# Negative Representation and Instability in Democratic Elections

Alexander Siegenfeld[1,2] and Yaneer Bar-Yam[2]

[1]*Department of Physics, Massachusetts Institute of Technology, 77 Massachusetts Ave., Cambridge, MA 02139, USA*
[2]*New England Complex Systems Institute, 277 Broadway, Cambridge, MA 02139, USA*

(Dated: October 29, 2018)

Motivated by the troubling rise of political extremism and instability throughout the democratic world, we present a novel mathematical characterization of the nature of political representation in democratic elections. We define the concepts of *negative representation*, in which a shift in electorate opinions produces a shift in the election outcome in the opposite direction, and *electoral instability*, in which an arbitrarily small change in opinion causes a large change in election outcome. Under very general conditions, we prove that unstable elections necessarily contain negatively represented opinions. Furthermore, increasing polarization of the electorate can drive elections through a transition from a stable to an unstable regime, analogous to the phase transition by which some materials become ferromagnetic below their critical temperatures. In this unstable regime, a large fraction of political opinions are negatively represented. Empirical data suggest that United States presidential elections underwent such a phase transition in the 1970s and have since become increasingly unstable.

It is generally believed that American democracy represents its citizens. However, the nature of this representation is unclear. Did the American electorate change so much between 2012 and 2016 that it is possible for candidates as different as Obama and Trump to have both been representative of that electorate? Indeed, many citizens believe Trump does not represent America [1], and others thought the same of Obama [2]. A clue to this puzzle of representation is the observation that a small change in the electorate can result in a large swing in the election outcome. For instance, in both the 2000 and 2016 presidential elections, a change in the political preferences of less than one percent of the electorate or the use of a national popular vote would have shifted the outcome between two very different candidates. Using a novel mathematical framework for analyzing elections and representation, we demonstrate how the instability described above implies a failure of representation.

Elections are, fundamentally, a means of aggregating many opinions into one—those of the citizens into that of the elected official. The opinions of the electorate and candidates must lie in some set of all potential opinions; for simplicity, we focus on the case in which this set can be embedded in a one-dimensional continuous space ($\mathbb{R}$), as in other studies [3–6], and extend the results to a multidimensional space in section S2 A. We define an *election* by $y[f(x)]$, a *functional* (a function of a function) that maps the distribution of electorate opinions $f(x)$—defined so that for any interval $[a, b] \in \mathbb{R}$, the number of voters with opinions in that interval is $\int_a^b f(x)dx$—to the election outcome $y \in \mathbb{R}$, the opinion of the elected official. Note that in this framework, candidacy is endogenous: candidate opinions—or equivalently, which candidates run—are themselves functions of the electorate opinions. Any electoral system, regardless of its detailed mechanisms (e.g. the number of candidates, the nature of political parties, etc.) can be conceptualized as such a process—a black box that outputs the opinion of the winning candidate based on the electorate opinions. The one restriction we place on this process is *translational invariance*: we assume that the election outcome will shift by $c$ if all opinions are shifted by $c$, or formally

$$y[f(x + c)] + c = y[f(x)] \tag{1}$$

for all $c$. (This assumption is technically stronger than necessary—see section S1 B.) If the model were not translationally invariant, it would *a priori* privilege certain opinions over others.

In order for representation to measure how much election outcomes change in response to changes in voter opinions [7], we define the *representation* of an opinion $x$ by

$$r_c(f, x) = \frac{\delta y}{c} \tag{2}$$

where $\delta y$ is the change in outcome that would occur if $x$ changes to $x' = x + c$. (Representation should *not* be defined using only the distance between a voter's opinion and that of the elected candidate: opinions without causal influence on the election outcome are not represented, even if they happen to align with that outcome.) It is convenient to measure representation by

$$r(f, x) = \lim_{c\to 0} r_c(f, x) \tag{3}$$

when the limit exists, since $r(f, x)$ does not depend on $c$.

For a large population, a number of results hold (see section S1 C for details). First, we can express representation in terms of functional derivatives of the election, assuming they exist:

$$r(f, x) = \frac{d}{dx}\frac{\delta y}{\delta f(x)} \tag{4}$$

(Note: the functional derivative $\frac{\delta y}{\delta f(x)}$ measures the change in the outcome $y$ with respect to a small increase in the distribution of opinions at $x$—see section S1 A.) Second, $r_c(f, x)$ is the average of $r(f, x)$ over the interval $[x, x+c]$, and thus representation of individual opinions can be measured by $r(f, x)$ alone. Third, the total representation of the electorate's opinions equals 1:

