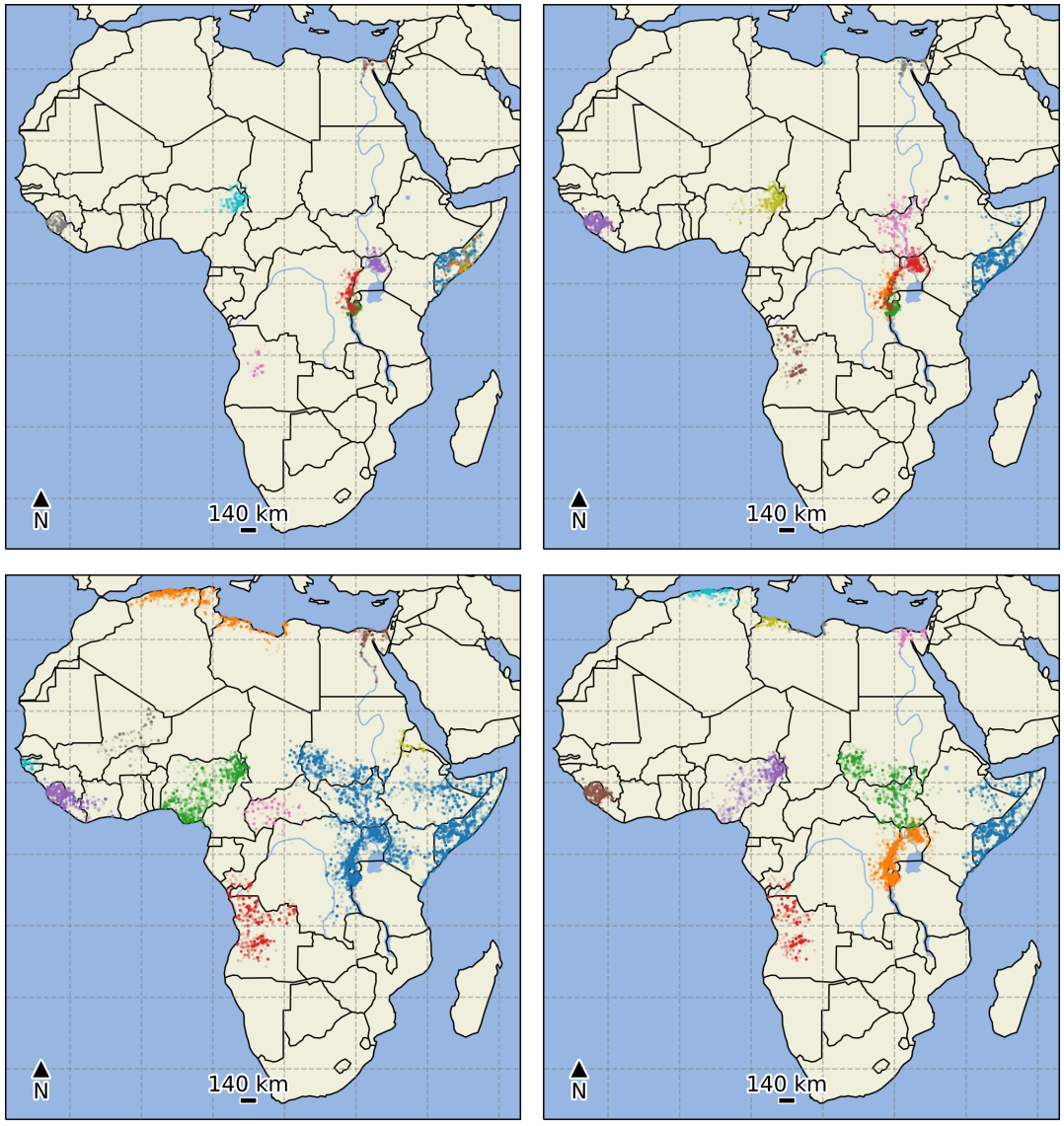


Figure E5: Spatial distribution of 10 largest Battle conflict avalanches for $b = 140$ km and multiple separation scales a. (clockwise from top left) $a = 16$ days, $a = 32$ days, $a = 64$ days, $a = 256$ days.

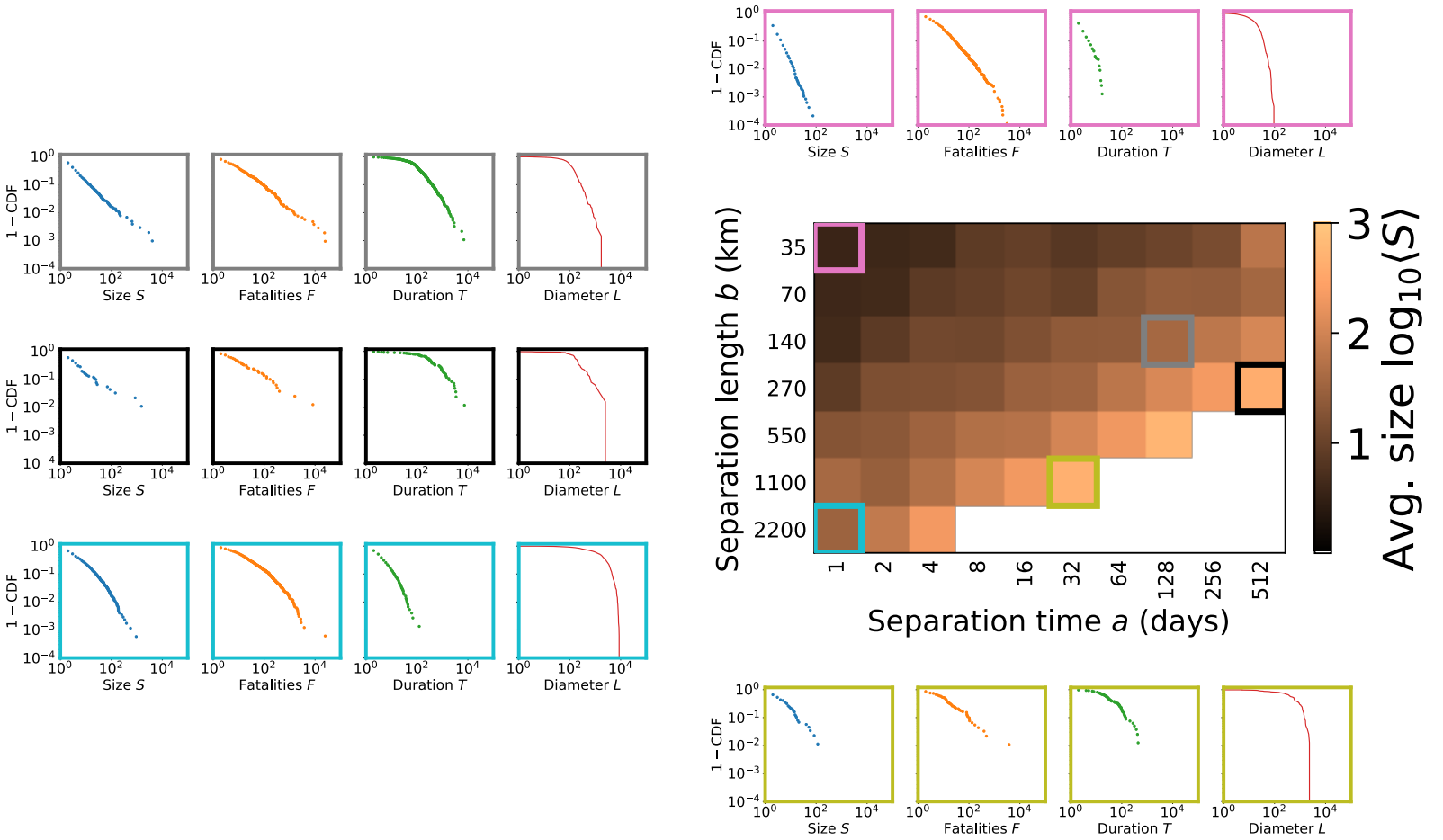


Figure E6: Battle distributions of scaling variables S, F, T, and L across a range of spatiotemporal scales $35 \leq b \leq 2,200$ km and $1 \leq a \leq 512$ days. (center) We show the average avalanche size $\langle S \rangle$ given a clustering spatiotemporal scale (b, a) to give a sense of the variation across all scales. Where we have K < 50 data points above the lower cutoff, the region is whited out. (top, pink) When b is small, avalanches are likewise small ($S < 10^2$ including the largest observed avalanche) and show little dynamic range ($T < 10^2$). (top left, gray) In a middle range of b, conflict avalanches exist for a wide range of scales, corresponding to the data that we analyze in the main text. (middle left, black) When the time scale a is comparable to the total duration of the data set (~ 8,000 days), a few avalanches approach the limits of the data, we have many fewer conflicts to examine, and we lose dynamic range in the statistics. (bottom left & right, teal & gold) When the length scale is comparable to the entire extent of the African continent (~ 8,000 km), most conflict avalanches span the system as visible from the diameters L, and the data set becomes sparse.

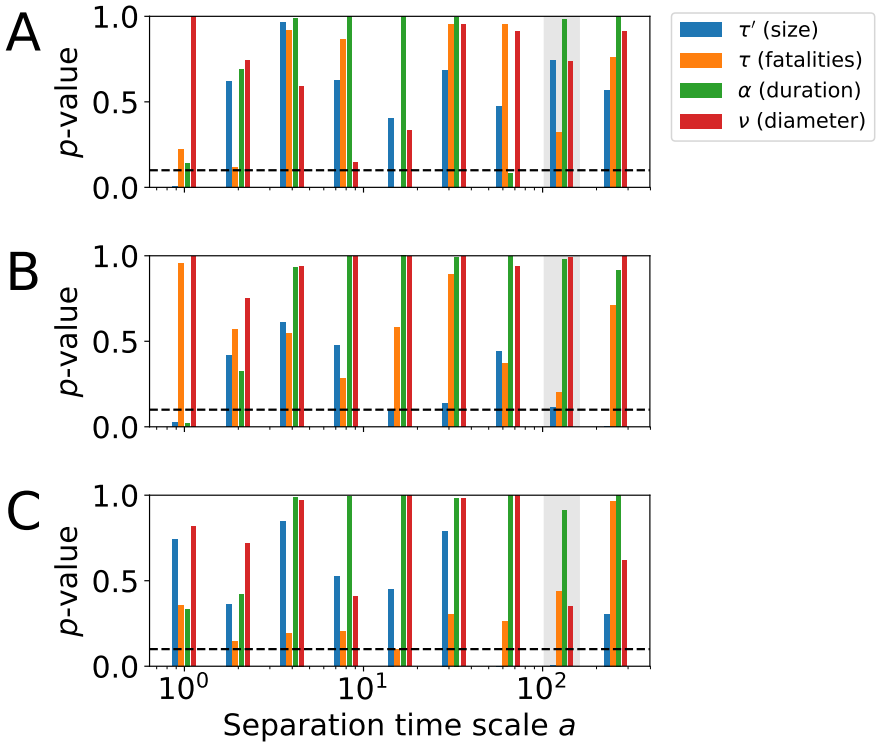


Figure E7: Results of statistical tests for power law fits to (A) Battles, (B) VAC, and (C) Riots/Protests. We consider distributions with $p \geq 0.1$ to be statistically indistinguishable from power laws [17]. The separation time we use in the main text $a = 128$ days is highlighted in gray.

senting $b = 35$ km and $a = 1$ day). Although most of the variables here show limited dynamic range, the distribution of fatalities is spread out across three orders of magnitude. This property reflects the prominence of conflict fatalities in the armed conflict literature given that fatalities are heavy-tailed even without accounting for spatiatemporal scales. When we go to much larger scales of $b = 1{,}080$–$2{,}060$ km and $a = 512$ days (black, teal and gold boxes), a few avalanches start to span the physical size of the African continent ($\sim 8{,}000$ km) and the time series ($\sim 8{,}000$ days). We would expect boundary effects to dominate in this regime and correspondingly avalanche space and time scales are compressed to a small region along either cutoff. As a result, we have many fewer conflict avalanches on which to estimate scaling parameters, so we avoid this regime. For a middle range of b around 10^2 km, we can probe a wide range of temporal scales for avalanches that display scale-invariant statistics in the tails while also accumulating a reasonable number of temporal profiles to evaluate.

Indeed, the choice of appropriate scale on which to define related events is a

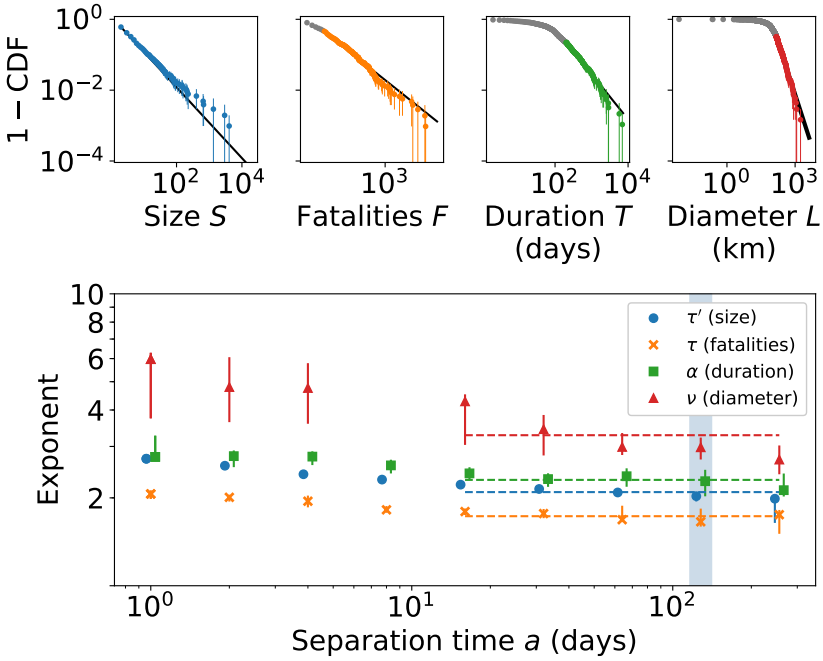


Figure E8: Measured exponents for VAC. All shown distributions for $a = 128$ days are indistinguishable from power law distributions at the $p \geq 0.1$ level according to the KS test. There is a missing point for ν at $a = 16$ days because the measured value exceeds the upper 90% confidence bound. Such an artifact can occur when the tail of the distribution is not sampled well as can happen with a large lower cutoff. In these cases, the measured exponent may be unreliable.

problem that has received much attention in the context of neural avalanches. For neural avalanches, researchers must determine appropriate interspike intervals and often must account for a fixed electrode spacing [6]—although new high-resolution, nearly single-cell optical techniques have become possible [61]. In principle, the physical layout of axonal and dendritic connections determines a causal network for neural spikes and so direct measurement of true (not only statistical) sequences should be possible. In practice, such measurements are not yet feasible and spatiotemporal proximity is often used as a proxy where a good rule-of-thumb is the average interspike interval as a measure of characteristic time scale. When electrode arrays that effectively define a coarse grid are used, the time scale defining related events must be scaled with the distance between the electrodes because the finite propagation velocity of neural signals sets a relevant scale [6]. Furthermore, other statistical techniques for detecting causality have been explored for constructing "causal networks" that induce very different distributions [84]. Such techniques for determining networks of related events present an opportunity for further work in armed conflict avalanches.

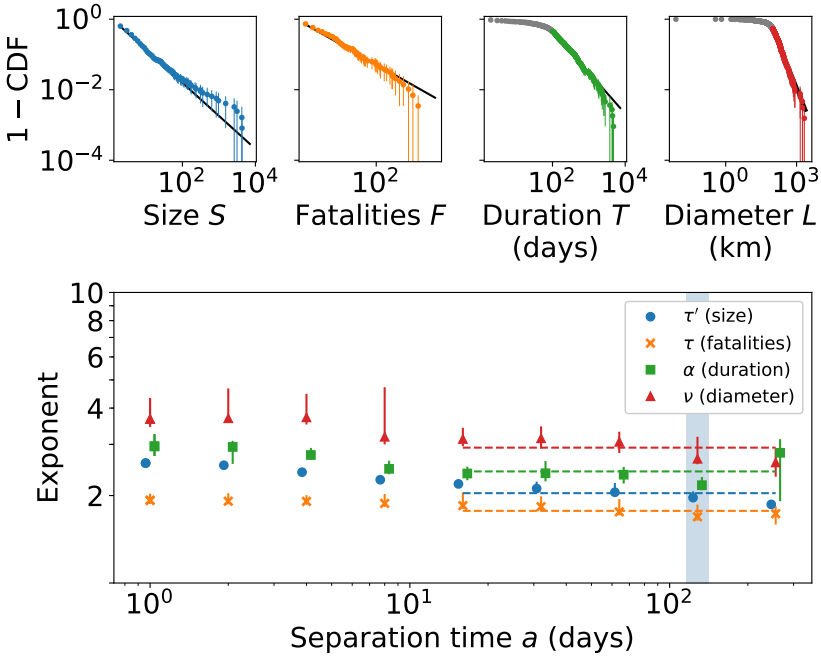


Figure E9: Measured exponents for Riots/Protests. All shown distributions for $a = 128$ days are indistinguishable from power laws at the $p \geq 0.1$ significance level according to the KS test except for $P(S)$. Yet, the separation times nearby $a = 128$ days, namely $a = 32$ days and $a = 256$ days, are in statistical agreement with the power law model which serve as bounds on the possible bias of the exponent estimate. Given that the bound is tight and continuing with the best estimate of the exponent, the scaling relations specified in Eq 8 are satisfied.

For our work, sociopolitical information could be used to cluster events into familiar notions of battles or wars, but such clustering is not deterministic and includes ambiguity both in identification of actors and attribution of responsibility [65]. We take the simplest (and neutral) approach where correlations can be imputed to *physical* spatiotemporal proximity, leading to the surprising conclusion that the spread of armed conflict might be described directly in the language of critical phenomena.

E.3 Power law fitting

Given the conflict avalanches for a given length scale b and time scale a, we extract the scaling variables S, F, T, and L to measure the distribution exponents τ', τ, α, and ν. To fit the exponents, we use the standard procedure described in reference [17]. First, we numerically find the maximum likelihood fits for a given distribution across a logarithmically spaced range of lower cutoffs. For each lower cutoff, we calculate the Kolmogorov-Smirnov (KS) statistic (the maximum distance between the cumulative distribution functions) and choose the lower

	Size τ'	Fatalities τ	Diameter ν	Duration α	S vs. T d_S/z	F vs. T d_F/z	L vs. T $1/z$
Battles	1.96 (1.91, 2.02)	1.65 (1.61, 1.87)	2.78 (2.60, 3.29)	2.44 (2.26, 2.67)	2.0 (1.7, 2.5)	2.5 (2.1, 3.2)	0.78 (0.64, 0.96)
Violence Against Civs.	2.03 (1.94, 2.08)	1.66 (1.59, 1.84)	2.98 (2.71, 3.22)	2.28 (2.02, 2.50)	1.9 (1.5, 2.2)	2.3 (1.9, 2.7)	0.69 (0.56, 0.85)
Riots/Protests	1.97* (1.92, 2.07)	1.69 (1.64, 1.86)	2.68 (2.56, 3.19)	2.17 (2.09, 2.32)	1.6 (1.4, 1.8)	1.5 (1.3, 1.7)	0.66 (0.57, 0.77)
Percolation growth 2D		2.05	2.87	2.65		1.57	0.88
Forest fires 2D		2.14 (2.11, 2.17)	1.28 (1.19, 1.37)	1.27 (1.20, 1.34)		1.89 (1.86, 1.92)	0.96 (0.94, 0.98)
Barkhausen 2D		2	2.09 (1.91, 2.27)	1.87 (1.81, 1.93)		1.55 (1.51, 1.59)	0.80 (0.69, 0.91)
Manna sandpile 2D		1.28 (1.26, 1.30)		1.47 (1.37, 1.57)		1.79 (1.72, 1.85)	
ARW 2D		1.3		1.5			
Neural		2.10 (2.09, 2.11)		2.86 (2.85, 2.87)		1.85 (1.82, 1.89)	
Neural [28]		1.7 (1.5, 1.9)		1.6 (1.4, 1.8)		1.3 (1.25, 1.35)	
Wars		1.53 (1.46, 1.60)					
Terrorism		2.38 (2.32, 2.44)					
Confrontation 2D		1.9*					

Table 4.1: Complete table of exponents showing more examples and uncertainty intervals compared to Table 2.2. For percolation growth, the exponents ν and α are calculated from the other exponents using scaling relations used in the main text. The power law model for sizes for Riots/Protests is not significant within our $p \geq 0.1$ threshold, as is indicated by an asterisk. Here, we also include another example of neural avalanches from a cortical culture [28], terrorism [18], and a coalescence-fragmentation model applied to confrontation [39]. For the latter model, Neil F. Johnson notes that this exponent can be found for confrontation on a two-dimensional space although the power law is not statistically significant [38]. For conflict avalanches, the uncertainty range corresponds to 90% bootstrapped confidence intervals. For the other examples, we take the error bars directly from the cited work.

cutoff with the smallest statistic. This procedure defines how to determine the exponents and lower bounds from the distributions shown in Figure 2.11.

To calculate significance, we sample from the power law fit. If there is a lower bound, we bootstrap sample from the data points below the lower cutoff to construct a full realization of a sample that is a combination of an unparameterized model below the cutoff and a power law above. We then run the same fitting procedure 2,500 times (again fitting the lower bound to each sample) to measure the distribution of the KS statistic. Thus, the KS statistic determines the p-value that we use for significance such that $p \geq 0.1$ indicates that the observed distribution has a KS statistic smaller than 90% of all bootstrapped samples. Across much of the data, the distributions that we find satisfy this stringent criterion for significance demonstrating that the power law form is a convincing model for armed conflict statistics.

For the data that we consider in the main text where $b = 140\,\text{km}$ and $a = 128\,\text{days}$, the distributions are statistically indistinguishable from power laws with $p \geq 0.1$. It is not the case, however, that every distribution for which we measure exponents satisfies this stringent criterion (Figure E7). In the cases where the statistical test fails, often the power law model is a reasonable fit to the tail of the distribution. As a result, we can still measure an exponent though it may be a biased estimate. Such biases appear to be small because the estimated exponents across a range of spatiotemporal scales all take similar values (Figure 2.11). Thus, across a large swathe of data, we find statistical evidence that power laws serve as accurate models when accounting for the spatiotemporal spread of conflict beyond individual events as have been investigated in other examples of armed conflict [15, 39, 60].

We measure the exponents for VAC and Riots/Protests and show them in Figs. E8 and E9 for fixed $b = 140\,\text{km}$ and across the same range of a as with Battles. In the top row of Figs. E8 and E9, the distributions for $a = 128\,\text{days}$ are all statistically indistinguishable from power laws except for $P(S)$ for Riots/Protests. Inspecting this distribution in more detail, we find a hump near the maximum sizes that deviates from the power law form.[12] Thus, the evidence of strict adherence to a power law form is less clear for this

[12]Similar deviations from the power law seem visible both for Battles and VAC size distributions—though they are possibly statistical artifact, such coincidence is noteworthy. Intriguingly, such humps are characteristic of finite-size effects in physical systems near the critical point where the largest avalanches "pile up" near the system size [13, 71]. Although we do not do so here, it is tantalizing to consider what signals of (universal) finite-size corrections may exist for armed conflict data.

particular distribution as we indicate in Table 2.2. Nevertheless, we point out in Figure E9 that the exponents for the adjacent separation times $a = 32$ days and $a = 256$ days tightly bound the range of possible values for $a = 128$ days, which falls in between.

E.4 Dynamical exponents

Next, we measure the dynamical exponents by regression on the appropriate pair of scaling variables. A simple parameterization of the scaling relation is

$$X = aT^{\delta} \tag{E1}$$

with coefficient parameter a and exponent parameter δ. If errors are multiplicative, the fitting procedure is equivalent to least-squares regression in logarithmic space. However, the typical regression problem only accounts for noise along the dependent variable (here X) which returns a solution that is not guaranteed to be symmetric about a fit to the inverse scaling relation

$$T = (X/a)^{1/\delta}. \tag{E2}$$

This asymmetry presents ambiguity in the choice of which regression to use to measure the scaling exponents.

Instead, we define a fitting procedure that ensures symmetry about the inversion of the scaling relation. We minimize a symmetrized cost function that treats both X and T as dependent variables in turn

$$\begin{aligned} C(a, \delta, \sigma_{\mathrm{X}}, \sigma_T) = \sum_{\mathrm{i}=1}^{\mathrm{K}} \left[\log X_{\mathrm{i}} - \delta \log T_{\mathrm{i}} - \log(a)\right]^2 / \sigma_{\mathrm{X}}^2 + \\ \left[(\log X_{\mathrm{i}} - \log(a))/\delta - \log T_{\mathrm{i}}\right]^2 / \sigma_T^2. \end{aligned} \tag{E3}$$

The variance parameters σ_{X} and σ_T account for the possibility that magnitude of the noise along the X dimension may be different than that of the noise along the T dimension. By numerical simulation, we find that the regression procedure using the symmetrized cost function shows similar or less bias than the simple least-squares fit with noisy data, and thus we adopt Eq E3 for fitting the dynamical exponents.

In Figure 2.12, we show the results of regression using Eq E3 to measure the dynamical scaling exponents for sizes and fatalities of conflict avalanches. As we write in the main text, we measure $d_S/z = 2.0$, $d_F/z = 2.5$, and $1/z = 0.8$ with the corresponding 90% confidence intervals in Table 2.2.

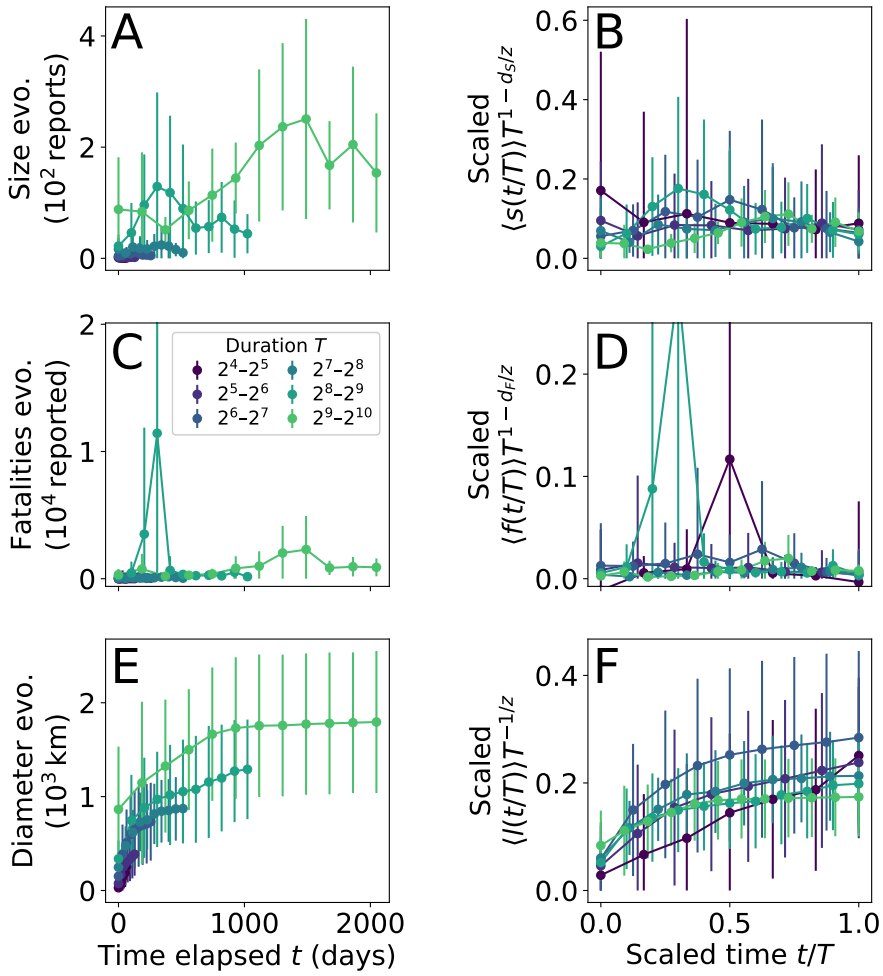


Figure E10: Temporal profiles for Battles with separation scale $a = 64$ days before collapse (A, C, E) and after (B, D, F). Scaling exponents d_S/z and d_F/z were fit by first taking the average profile over bins then minimizing the logarithm of the geometric variance. Even this particularly nice collapse is noisy, but the exponents and that we measure from the collapse agree with those from the scaling relations. For all these profiles, we removed the single conflict avalanche of duration $T > 2^{10}$ days, which showed variability of about an order of magnitude above the shown curves. These events in this large conflict avalanche collectively span much of the Horn of Africa and eastern Central Africa including conflicts related to events in Darfar and all the way to eastern Somalia. There are also reports included from the Second Congo Civil War and the Angolan Civil War to the west. Some of these unusually large conflicts had started well before the scope of our data set and correspondingly very large events in truncated dynamical profiles would be consistent with the scaling description. These events are included in the cumulative profiles in the main text.

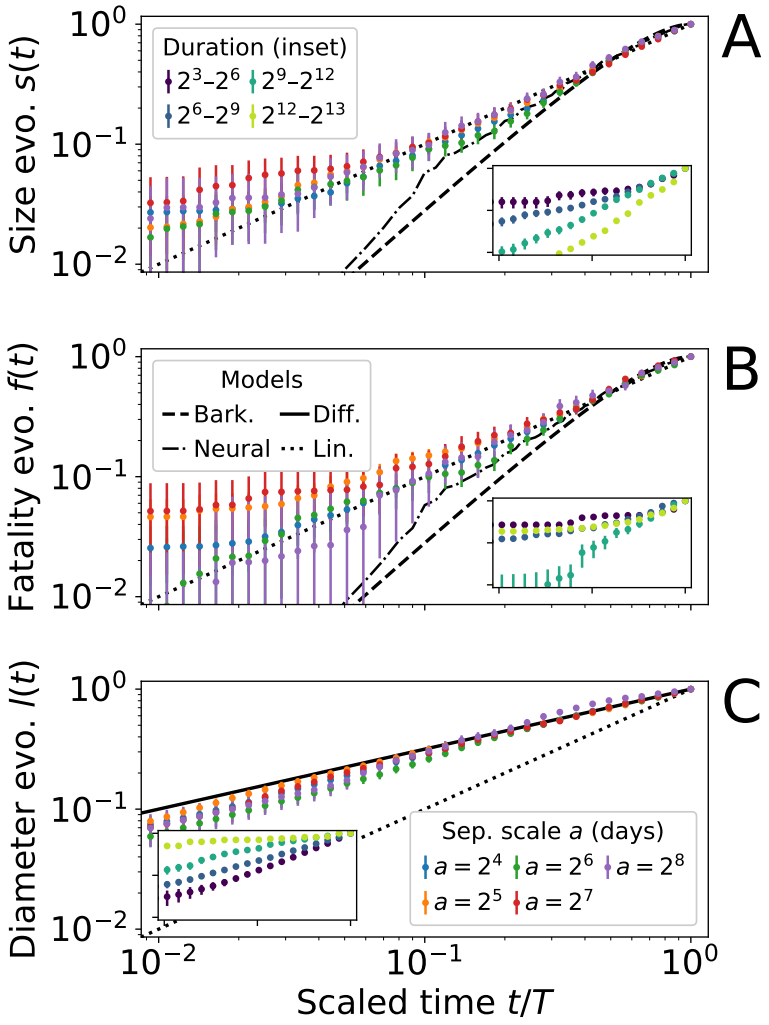


Figure E11: Temporal profiles for VAC events converge to universal profiles similar to those of Battles from Figure 2.13.

E.5 A cumulative temporal profile

We take a non-parametric approach to showing the collapse of conflict temporal profiles, using a cumulative curve because of our small data set. In contrast, rate profile curves are often shown elsewhere such as with neural avalanches [28, 59, 61, 78]. Whereas controlled experiments permit observation of multiple systems with $\gtrsim 10^4$ avalanches, we have at most $< 10^3$ avalanches for ($T \geq 8$ days, $S > 2$ reports, $F > 2$ fatalities), and many fewer for large b and a. For the temporal bins shown in Figure 2.13, the number of samples ranges from < 10 to a few hundred in the best sampled bins even with logarithmic spacing. As a result, the rate temporal profiles show considerable variability as in Figure E10, and the statistical similarity between the profiles is overshadowed by visible noise.

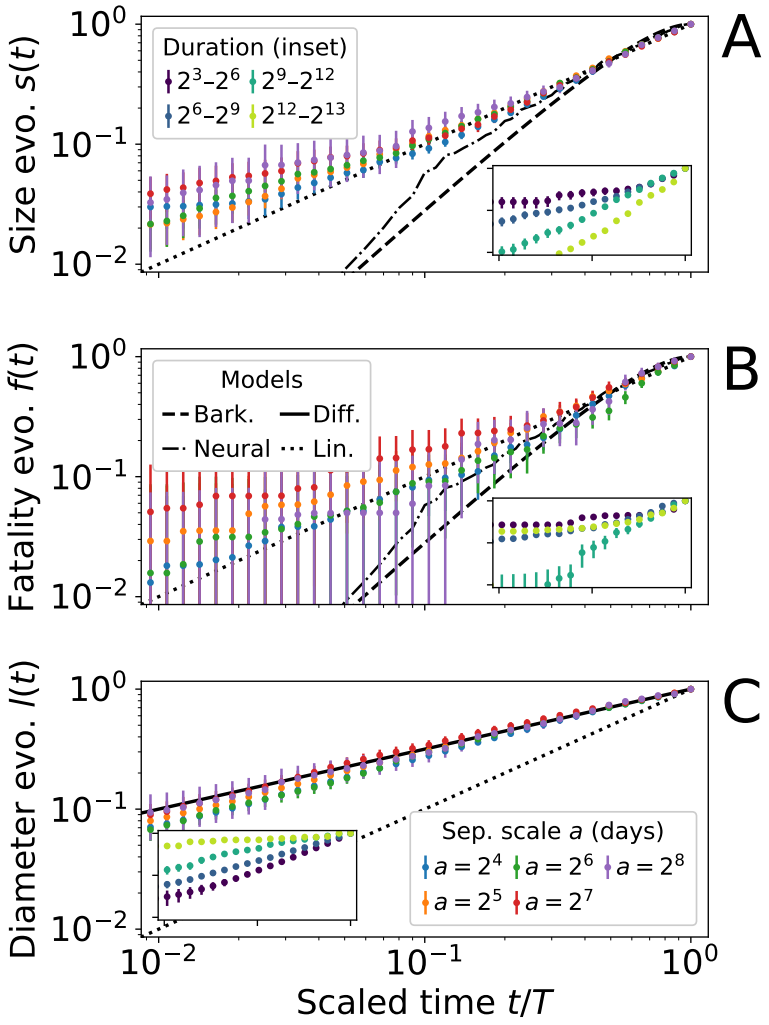


Figure E12: Temporal profiles for Riots/Protests.

The importance of noise is apparent in the example of a rate profile collapse in Figure E10. Despite such noise, we can use an aggregate of many such noisy profiles to measure the dynamical scaling exponents via collapse and compare them with the exponents determined from the scaling relations S vs. T, F vs. T, and L vs. T. In the spirit of the standard rate profile collapse procedure, we average across bins spaced logarithmically by factors of 2 and measure the exponents that return the best collapse by minimizing the logarithm of the geometric variance between the curves. We find agreement with the dynamical exponents measured from comparing the scaling variables directly against one another—though we note that we were not able to successfully collapse the largest conflict avalanches as described in more detail in Figure E10. Thus, various measurements of the dynamics of conflict through exponent relations, direct comparison of scaling variables, and profile collapse are all in agreement across sequences of conflict events lasting from weeks to years.

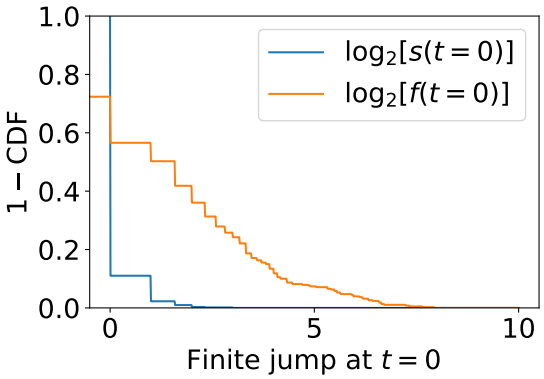


Figure E13: Distribution of finite jump at $t = 0$ in temporal profiles. For sizes (blue), nearly all conflict only involve a single report on the first day which can be accounted for as a lattice bias. For fatalities (orange), a much wider distribution can be accounted for by a lattice bias taken over the ensemble of all fatality profiles for a given time bin $\langle 1/S \rangle (T)$.

To construct the cumulative size profiles $s(t)$, we use the right-handed cumulative distribution function counting the number of events scaled by total size of the conflict, S, and by the total duration, T, on the time axis. This profile is a series of step functions that increments first at $t = 0$, subsequently at every later day on which at least one report is observed, and finally at $t = T$. By definition, all the *size* profiles must end at 1 and they must start at $1/S$ (Figure E13). This offset constitutes a *lattice* bias that disappears geometrically as $S \to \infty$, but many of our profiles involve small avalanches. To account for this bias, we subtract from the profile the value $1/S$, again subtract $1/S$ from $t = T$, and then scale the profile such that it ends at 1. As a result, profiles of avalanches of size $S = 2$, by definition, are flat, and we exclude them from this analysis. The same lattice bias appears in rate profiles since avalanches by definition start with at least one event per time bin. As with the cumulative profile, the finite jump decays geometrically to 0 with the size of the avalanche although we are not aware of any explicit mention of such corrections with neural avalanches—they are typically left uncorrected in collapsed profiles—perhaps because neural avalanches are much bigger. Here, the prevalence of few events in conflict avalanches means that accounting for such biases is essential for capturing the temporal profile collapse.

For fatalities, however, subtracting a similar $1/F$ bias per avalanche is an ill-posed solution to addressing the lattice bias because some reports include no fatalities. Indeed, any number of fatalities may occur at $t = 0$ so there is no *a priori* reason to account for a lattice effect of $1/F$ (Figure E13). Yet, we find a substantial fraction of events occur on the first day, accounting for about 30% of all fatalities for conflicts of duration $T \leq a$ and 10–20% in conflicts $T > a$ and

decreasing in a roughly geometric manner with conflict duration. Motivated by the nearly linear profile between the endpoints, we first take the average over cumulative fatality profiles $\langle f(t) \rangle$ and assume that fatalities occurred with uniform probability across all S reports filed during a conflict avalanche. In other words, such a null model would imply that an average fraction of $\langle 1/S \rangle (T)$ fatalities on the first and last days of a conflict avalanche of duration T. Similar to size profiles, we find that the finite jumps at $t = 0$ and $t = T$ can be largely accounted for by a $1/S$ lattice bias subtracted from the averaged trajectories. Thus, we find a collapse of the temporal profiles for both sizes and fatalities after accounting for substantial lattice bias incurred by the discrete nature of conflicts in the data.

The growth in diameter $l(t)$, however, is not naturally discrete and for which a piecewise interpolation function would serve as a poor approximation. Instead, we treat diameter growth as a continuous function that we approximate using minimal, linear interpolation between the observed distances.

We find that the temporal profiles for VAC and Riots/Protests resemble those of Battles as pictured in Figures E11 and E12. Although the smaller size of Riots/Protests events introduces more variability, we find most of the profiles are largely consistent with those of Battles: the temporal profiles for long avalanches are nearly linear in size and resemble diffusion in diameter. Overall, this coincidence in temporal profiles leads to the surprising possibility that the dynamics of armed conflict are similar across multiple kinds of conflict when observed over a sufficiently large scale.

E.6 Temporal shuffle

As a test of the apparently uniform rate at which conflicts grow in size and fatalities, we show the result of randomly and uniformly placing reports of conflict events on any day between the first and last days within each conflict avalanche in Figure E14 (stipulating that events must occur at both the first and last days). As we might have expected from the flat profiles of size and fatality evolution, $s(t)$ and $f(t)$, respectively, this shuffling procedure does not substantially alter the collapse. For diameter growth $l(t)$, however, the profile changes, deviating in a slightly curving fashion from the scaling behavior observed in the data. These results show that though sizes and fatalities may accumulate for Battles in a way consistent with a uniform rate, the geographic extent of a typical conflict avalanche grows nonlinearly.

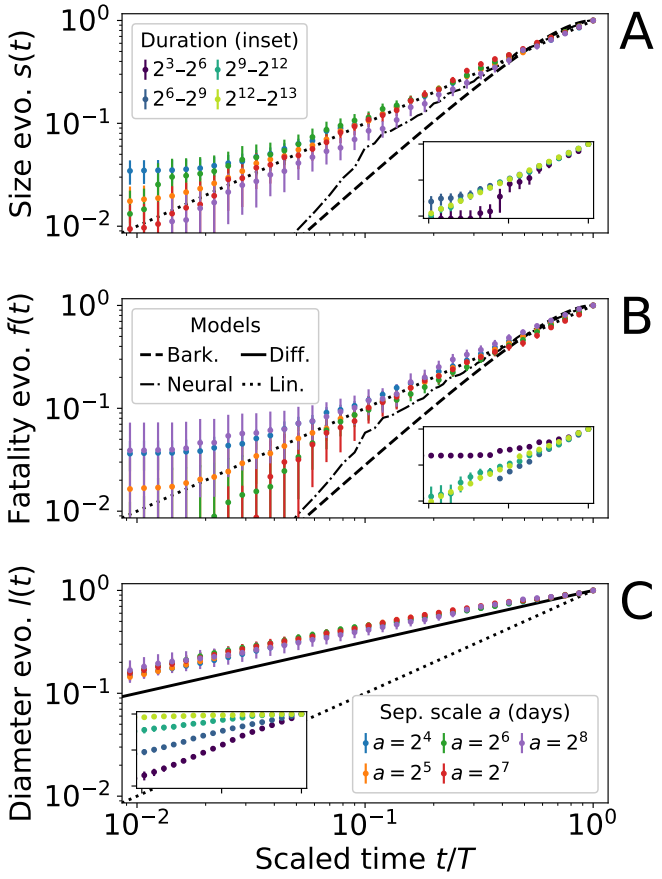


Figure E14: Battle conflict avalanche temporal profiles after time shuffle remained largely unchanged from Figure 2.13 except for the growth in diameter $l(t)$.

E.7 Activated random walkers & percolation growth

We simulate the activated random walkers (ARW) model described in reference [23]. The model consists of "walkers," or particles, living on lattice sites that are inactive when alone but are activated when there are multiple sites on the same site. At every site with multiple walkers, two walkers move to randomly chosen neighbors. As long as any walkers are active, the cascade continues and grows in size S, measured by the cumulative number of walkers that move at each step, and duration T, measured by the number of simultaneous updates over the entire lattice. To produce the distributions we show in Figure E15, we used a square lattice with edge length $l = 10^3$ with free boundary conditions such that walkers that left the boundaries disappeared. Whenever the dynamics stopped, we added a walker at a random site to start the dynamics again and collected 10^4 samples. Using maximum likelihood, we find the distribution exponents for size $\tau = 1.31$ with lower cutoff of 60 and duration $\alpha = 1.55$ with lower cutoff of 45.

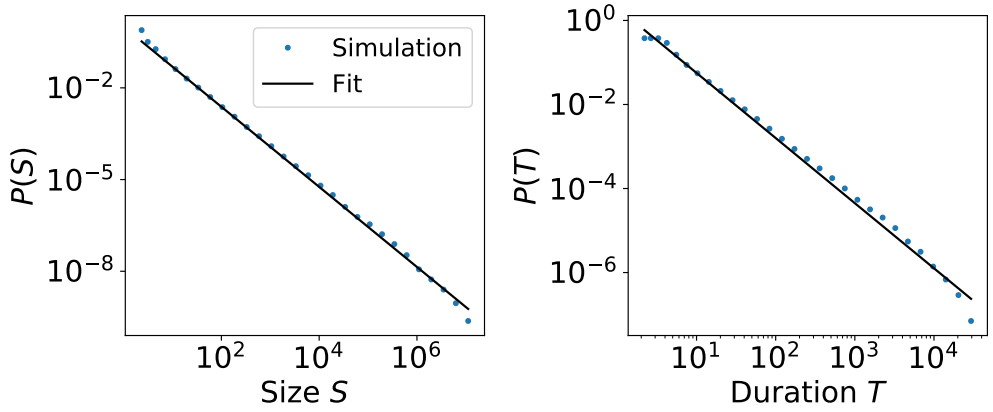


Figure E15: Distributions of size and duration for activated random walkers model in 2D [23].

"Percolation growth" in Table 2.2 refers to the growth of a percolation cluster on a square lattice [14, 76]. This model is akin to the way that forest fires in the model grow on connected clusters of trees at the critical point. After seeding a lattice with an occupied site at the origin, we grow a percolation cluster by occupying the neighbors of any occupied site with some probability p that the connecting bond is "open." We count time in units of shells such that the unoccupied neighbors of any of the currently occupied sites are simultaneously occupied in one time step. Using a square lattice with edge length $l = 10^4$, open bond probability $p = 0.49$, $K = 10^3$ samples, and fitting to trajectories with duration $T \geq 10$, we recover the dynamical exponents mentioned in reference [14]. To recover the distribution exponents, we assume that the percolation clusters live a two-dimensional lattice, yielding $\tau = 2$, and calculate the remaining exponents ν and α using the exponent relations from the main text.

F Appendix for Chapter 3

F.1 Experimental protocol

All subjects were informed about the purpose and goal of the study at the beginning of the experiment and gave consent. After a preliminary survey about experience in sports or performing arts and questions about any conditions that would exclude them from the study (including vision, hearing, and arm motion problems and history of poor experience with virtual reality headsets), they were shown how to use the motion capture suit and virtual reality headset comfortably. The subject was familiarized with the mirror game outside of the virtual reality environment through two quick practice rounds (one hand at a time) with the researcher. Subjects were then instructed to "mirror [simultaneously] the motion, or velocity, of the avatar" where the word "simultaneously" was included in the training conditions because it was unclear if all subjects understood what was implied by mirroring in the untrained condition. When audio cues were used, they were also told, "Try to use the sound to predict the motion of the avatar's hand." Immediately previous to the start of the mirroring task, they were reminded visually by a floating script to "Mirror the hand." Periodically throughout the trial, the comfort of subjects in the virtual environment was assessed verbally. At the end of the experiment, all subjects filled out a post-experiment survey to assess the comfort of the suit and virtual headset, importance of fatigue, clarity of instructions, and to check if they had been following instructions.

A sequence of trials for a single hand consisted of 16 different 30 s trials where the first and last trials were always a fully visible condition. During the experiments, the task was paused every 2–3 minutes to assess the subject for any poor reactions to the virtual environment and to ask explicitly about fatigue. If the subject expressed any sign of fatigue, a rest of time of at least 15 s was taken.

We tested four different experimental conditions including no training and no audio (Visual Only), no training with audio (Audio), training without audio (Train), and training with audio (Train+Audio) each with subject sample size N and unique subject and hand combinations M: $(N = 10, M = 17)$, $(N = 10, M = 10)$, $(N = 7, M = 13)$ and $(N = 8, M = 15)$, respectively. Nearly all subjects participated in two experiments, one with each hand and the first hand chosen randomly. The exceptions were when coding bugs prevented us from continuing the experiment.

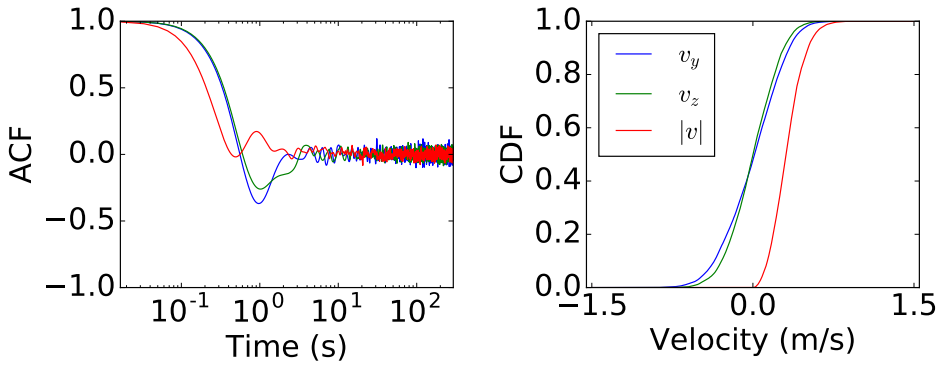


Figure F1: Statistics on the avatar's motion. (left) Autocorrelation function (ACF) of the velocity along the y and z axes on which we assessed performance and the norm velocity $|v|$. There is little structure in the velocities after the 1 s time scale because the motion is aperiodic. (right) Cumulative distribution function (CDF) of the velocities. The velocities are relatively slow and nearly all within a speed of 1 m/s.

For the Train and Train+Audio conditions, subjects were told that the first 5 minutes of the experiment would consist of a practice round with a single break in the middle. During the break, subjects were asked if they had any questions about their performance. When audio cues were used, the experimenter emphasized the instruction to use the audio cue and asked the subjects to explain how they were using the audio cues. If they made incorrect inferences about how the audio corresponded to the motion—for example, one subject thought the volume of the audio changed with the location of the avatar's hand—the experimenter explained to them how they were incorrect. To all subjects, the experimenter explained that the audio cue had pitch proportional to the speed of the avatar and became higher in pitch when the avatar was moving faster and lower when the avatar was slowing down or changing directions.

We collected data from 35 participants, but one subject was excluded from the analysis because of professed disinterest in the experiment and cursory completion of the post-experimental survey that included answering an inapplicable question without any mention or question to the experimentalist. All subjects were assigned to one experimental condition per visit. Subjects ranged in ages from 18–42 with varying levels of experience in physical activities requiring coordination with others. Experimental protocol was approved by the institution's IRB and the HRPO at the DoD.

The motion of the avatar was generated by the experimenter with the goal of keeping it aperiodic and within velocity bounds that would be well tracked by

the PN motion capture suit. In Figure F1, we show the autocorrelation function (ACF) of the avatar's motion and the distribution of velocities. The autocorrelation function shows small periodicities at the 1 Hz time scale but otherwise little other periodicity at longer time scales. The CDFs show that the velocity of the avatar was limited to a small regime bounded by 1 m/s.

F.2 Experimental apparatus architecture

To run the experiment, we combine several commercially available or open source platforms to run the virtual environment, capture the motion of the subject, and train the online learning algorithm. We discuss how these are combined in an overview and then discuss details of the platforms in more detail.

The main components of the system are detailed in Figure 3.2. The apparatus involves running a virtual reality environment on Unreal Engine 4 (UE4), a game development engine. Subjects are immersed in the environment with the Oculus Rift virtual reality headset. We capture their motion using the Perception Neuron motion capture suit. We compare the subject's motion with the prerecorded avatar's on a Python backend and learn the subject's performance landscape the results of which are sent to UE4 to determine the course of the experiment.

UE4 is a standard game development engine used to develop applications for virtual reality environments built on a C++ backend [25]. Since it is widely used, many plugins and features are ready to use, and the Oculus Rift requires no further programming to interact with the three-dimensional environment that we build. We use the environment to display visual instructions to the subjects, manipulate the visual appearance of the avatar, play the audio cue, and provide feedback to the subjects on their performance in the form a green "health bar" above the avatar's head. This environment is displayed in the Oculus Rift virtual reality headset that was originally designed for gaming and is available on the consumer market. It provides a 3-dimensional perspective through two lenses that refresh the visual field at 90 Hz. Although each eye has a high definition 1080p view of the world, the width of the field of view means that pixelation is visible.

To check that this environment was adequate for controlling the visual appearance of the avatar during our experiments, we verified that the internal loop controlling whether or not the avatar was visible was accurate to the tens of milliseconds level. We did this by recording the system clock time every iteration of the loop found it to be accurate within tens of milliseconds. Instead, the limiting factor in how low we can reduce the shortest visual gap or visual ap-

pearance of the avatar is the refresh rate of the headset. This pins us at a lower limit of about 0.1 s which is close to the minimum for human reaction time.

Perception Neuron (PN) is a motion capture suit developed by Noitom. Instead of relying on optical marker tracking, PN is based on a network of inertial measurement units (IMUs) that measure local acceleration and angular velocities. This is a relatively new technology because drift error can become a serious problem for systems not tethered to a fixed coordinate system.[13] Nevertheless, it is the case that in recent years IMU-based systems have made notable advances and easily portable, energy efficient, and significantly cheaper than most optical marker tracking systems.

PN comes with software that rapidly (within a delay of 15 ms) computes and transmits via port the motion of the subject including position, velocity, acceleration, and rotation angles [82]. However, these measurements are processed by a custom algorithm based on proprietary technology and the raw acceleration and orientation data from the suit sensors are inaccessible. Although we cannot inspect the algorithm in detail, we note that the widely-used algorithm used to calculate lower-order moments of motion (velocity and position) are almost always variations on the Kalman filter [83]. Typically, results are more biased by particular assumptions of the algorithm the higher the order of the integral or derivative one takes from the data, so we focus on measuring the velocities of the subjects and do not consider positions or orientations of the body.

For our analysis, it is important that the total latency in our system be below 100 ms which is the lower limit to human reaction time. Across a few tests, we find that the PN system compares remarkably well to other well-tested equipment systems and latency errors are easily below 100 ms.

First, we compare the PN suit with a known standard and well-tested Vicon optical marker tracking system in a local facility. This system provides a different way of measuring the motion of the subject because it tracks the location of each of the markers which can be used to calculate the velocities instead of the acceleration. When properly calibrated, the Vicon system can measure the position of its markers down to millimeter precision and with a latency of single milliseconds. We find that on a computer system with sufficient processing power (otherwise a significant time varying delay is incurred) and when the PN

[13] Drift error refers to the fact that the measurement components cannot measure directly the position or velocity of the IMUs directly, but that they must be calculated from integration of noisy measurements.

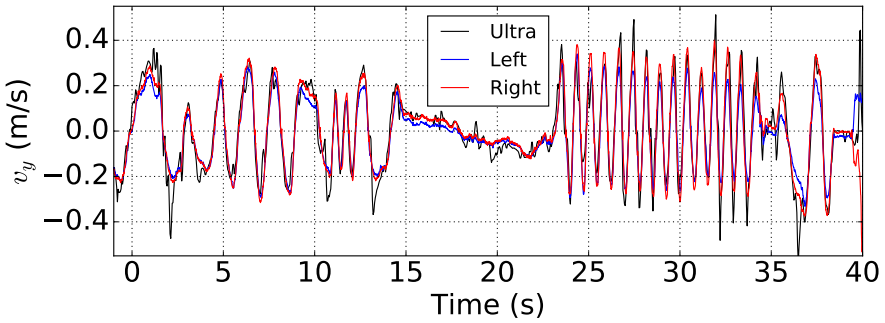


Figure F2: Comparison of velocity with an ultrasound distance meter. A box was held between the two hands of the experiment and moved towards and away from the detector. We show the resulting estimates of the velocities from the detector (black), the Perception Neuron left hand (blue), and the right hand (red). Zero crossings in the velocities agree within 50 ms, below the lower limit of our analysis in performance (Figure 3.3).

suit is physically connected to the computer, latency is well below 50 ms as advertised. The values of the velocities do not agree with those estimated from the Vicon system but they constitute roughly a scaled transformation such that velocities < 1 m/s like those encountered in the avatar's motion do not incur more than 10% error in the conditions we explored. Reassurringly, the zero velocity crossings match almost exactly in the two systems. Given the high accuracy and precision in timing of direction changes but relatively significant errors in the magnitudes, we do not consider directly the magnitudes of the velocities in our analysis.[14]

We furthermore compare the timing of the suit with an ultrasound Vernier Motion Detector. The ultrasound distance meter is reportedly accurate to a single millimeter with a maximum recording frequency of 30 Hz [53]. To compare the PN suit with the detector, we held a box between both hands and moved it towards and away from the detector and measured the velocities along this axis of motion as shown in Figure F2. By interpolating the measured velocities to estimate the zero crossings, we find that disagreements were below 50 ms. Again, we found nothing to suggest that the latency of the PN suit was large enough to affect our results when estimating the velocities. Indeed, we find close agreement between the two systems and the timing of direction changes is precise within few tens of milliseconds.

[14]For the alignment of the subject's motion with the avatar's, we use the magnitude of the measured velocities but we design our cost function to rely on linear differences between the velocities to minimize the effects of scale. In principle, we could also account for such scaling errors by introducing a scaling error parameter, but we did not find this necessary.

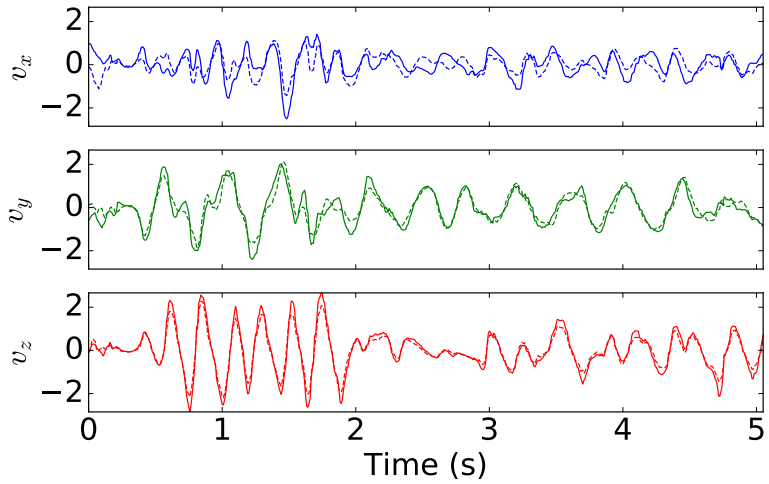


Figure F3: Example velocity trajectories as measured using the Perception Neuron suit when the subjects hands are clasped together. Errors in the inferred orientation of the hands lead to larger relative errors in the x direction which corresponds to the axis pointing from the subject to the avatar (Figure 3.1). We ignore this axis for our analysis.

Finally, we tested the suits by fixing the hands together checking for consistency between the two velocity trajectories, an example of which is shown in Figure F3. Here, we found that rotational errors in the orientation of the arms would leads to differences in timing and velocity along the x axis (pointing from the subject to the avatar). The other axes y and z seemed to be less affected by this problem. For our analysis, we do not consider the x direction.

F.3 Dynamic time warping (DTW)

Although spectral techniques provide one way to compare motion in coordination tasks (including variants of cross correlation, wavelet analysis, and recurrence plots), we use DTW to align the velocity trajectories of the subject with that of the avatar [36, 54, 67]. One major issue with using spectral techniques to identify temporal delays in aperiodic motion is ambiguity in deciding which local peak in the time-lagged cross-correlation corresponds to the time delay especially when individuals are failing to mirror the partner well. DTW, on the other hand, finds the globally optimal alignment and thus can use global information to resolve these ambiguities in local alignment. Overall, DTW is a computationally efficient way of accounting for strong local nonlinearities in multiple dimensions when comparing motion trajectories.

The goal of DTW is to align two curves by allowing local temporal stretching. This is accomplished by a combinatorial algorithm that involves finding the

optimal path in the matrix defined by Eq F1

$$D_{ij} = \left|\vec{v}_s(t_i) - \vec{v}_a(t_j)\right| + \lambda g(i,j) \tag{F1}$$

The corresponding path minimizes the total accumulated distance between the warped curves with some extra cost g and strength of regularization λ for disfavoring unrealistic trajectories. The resulting path defines a warped time $\tilde{t}_i$ that gives a measure of the local time delay (or anticipation) that the subject shows while tracking the avatar.

The first term in Eq F1, what we call the cost, is often quadratic in the distance. For our system, the linear distance between the two velocities is essential because of both human motion and limitations of the experimental apparatus that we are using to capture motion. Some subjects change directions very rapidly to correct for errors in direction and this results in large velocities with temporal profiles that are almost correct while velocity magnitude deviations can be large. With a cost function that grows superlinearly with the velocity, error peaks would be aligned even at the expense of many features smaller in magnitude but indicative of mirroring. Furthermore, we have found that the motion capture system can overestimate absolute velocities especially when the acceleration is large. Thus, peaks in velocities are especially prone to systematic error. In both these cases, a superlinear distance measure between the velocities would favor weight large peak matching instead of aligning the many smaller features of trajectories, and so we rely on a cost linear with distance between the trajectories.

As for second term in Eq F1, the regularization, we design g to avoid situations in which the subject is impossibly anticipating the motion of the avatar (as can happen when motion seems briefly periodic) and when the inferred delay is so large that subjects would have to remember far into the past while memorizing new motion simultaneously. To design a sensible regularization function in Eq F1, we find that when $\lambda = 0$ DTW will find some trajectories where the subject is leading the avatar by seconds or is behind the avatar by seconds. These trajectories tend to appear in cases where the subject is doing very poorly and so it is difficult to find a temporally local trajectory that resembles the avatar's motion. They also occur where brief periodicities mean a phase shift of 2π overlays the trajectories. Noting that when the avatar is fully visible, subjects infrequently venture outside a time delay of $3/2\,\mathrm{s}$ or are ahead by more than $1/2\,\mathrm{s}$, we define g to be zero within the interval $\Delta t \in [-1/2\,\mathrm{s}, 3/2\,\mathrm{s}]$ and then sharply increasing outside of that range.

$$g(i,j) = \begin{cases} 0, & |t_i - t_j + 1/2| < 1 \\ |t_i - t_j + 1/2|^6, & |t_i - t_j + 1/2| \geq 1 \end{cases} \tag{F2}$$

with $\lambda = 10^{-3}$ controlling the strength of regularization.

To calculate alignment, we first use FastDTW which can calculate the time warp in nearly linear time instead of quadratic time [69]. If the found trajectory ventures outside of the bounding interval $\Delta t \in [-1/2\,\mathrm{s}, 3/2\,\mathrm{s}]$, we then solve the problem using our own (slower) implementation including the regularization. We find that about 60% of the untrained trials were regularized whereas only 35% of the trained trials were. We might expect this difference because untrained individuals typically do not replicate the trajectory of the avatar as well and the algorithm is more prone to misaligning stretches of motion.

F.4 Velocity error thresholds

In the main text, we only consider the temporal delays ϵ^* to characterize the performance of the subjects. Here, we explore the effect of a threshold in the alignment of the velocities ϵ_v^*. In agreement with our results when only considering the time delay threshold as shown in Figure 3.3, we find that Visual Only performance is much worse when compared to the other conditions, there is a range of timescales from about 200–800 ms where the largest variation in performance between conditions appear, and that beyond those limits performance variation is small. We also find that audio cues have a larger effect on performance for the shortest time scales, in contrast with training where performance is worse at faster time scales. Overall, inclusion of the velocity error threshold reaffirms our results about the change in mean performance in the main text where we only consider the time delay threshold.

To measure velocity error, we focus only on normalized velocities. We ignore the speed because the size of the avatar does not scale with the size of subject and because we find that the PN suit system is prone to scaling errors with velocity estimation (See SI Section F.2). To compare the velocity directions, we define the error to be

$$\epsilon_v(\tilde{t}) = \frac{1}{2} - \frac{1}{2}\frac{\vec{v}_a(\tilde{t}) \cdot \vec{v}_s(\tilde{t})}{\left|\vec{v}_a(\tilde{t})\right| \left|\vec{v}_s(\tilde{t})\right|} \tag{F3}$$

that is 1 when the velocities are anti-aligned, 1/2 when they are orthogonal, and 0 when they are exactly parallel. As with the timing delays, we choose a threshold ϵ_v^* and measure when subjects are below the fixed threshold. Now instead of a single threshold, we have two thresholds in both velocity ϵ_v^* and

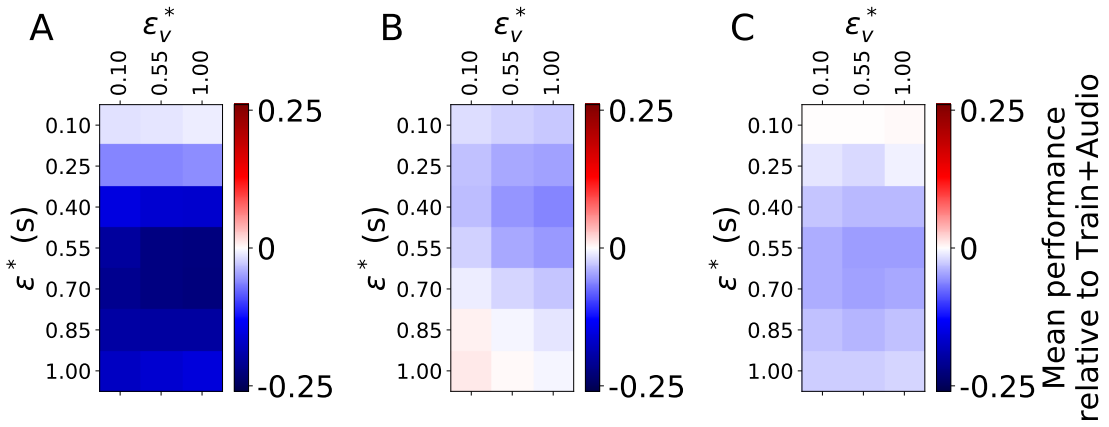


Figure F4: (A) Difference in average performance of Visual Only relative to Train+Audio for different combinations of the time delay ϵ^* and velocity direction ϵ_v^* threshold. Rightmost column corresponds to the difference between the mean performance values shown in the bottom of Figure 3.3. (B) Comparison of Train with Train+Audio. (C) Comparison of Audio with Train+Audio. Negative values (blue) indicate that performance is worse relative to Train+Audio.

timing ϵ^*,

$$\hat{\pi}(\tau, f) = \frac{1}{\tilde{T}+2}\left(1 + \sum_{\tilde{t}}^{\tilde{T}} \Theta\left[\epsilon^* - \left|\epsilon\left(\tilde{t}\right)\right|\right] \times \Theta\left[\epsilon_v^* - \left|\epsilon_v\left(\tilde{t}\right)\right|\right]\right). \tag{F4}$$

We calculate the mean performance $\langle \pi_i \rangle_{\epsilon^*}$ over the average per subject as we change the thresholds. To do this, we infer the entire predicted landscape given the data points from Eq F4 for every combination of ϵ^* and ϵ_v^* of interest. We summarize the results of the predicted landscapes in Figure F4 where we show the change in average performance from (A) Visual Only to Train+Audio, (B) from Train to Train+Audio, then (C) from Audio to Train+Audio. The deepness of the blue indicates how much worse average performance in the shown condition is relative to Train+Audio, whereas red indicates relatively better performance. Consistent with results in the main text, subjects do much better with either training or audio than in the Visual Only condition as indicated by the blue-dominated leftmost graph. The rightmost column of these graphs corresponds to the differences in the mean performance values shown in Figure 3.3. In the Train and Audio conditions in Figs. S5B and C, the enhancement for Train+Audio is concentrated at $0.2\,\mathrm{s} < \epsilon^* < 0.8\,\mathrm{s}$ across all ϵ_v^*. At the smallest

shown $\epsilon_v^* = 0.1$, we again are at the limit where subjects all perform poorly because the threshold for error is so low, and so we find, as expected, a narrowing the range of performance across all conditions.

F.5 Learning the performance landscape

Mapping the topology of the performance landscape by measuring every combination of parameters (τ, f) is infeasible. If, however, we assume that that the average performance landscape changes smoothly as we change the visual appearance of the avatar with parameters τ and f, we can measure a few key points and interpolate the missing ones. We model any particular measurement of the subject i's performance as a stochastic variable.

$$\pi_{\mathrm{i}}^*(\tau, f) = p_{\mathrm{i}}(\tau, f) + \eta_{\mathrm{i}}, \tag{F5}$$

where $\pi \in [0, 1]$ has been mapped to the real line with the inverse logistic transform $\pi^* = -\log[1/\pi - 1]$ such that $0 \to -\infty$ and $1 \to \infty$. The first term in Eq F5 refers to the variation inherent to the subject, embodying how fluctuations in the performance landscape are correlated across different τ and f. It has mean $\langle p_{\mathrm{i}} \rangle = \mu_{\mathrm{i}}$. The second term in Eq F5 refers to an independent source of statistical noise η_{i} with mean $\langle \eta_{\mathrm{i}} \rangle = 0$ and width $\left\langle \eta_{\mathrm{i}}^2 \right\rangle = \alpha_{\mathrm{i}}^2$. The expected covariance between any two measurements is then

$$\begin{aligned}\langle \pi_{\mathrm{i}}(\tau, f)\pi_{\mathrm{i}}(\tau', f') \rangle = \langle [p_{\mathrm{i}}(\tau, f) - \mu_{\mathrm{i}}]\,[p_{\mathrm{i}}(\tau', f') - \mu_{\mathrm{i}}] \rangle + \\ \delta_{\tau,\tau'}\delta_{f,f'}\alpha_{\mathrm{i}}^2\end{aligned} \tag{F6}$$

with delta function $\delta_{x,x'} = 1$ only if $x = x'$ and 0 otherwise.

We model the distribution characterized by the covariance in Eq F6 using Gaussian process regression (GPR). This technique is equivalent to a multivariate normal distribution of the observed data points with covariance that typically decays with increasing distance between two parameter sets (τ, f) and (τ', f') [9, 66], where the decay length determines the typical size of local features in the performance landscape.

When modeling the covariance function during the course of an experiment, we used different formulations for running the experiments including radial basis kernel G or an exponential kernel K, which are both common parameterizations of the kernel function. They are, respectively,

$$K_{\mathrm{i}}(d) = \theta_{\mathrm{i}} \exp\left(-d^2/2s_{\mathrm{i}}^2\right) \tag{F7}$$

$$G_{\mathrm{i}}(d) = \phi_{\mathrm{i}} \exp\left(-d/\lambda_{\mathrm{i}}\right) \tag{F8}$$

with coefficients ϕ_i and θ_i and scale parameters λ_i and s_i. Typically, the diagonal terms representing the noise are considered separate from the kernel function such that the covariance is the sum of the two:

$$\langle \pi_i(\tau, f)\pi_i(\tau', f')\rangle = K_i(d) + \alpha_i^2 \delta_{\tau,\tau'}\delta_{f,f'} \tag{F9}$$

In addition to the kernel, we must also decide on a geometry for the performance landscape that determines the distance d in Eq F9. We use the geodesic distance on a hemisphere to observe the singularities at $f = 1$ and $f = 0$ corresponding to the north and south poles, where the longitude lines of performance at different τ all converge [31].

Combining these elements, the log-likelihood of the set of observed data points for subject i is given by the multivariate normal distribution

$$\log L_i \propto -\sum_{\substack{x=(\tau,f)\\ x'=(\tau',f')}} \pi_i(x) K_{x,x'} \pi_i(x') - \frac{1}{2}\log|K| \tag{F10}$$

If the hyperparameters are not optimized at every step, the parameter combination (τ, f) of maximal predicted uncertainty is deterministic after every measurement because the log-likelihood does not depend on the value of performance measured at that point. If the hyperparameters are optimized, then the parameter combination with maximal uncertainty can change, but the computational cost of the calculation can be much higher.

We used different formulations of GPR depending on the experimental condition. For the untrained conditions, we used a radial basis kernel function without hyperparameter optimization at every trial. Thus for all the untrained trials, all the same points were measured on the performance landscape in the same order, though ensuring that the parameter combination with maximum uncertainty was selected next. For the trained conditions, we used an exponential kernel with online hyperparameter optimization. The difference in procedures means that the sets of points collected for the untrained trials are fixed throughout all subjects whereas for the trained subjects the measured points are scattered differently throughout the parameter space. When we model the aggregate landscapes at the end, however, we model all performance landscapes in the same way.

F.6 Aggregate performance landscape

We combine the measured values across all subjects for a given experimental condition to construct an aggregate performance landscape that captures

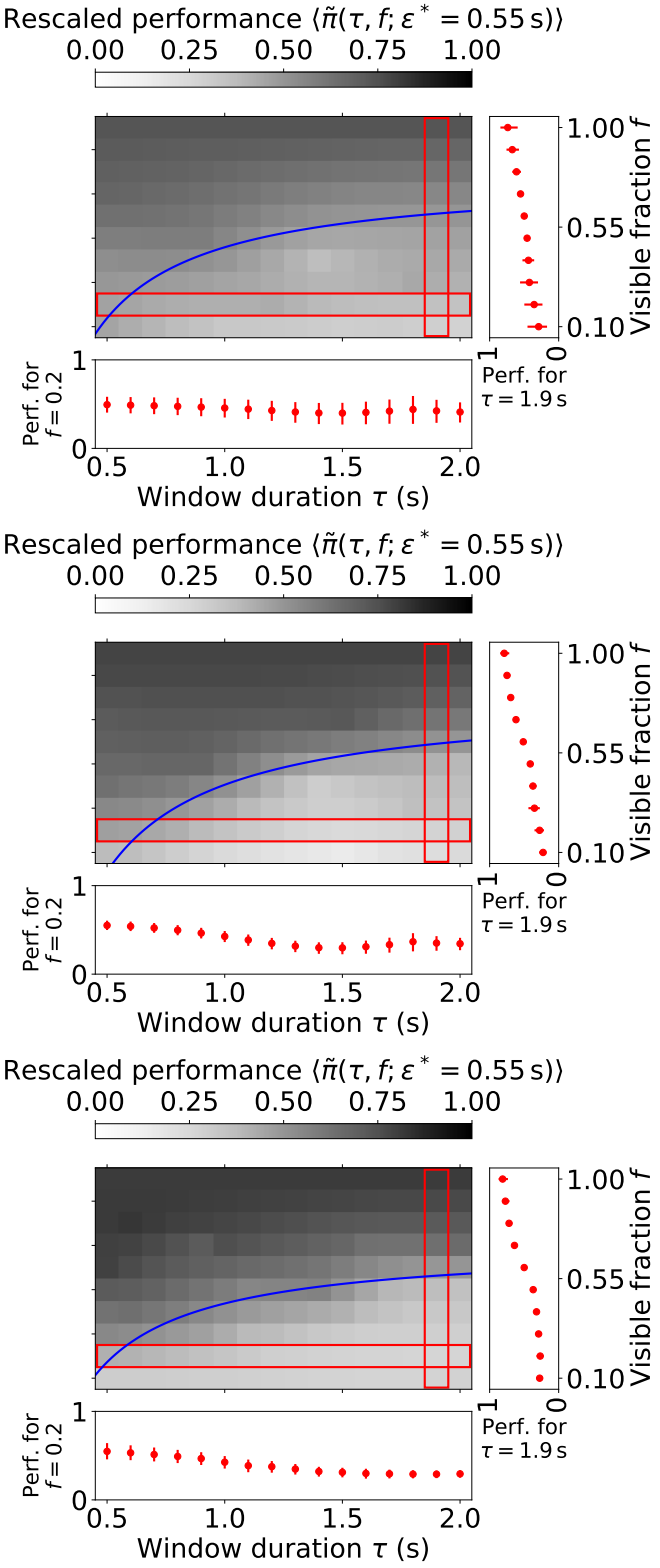


Figure F5: Aggregated performance landscapes for the Visual Only, Audio, and Train conditions from top to bottom as predicted using Gaussian process regression. Blue lines are best fits to level contours of the rescaled performance $\langle\tilde{\pi}\rangle = 1/2$ of the form given by Eq 3.1.

subject-specific fluctuations and structure common across subjects. We add onto Eq F5 a common landscape P across all subjects and noise H that is iid for every observation made with $\langle H \rangle = 0$ and $\left\langle H^2 \right\rangle = \beta$.

$$\pi_{\mathrm{i}}(\tau, f) = p_{\mathrm{i}}(\tau, f) + \eta_{\mathrm{i}} + P(\tau, f) + H \tag{F11}$$

The terms P and H do not have indices because they are shared across all subjects.

Then, we assume the corresponding covariance matrix has the block form

$$\begin{aligned} K_{\mathrm{ij}}(d) = a\, K_{\mathrm{co}}(d) + \delta_{\mathrm{ij}} \left(b_{\mathrm{i}}\, K_{\mathrm{i}}(d) + \delta_{\tau,\tau'} \delta_{f,f'} \alpha_{\mathrm{i}}^2 \right) + \\ \delta_{\mathrm{ij}} \delta_{\tau,\tau'} \delta_{f,f'} \delta_{\mathrm{n,n'}} \beta^2 \end{aligned} \tag{F12}$$

with common kernel K_{co}, weight coefficients a and b, and each data point has a unique index n. Here, we use the more flexible Màtern kernel,

$$K(d) = \Theta \frac{2^{1-\nu}}{\Gamma(\nu)} (d/\lambda)^{\nu} \kappa_{\nu}(d/\lambda) \tag{F13}$$

where κ is the modified Bessel function of the second kind [40]. The Màtern kernel has a smoothness parameter ν such that when $\nu = 1/2$ (with $0 \leq \nu \leq 1/2$ on a spherical surface) it is the exponential kernel from Eq F8.

Given the large number of parameters, we regularize the problem by imposing a sparseness constraint on the coefficients of the subjects when maximizing likelihood

$$f(\{\theta_{\mathrm{i}}\}) = \frac{1}{\mathrm{N}} \sum_{\mathrm{i}=1}^{\mathrm{N}} |\theta_{\mathrm{i}}| \tag{F14}$$

such that subject specific terms in Eq F12 are driven to zero when the common landscape is sufficient to describe the subject's behavior. When the coefficient θ_{i} is driven to 0, the other parameters for that subject's kernel become irrelevant so this is an efficient way of reducing the dimensionality of the parameter space. We also find that the noise terms are often driven to 0 if they are not constrained even though fluctuations in the data seems to be strongly important even when comparing the same subject's performance for $f = 1$. Therefore, we add a weak regularization for the noise terms

$$\frac{1}{10^3 \mathrm{N}} \sum_{\mathrm{i}=1}^{\mathrm{N}} |\alpha_{\mathrm{i}} - 1/2| \tag{F15}$$

As validation of our choice of the structure of the covariance matrix, we find that the ratio of coefficients $a/\langle b \rangle_i$ is not driven to 0, but varies from 0.2 to 4.8 indicating that shared structure in the performance landscape is important.

The maximum likelihood parameters we find describe a model that agrees well with the measured data points across all 24 combinations of the 4 experimental conditions and 6 values of ϵ^* with correlation coefficient of $\rho = 0.98$ when comparing π with $\hat{\pi}$. We also perform a cross-validation test by leaving one data point out of the data (such that the covariance matrix is one row and one column smaller) and comparing the prediction with the measured data point and find still yet $\rho = 0.95$ [1]. For each single landscape (for a fixed ϵ^* and experimental condition), we check directly the prediction error of this cross-validation procedure and find that the average norm error per landscape is less than 0.01. These statistics show that we have found a good fit to the performance landscape.

As we describe in the main text, we must aggregate over rescaled performance landscapes to show the transition curve shown in Figure 3.1C. First, we rescale them such that they all reach the value of $\tilde{\pi} = 1/2$ at $(\tau, f) = (2.0, 0.6)$. Then, we set $\tilde{\pi}(f = 1) \approx 1$ and $\tilde{\pi}(f = 0) \approx 0$. The precise values for this last step in rescaling are chosen for maximum contrast, but the shape of the transition region does not depend on the upper and lower limits of the rescaled performance landscape. The result of this aggregation for the Train+Audio condition is in Figure 3.1C and the other conditions are shown in Figure F5.

F.7 Modeling the transition

In the main text, we fit the level curve of performance in the region between high and low performance using a linear relation between the visible duration τ_{vis} and the invisible duration τ_{inv} parameterized by the two constants a and b in Eq 3.1. We find this curve by minimizing the total value of the boxes that the contour passes through

$$C = \sum_{k=0}^{K} \pi\left(\tau_k, f = \frac{1}{1+b} - \frac{a}{(1+b)\tau}\right)^2 \tag{F16}$$

where $\tau_K = 2\,\text{s}$. Since we restrict our contour to the limits of the parameter space we explore in this work, however, Eq F16 can be minimized by simply reducing the length of the contour.

In order to ensure that the contour passes through the grid and does not minimize length at the expense of fitting the level curve, we normalize by the length

Transition contour parameters

Condition	a	b
Visual Only	0.41	0.28
Train	0.38	0.42
Audio	0.49	0.25
Train+Audio	0.34	0.50

Table 4.2: Transition contour parameters. Parameters found for Eq 3.1 using objective function in Eq F19. Compared to the other conditions, Train+Audio has a flatter transition zone showing that the transition to poor performance varies less with τ.

of the contour on the grid. This length is

$$L = \int_{\tau=\tau_0}^{\tau=2} \sqrt{1 + \frac{df}{d\tau}^2}\, d\tau \tag{F17}$$

where τ_0 corresponds to value of τ where the contour crosses the bottom limit of the measured performance landscape where $f = 0.1$,

$$\tau_0 = \frac{a}{-0.1b - 0.9} \tag{F18}$$

Combining Eqs F16 and F17, we have the objective function that we use to find the parameters in Eq 3.1

$$\min_{a,b} C(a,b) = C(a,b)/L(a,b). \tag{F19}$$

The discreteness of the landscape means that a gradient-based minimization routine will fail. Instead, we evaluate the function across a grid of a and b to find the optimal solution. For all the experimental conditions, we find the values of a and b to yield very similar transition curves. The values are shown in Table 4.2.

F.8 Decay distributions

When we inspect the durations of time t that subjects are tracking the avatar closely, we find that the distributions $p(t)$ can be described by three main classes characterized by either an exponential tail, a gamma-like function with a dearth of the shortest decay times, or a heavier-tailed distribution. Although the exponential decay is a signature of a memoryless process, the remaining two distributions indicate that the dynamics of how subjects are tracking the motion of

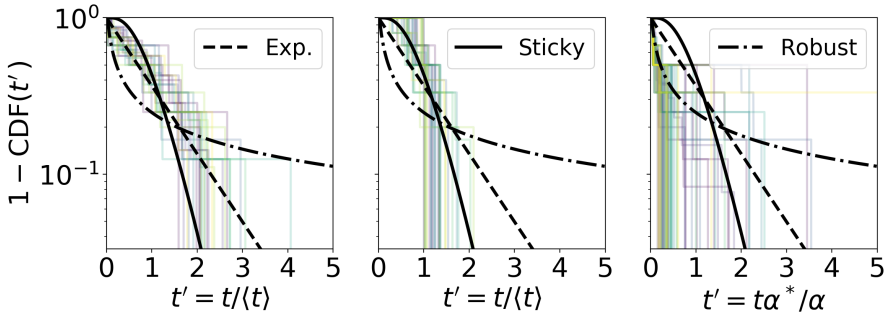


Figure F6: Distribution of decay times for stable runs as indicated by shaded regions in Figure 3.1B. (left) The most frequently occurring distribution is close to an exponential decay. (middle) "Sticky" distributions show a dearth of very short decay times. (right) "Robust" distributions show a power-law like decay with a heavy tail.

the avatar are not generated from an time-independent process. Here, we provide more detail on a few possible explanations for the form and origin of these distributions.

The predominant class of distributions are exponential. In the language of control theory, exponential decays are considered a "first to failure" process in a multicomponent system where failure manifests when the first component fails. This intuition suggests that while the average decay rate of the subject might depend on subject handedness, fatigue, difficulty of the task, or other factors, much of the variation around the average can be explained by a memoryless process as if the subject is susceptible to random fluctuations that lead to failure.

In contrast, we find another class of decay distributions that show a characteristic depletion of short decay times as indicated by a flat region of the CDF at small durations. If it were the case that subjects were not to decay straight from success (S) to failure (F) at mirroring but first had to decay to intermediary states behaviorally indistinguishable from S, we say subjects show "sticky" mirroring:

$$\mathrm{S}_0 \overset{k_0}{\rightarrow} \mathrm{S}_1^* \overset{k_1}{\rightarrow} \cdots \overset{k_{N-1}}{\rightarrow} \mathrm{S}_N^* \overset{k_N}{\rightarrow} \mathrm{F} \tag{F20}$$

with decay rate constants k_i. For $N > 1$, we would expect a flat region in the complementary CDF near $t = 0$ whose extent depends on the number of intermediary states before decay. Since the average time to decay is only determined by the sum of the rate constants $K = \sum_{i=0}^{N} k_i$, we write the complementary CDF of decay times, otherwise known as the survival function, as a function of a single

rate constant

$$1 - \mathrm{CDF}(t') = e^{-Kt'} \sum_{\mathrm{n}=0}^{\mathrm{N}} \frac{K^{\mathrm{n}} t'^{\mathrm{n}}}{\mathrm{n}!} \tag{F21}$$

where the distributions have been rescaled such that $K = 1$ in Figure F6. In the limit of $\mathrm{N} \to \infty$, we recover the gamma distribution. We find that the measured values of N as calculated with maximum likelihood are concentrated at smaller values. Over 50% of the observed values smaller than or equal to 5 when $\epsilon^* = 1/2\,\mathrm{s}$, suggesting that enhanced dynamical stability corresponding to the "sticky" distribution is slight.

In the remaining trials, we find distributions dominated by very short decay times but with a heavy tail of "robust" long runs. With the exponential and "sticky" distributions, we observe dynamics that are consistent with subjects occupying success or failure states separated by "energy barriers." When the dynamics are dominated by the time it takes to escape a local energy minimum, we may expect an exponential distribution for decay times. When one energy minimum becomes strongly dominant such that there is little switching, however, the dynamics will instead be dominated by the width of the local energy basin. Using this intuition, we might expect that the first passage time for simple diffusion as shown in Figure F6 model the data better than the other distributions,

$$1 - \mathrm{CDF}\,(t') = 1 - \sqrt{\frac{\alpha}{\pi}} \int_{1/30}^{t'=t\alpha^*/\alpha} t^{-3/2} e^{-\alpha/t}\, dt. \tag{F22}$$

Here, the lower limit is important and is given by our interpolation of the velocity trajectories at 30 Hz. We cannot get a scaling collapse by rescaling by the mean. Instead, we rescale by the parameter α using a constant $\alpha^* = 0.05$ as we show in Figure F6. For robust mirroring, it is as if the subject is trapped in some wide region characterized by successful mirroring.

Appendices references

[1] David M Allen. "Mean Square Error of Prediction as a Criterion for Selecting Variables". In: *Technometrics* 13.3 (Aug. 1971), pp. 469–475.

[2] *An Introduction to Linear Programming and Game Theory*. en. 3rd. Hoboken, NJ, USA: John Wiley & Sons, Inc., 2008.

[3] Erik Aurell and Magnus Ekeberg. "Inverse Ising Inference Using All the Data". en. In: *Physical Review Letters* 108.9 (Mar. 2012), p. 090201. ISSN: 0031-9007, 1079-7114. DOI: 10.1103/PhysRevLett.108.090201.

[4] John Barton and Simona Cocco. "Ising Models for Neural Activity Inferred via Selective Cluster Expansion: Structural and Coding Properties". en. In: *Journal of Statistical Mechanics: Theory and Experiment* 2013.03 (Mar. 2013), P03002. ISSN: 1742-5468. DOI: 10.1088/1742-5468/2013/03/P03002.

[5] Lawrence Baum. "Comparing the Policy Positions of Supreme Court Justices from Different Periods". en. In: *Western Political Quaterly* (1988), pp. 509–521.

[6] John M. Beggs and Dietmar Plenz. "Neuronal Avalanches in Neocortical Circuits". en. In: *The Journal of Neuroscience* 23.35 (Dec. 2003), pp. 11167–11177. ISSN: 0270-6474, 1529-2401. DOI: 10.1523/JNEUROSCI.23-35-11167.2003.

[7] William S. Bialek. *Biophysics: Searching for Principles*. en. Princeton, NJ: Princeton University Press, 2012. ISBN: 978-0-691-13891-6.

[8] William Bialek and Rama Ranganathan. "Rediscovering the Power of Pairwise Interactions". en. In: *arXiv:0712.4397 [q-bio]* (Dec. 2007). arXiv: 0712.4397 [q-bio].

[9] Christopher M Bishop. *Pattern Recognition and Machine Learning*. Singapore: Springer Verlag, Aug. 2006.

[10] Michael J. Bommarito and Ahmet Duran. "Spectral Analysis of Time-Dependent Market-Adjusted Return Correlation Matrix". en. In: *Physica A: Statistical Mechanics and its Applications* 503 (Aug. 2018), pp. 273–282. ISSN: 03784371. DOI: 10.1016/j.physa.2018.02.091.

[11] Tamara Broderick et al. "Faster Solutions of the Inverse Pairwise Ising Problem". en. In: *arXiv:0712.2437 [cond-mat, q-bio]* (Dec. 2007). arXiv: 0712.2437 [cond-mat, q-bio].

[12] J.O. Caldecott. *An Ecological and Behavioral Study of the Pigtailed Macaque*. Karger, 1986.

[13] John L. Cardy, ed. *Finite Size Scaling*. Vol. 2. New York: North-Holland, 1988.

[14] S. Clar, B. Drossel, and F. Schwabl. "Scaling Laws and Simulation Results for the Self-Organized Critical Forest-Fire Model". en. In: *Physical Review E* 50.2 (Aug. 1994), pp. 1009–1018. ISSN: 1063-651X, 1095-3787. DOI: 10.1103/PhysRevE.50.1009.

[15] Aaron Clauset. "Trends and Fluctuations in the Severity of Interstate Wars". en. In: *Science Advances* 4.2 (Feb. 2018), eaao3580. ISSN: 2375-2548. DOI: 10.1126/sciadv.aao3580.

[16] Aaron Clauset, M. E. J. Newman, and Cristopher Moore. "Finding Community Structure in Very Large Networks". en. In: *Physical Review E* 70.6 (Dec. 2004), p. 066111. ISSN: 1539-3755, 1550-2376. DOI: 10.1103/PhysRevE.70.066111.

[17] Aaron Clauset, Cosma Rohilla Shalizi, and M. E. J. Newman. "Power-Law Distributions in Empirical Data". en. In: *SIAM Review* 51.4 (Nov. 2009), pp. 661–703. ISSN: 0036-1445, 1095-7200. DOI: 10.1137/070710111.

[18] Aaron Clauset, Maxwell Young, and Kristian Skrede Gleditsch. "On the Frequency of Severe Terrorist Events". en. In: *Journal of Conflict Resolution* 51.1 (Feb. 2007), pp. 58–87. ISSN: 0022-0027, 1552-8766. DOI: 10.1177/0022002706296157.

[19] S. Cocco and R. Monasson. "Adaptive Cluster Expansion for Inferring Boltzmann Machines with Noisy Data". en. In: *Physical Review Letters* 106.9 (Mar. 2011), p. 090601. ISSN: 0031-9007, 1079-7114. DOI: 10.1103/PhysRevLett.106.090601.

[20] Thomas M. Cover and Joy A. Thomas. *Elements of Information Theory*. 2nd. Hoboken: John Wiley & Sons, 2006.

[21] Bryan C. Daniels, David C. Krakauer, and Jessica C. Flack. "Control of Finite Critical Behaviour in a Small-Scale Social System". en. In: *Nature Communications* 8 (Feb. 2017), p. 14301. ISSN: 2041-1723. DOI: 10.1038/ncomms14301.

[22] Jason Davies. "Poisson-Disc Sampling". In: ().

[23] Ronald Dickman et al. "Paths to Self-Organized Criticality". en. In: *Brazilian Journal of Physics* 30.1 (Mar. 2000), pp. 27–41. ISSN: 0103-9733. DOI: 10.1590/S0103-97332000000100004.

[24] Z Eisler, I Bartos, and J Kertész. "Fluctuation Scaling in Complex Systems: Taylor's Law and Beyond". In: *Advances in Physics* 57 (2008), pp. 89–142.

[25] Epic Games. "Performance and Profiling". In: (2017).

[26] J. C. Flack. "Multiple Time-Scales and the Developmental Dynamics of Social Systems". In: *Philosophical Transactions of the Royal Society B: Biological Sciences* 367.1597 (July 2012), pp. 1802–1810. DOI: 10.1098/rstb.2011.0214.

[27] Santo Fortunato and Claudio Castellano. "Scaling and Universality in Proportional Elections". en. In: *Physical Review Letters* 99.13 (Sept. 2007), p. 138701. ISSN: 0031-9007, 1079-7114. DOI: 10.1103/PhysRevLett.99.138701.

[28] Nir Friedman et al. "Universal Critical Dynamics in High Resolution Neuronal Avalanche Data". en. In: *Physical Review Letters* 108.20 (May 2012), p. 208102. ISSN: 0031-9007, 1079-7114. DOI: 10.1103/PhysRevLett.108.208102.

[29] Andrea Giometto et al. “Scaling Body Size Fluctuations”. In: *Proceedings of the National Academy of Sciences* 110.12 (Mar. 2013), pp. 4646–4650.
[30] Nils Petter Gleditsch et al. “Armed Conflict 1946-2001: A New Dataset”. en. In: *Journal of Peace Research* 39.5 (Sept. 2002), pp. 615–637. ISSN: 0022-3433, 1460-3578. DOI: 10.1177/0022343302039005007.
[31] Tilmann Gneiting. “Strictly and Non-Strictly Positive Definite Functions on Spheres”. In: *Bernoulli* 19.4 (Sept. 2013), pp. 1327–1349.
[32] Roger Guimerà and Marta Sales-Pardo. “Justice Blocks and Predictability of U.S. Supreme Court Votes”. en. In: *PLoS ONE* 6.11 (Nov. 2011). Ed. by Yamir Moreno, e27188. ISSN: 1932-6203. DOI: 10.1371/journal.pone.0027188.
[33] Gavin Hall and William Bialek. “The Statistical Mechanics of Twitter”. en. In: *arXiv:1812.07029 [physics]* (Dec. 2018). arXiv: 1812.07029 [physics].
[34] R. Hanel, S. Thurner, and M. Gell-Mann. “How Multiplicity Determines Entropy and the Derivation of the Maximum Entropy Principle for Complex Systems”. en. In: *Proceedings of the National Academy of Sciences* 111.19 (May 2014), pp. 6905–6910. ISSN: 0027-8424, 1091-6490. DOI: 10.1073/pnas.1406071111.
[35] Aapo Hyvärinen. “Some Extensions of Score Matching”. en. In: *Computational Statistics & Data Analysis* 51.5 (Feb. 2007), pp. 2499–2512. ISSN: 01679473. DOI: 10.1016/j.csda.2006.09.003.
[36] F Itakura. “Minimum Prediction Residual Principle Applied to Speech Recognition”. In: *IEEE Trans Acoust, Speech, Signal Process* 23.1 (Feb. 1975), pp. 67–72.
[37] E. T. Jaynes. “Information Theory and Statistical Mechanics”. en. In: *Physical Review* 106.4 (May 1957), pp. 620–630. ISSN: 0031-899X. DOI: 10.1103/PhysRev.106.620.
[38] Neil F Johnson. *Personal Communication*. Mar. 2019.
[39] Neil F. Johnson et al. “Simple Mathematical Law Benchmarks Human Confrontations”. en. In: *Scientific Reports* 3.1 (Dec. 2013), p. 3463. ISSN: 2045-2322. DOI: 10.1038/srep03463.
[40] Charles F F Karney. “Algorithms for Geodesics”. In: *J Geod* 87.1 (2013), pp. 43–55.
[41] Daniel Martin Katz, Michael J. Bommarito, and Josh Blackman. “A General Approach for Predicting the Behavior of the Supreme Court of the United States”. en. In: *PLoS ONE* 12.4 (Apr. 2017). Ed. by Luís A. Nunes Amaral, e0174698. ISSN: 1932-6203. DOI: 10.1371/journal.pone.0174698.
[42] A.Y. Khinchin. *Mathematical Foundations of Information Theory*. Dover Books on Mathematics. Dover Publications, 1957.
[43] Joakim Kreutz. “How and When Armed Conflicts End: Introducing the UCDP Conflict Termination Dataset”. In: 47.2 (Mar. 2010), pp. 243–250.

[44] Edward D. Lee. "Partisan Intuition Belies Strong, Institutional Consensus and Wide Zipf's Law for Voting Blocs in US Supreme Court". en. In: *Journal of Statistical Physics* 173.6 (Dec. 2018), pp. 1722–1733. ISSN: 0022-4715, 1572-9613. DOI: 10.1007/s10955-018-2156-0.

[45] Edward D. Lee, Chase P. Broedersz, and William Bialek. "Statistical Mechanics of the US Supreme Court". en. In: *Journal of Statistical Physics* 160.2 (July 2015), pp. 275–301. ISSN: 0022-4715, 1572-9613. DOI: 10.1007/s10955-015-1253-6.

[46] Edward D. Lee and Bryan C. Daniels. "Convenient Interface to Inverse Ising (ConIII): A Python 3 Package for Solving Ising-Type Maximum Entropy Models". en. In: *Journal of Open Research Software* 7.1 (Mar. 2019), p. 3. ISSN: 2049-9647. DOI: 10.5334/jors.217.

[47] Jeffrey Lewis. *California Assembly and Senate Roll Call Votes, 1993 to the Present*. http://amypond.sscnet.ucla.edu/california/. 2019.

[48] Jeffrey B. Lewis et al. *Voteview: Congressional Roll-Call Votes Database*. https://voteview.com/. 2019.

[49] Dwayne Liburd and Sonia Barbosa. *State Supreme Court Data Project*. en. Jan. 2009. DOI: 10.7910/DVN/Z80F7P.

[50] David J C MacKay. *Information Theory, Inference, and Learning Algorithms*. en. Cambridge University Press, 2003.

[51] Andrew D. Martin and Kevin M. Quinn. "Dynamic Ideal Point Estimation via Markov Chain Monte Carlo for the U.S. Supreme Court, 1953–1999". en. In: *Political Analysis* 10.02 (2002), pp. 134–153. ISSN: 1047-1987, 1476-4989. DOI: 10.1093/pan/10.2.134.

[52] Lina Merchan and Ilya Nemenman. "On the Sufficiency of Pairwise Interactions in Maximum Entropy Models of Networks". en. In: *Journal of Statistical Physics* 162.5 (Mar. 2016), pp. 1294–1308. ISSN: 0022-4715, 1572-9613. DOI: 10.1007/s10955-016-1456-5.

[53] "Motion Detector User Manual". In: (2018).

[54] Meinard Müller. *Information Retrieval for Music and Motion*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2007.

[55] Ilya Nemenman, F. Shafee, and William Bialek. "Entropy and Inference, Revisited". In: *Advances in Neural Information Processing Systems 14*. Ed. by T. G. Dietterich, S. Becker, and Z. Ghahramani. MIT Press, 2002, pp. 471–478.

[56] M. E. J. Newman and G. T. Barkema. *Monte Carlo Methods in Statistical Physics*. en. Oxford : New York: Clarendon Press ; Oxford University Press, 1999.

[57] Mark E. J. Newman. "Power Laws, Pareto Distributions and Zipf's Law". en. In: *Contemporary Physics* 46.5 (Sept. 2005), pp. 323–351. ISSN: 0010-7514, 1366-5812. DOI: 10.1080/00107510500052444.

[58] H. Chau Nguyen, Riccardo Zecchina, and Johannes Berg. "Inverse Statistical Problems: From the Inverse Ising Problem to Data Science". en. In:

Advances in Physics 66.3 (July 2017), pp. 197–261. ISSN: 0001-8732, 1460-6976. DOI: 10.1080/00018732.2017.1341604.

[59] Stefanos Papanikolaou et al. "Universality beyond Power Laws and the Average Avalanche Shape". en. In: *Nature Physics* 7.4 (Apr. 2011), pp. 316–320. ISSN: 1745-2473, 1745-2481. DOI: 10.1038/nphys1884.

[60] S. Picoli et al. "Universal Bursty Behaviour in Human Violent Conflicts". en. In: *Scientific Reports* 4.1 (May 2015), p. 4773. ISSN: 2045-2322. DOI: 10.1038/srep04773.

[61] Adrián Ponce-Alvarez et al. "Whole-Brain Neuronal Activity Displays Crackling Noise Dynamics". en. In: *Neuron* 100.6 (Dec. 2018), 1446–1459.e6. ISSN: 08966273. DOI: 10.1016/j.neuron.2018.10.045.

[62] Keith T. Poole and Howard Rosenthal. "A Spatial Model for Legislative Roll Call Analysis". en. In: *American Journal of Political Science* 29.2 (May 1985), p. 357. ISSN: 00925853. DOI: 10.2307/2111172.

[63] Keith Poole et al. "Scaling Roll Call Votes with Wnominate in R". en. In: *Journal of Statistical Software* 42.14 (June 2011), p. 21.

[64] Katherine N. Quinn. "Patterns of Structural Hierarchies in Complex Systems". PhD thesis. Cornell University, 2019.

[65] Clionadh Raleigh et al. "Introducing ACLED: An Armed Conflict Location and Event Dataset: Special Data Feature". en. In: *Journal of Peace Research* 47.5 (Sept. 2010), pp. 651–660. ISSN: 0022-3433, 1460-3578. DOI: 10.1177/0022343310378914.

[66] C E Rasmussen and C K I Williams. *Gaussian Processes for Machine Learning*. Cambridge: MIT Press, 2006.

[67] Chotirat Ann Ratanamahatana and Eamonn Keogh. "Everything You Know about Dynamic Time Warping Is Wrong". In: *KDD/TDM 2004*. Aug. 2004, pp. 1–11.

[68] H Risken. *The Fokker-Planck Equation*. 1984.

[69] Stan Salvador and Philip Chan. "Toward Accurate Dynamic Time Warping in Linear Time and Space". In: *Intell Data Anal* 11.5 (2007), pp. 561–580.

[70] M R Sarkees and F W Wayman. *Resort to War: A Data Guide to Inter-State, Extra-State, Intra-State, and Non-State Wars, 1816-2007*. Correlates of War Series. CQ Press, 2010.

[71] James P. Sethna, Karin A. Dahmen, and Christopher R. Myers. "Crackling Noise". En. In: *Nature* 410.6825 (Mar. 2001), p. 242. ISSN: 1476-4687. DOI: 10.1038/35065675.

[72] John Shawe-Taylor and Nello Cristianini. *Kernel Methods for Pattern Analysis*. Cambridge University Press, June 2004.

[73] Jascha Sohl-Dickstein, Peter B. Battaglino, and Michael R. DeWeese. "New Method for Parameter Estimation in Probabilistic Models: Minimum Probability Flow". en. In: *Physical Review Letters* 107.22 (Nov. 2011), p. 220601. ISSN: 0031-9007, 1079-7114. DOI: 10.1103/PhysRevLett.107.220601.

[74] Jascha Sohl-Dickstein, Peter Battaglino, and Michael R. DeWeese. "Minimum Probability Flow Learning". en. In: *arXiv:0906.4779 [physics, stat]* (June 2009). arXiv: 0906.4779 [physics, stat].

[75] Harold J. Spaeth et al. *Supreme Court Database*. http://Supremecourtdatabase.org. 2016.

[76] Dietrich Stauffer and Ammon Aharony. *Introduction to Percolation Theory*. CRC Press, 1994.

[77] Sattar Taheri-Araghi et al. "Cell-Size Control and Homeostasis in Bacteria". In: *Current Biology* 25.3 (2015), pp. 385–391.

[78] Nicholas M. Timme et al. "Criticality Maximizes Complexity in Neural Tissue". en. In: *Frontiers in Physiology* 7 (Sept. 2016). ISSN: 1664-042X. DOI: 10.3389/fphys.2016.00425.

[79] Gasper Tkacik et al. "Spin Glass Models for a Network of Real Neurons". en. In: *arXiv:0912.5409 [q-bio]* (Dec. 2009). arXiv: 0912.5409 [q-bio].

[80] Mark K. Transtrum and Peng Qiu. "Model Reduction by Manifold Boundaries". en. In: *Physical Review Letters* 113.9 (Aug. 2014), p. 098701. ISSN: 0031-9007, 1079-7114. DOI: 10.1103/PhysRevLett.113.098701.

[81] N G Van Kampen. *Stochastic Processes in Physics and Chemistry*. Elsevier, Aug. 2011.

[82] Vicon. "Personal Communication". July 2016.

[83] G Welch and G Bishop. *An Introduction to the Kalman Filter*. Tech. rep. 2006.

[84] Rashid V. Williams-García, John M. Beggs, and Gerardo Ortiz. "Unveiling Causal Activity of Complex Networks". en. In: *Europhysics Letters* 119.1 (July 2017), p. 18003. ISSN: 0295-5075, 1286-4854. DOI: 10.1209/0295-5075/119/18003.

[85] R. K. P. Zia, Edward F. Redish, and Susan R. McKay. "Making Sense of the Legendre Transform". en. In: *American Journal of Physics* 77.7 (July 2009), pp. 614–622. ISSN: 0002-9505, 1943-2909. DOI: 10.1119/1.3119512. arXiv: 0806.1147.