$$\int_{-\infty}^{\infty} f(x)r(f, x)dx = 1 \tag{5}$$

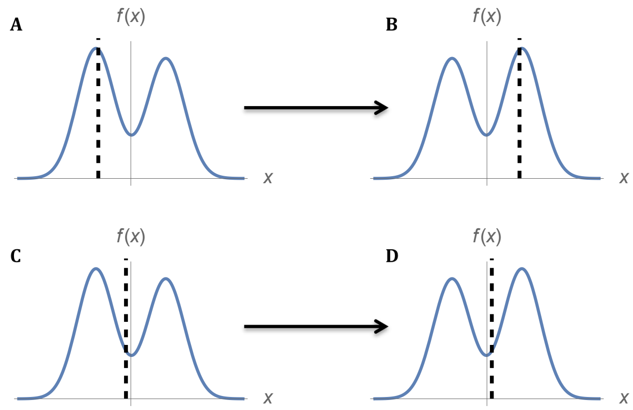


FIG. 1. Changes in election outcomes (vertical dashed lines) resulting from changes in distributions of voter opinions ($f(x)$, denoted by the blue curves), where the horizontal axes ($x$) denote political opinion. (A to B) When not everyone votes, an election can be unstable (eq. (12))—a small shift in voters to the right causes a large swing in the election outcome. (C to D) For median voting, by contrast, a small shift in voters causes a similarly small shift in outcome.

Because the total amount of representation is fixed, individual voter opinions are on average more represented in smaller democracies than in larger ones. Although (or perhaps because) this fact is intuitive, its implications are largely ignored in public discourse.

Before discussing our main results concerning negative representation and instability, we apply our formalism to two examples. The Median Voter Theorem [4] states that when political opinions lie in one dimension and everyone votes for the candidate whose opinion is closest to his, the optimal position for a candidate to take is that of the median voter. For this type of election, $y[f] = \text{median}[f] \equiv m$, yielding $r(f,x) = \frac{d}{dx}\frac{\delta y}{\delta f(x)} = \frac{d}{dx}\frac{1}{2f(m)}\,\text{sign}(x-m) = \frac{\delta(x-m)}{f(m)}$. Only the median voter's opinion is represented—other opinions matter in that they establish who the median voter is. This concentration of representation is consistent with the Owen-Shapley index [8] of voting power (section S2 B). By contrast, for an election in which the mean opinion is selected ($y[f] = \frac{\int_{-\infty}^{\infty} f(x)x dx}{\int_{-\infty}^{\infty} f(x)dx}$), all opinions are equally represented ($r(f,x) = \frac{1}{N}$ where $N$ is the number of voters). Despite this equal representation, mean voting has its problems: the presence of an extreme opinion can have an outsized effect, and voters therefore have an incentive to exaggerate their opinions. Median voting, on the other hand, is robust: the presence of an additional opinion, regardless of its extremity, has an equal effect of $\frac{\delta y}{\delta f(x)} = \pm\frac{1}{2f(m)}$. For actual elections, in contrast with the assumptions of the Median Voter Theorem, not everyone votes. As we later show, when people with opinions farther from both candidates are more likely to vote, elections can lean towards mean voting; when those people are less likely to vote, negative representation and instability can arise.

Instability and negative representation are comorbid. An election is unstable if an arbitrarily small change in opinion can cause a sizable change in the election outcome (fig. 1), i.e. for some $f$ and $x$,

$$\lim_{c\to 0} cr_c(f,x) \neq 0 \tag{6}$$

If an election $y[f]$ is unstable for an opinion distribution $f_0$, then if some opinion $x_0$ changes by a small amount $\epsilon$ (call the resulting opinion distribution $f_1$), the election outcome changes by a larger amount $C$, i.e. $\delta y_1 \equiv y[f_1] - y[f_0] = C$ with $|C| > |\epsilon|$. Now consider starting with $f_0$ and shifting all opinions except $x_0$ by $-\epsilon$ (call the resulting opinion distribution $f_2$). Since $f_2(x) = f_1(x+\epsilon)$, $\delta y_2 \equiv y[f_2]-y[f_0] = C-\epsilon$ by translational invariance (eq. (1), see section S1 B). Thus, $\delta y_1$ and $\delta y_2$ have the same sign (since $|C| > |\epsilon|$), despite being caused by changes of opinion in opposite directions, so one of the two changes in opinions must be negatively represented. Note that eq. (5), which holds only when the population is large and the election differentiable, was not used, as this result involves a lack of continuity. (Specifically, the effects of individual changes in opinion on the election outcome need not be additive, while eq. (5) assumes that these effects can be expanded to linear order.) This proof that instability implies negative representation relies only on translational invariance; therefore, we expect it to hold generally in real-world conditions, regardless of the size of the electorate, the number of candidates, the existence of primaries, the effects of the electoral college, etc.

To illustrate how negative representation can arise and to show how polarization drives elections through a phase transition into an unstable regime, we now examine probabilistic voting models [9–12]. Consider an election with two candidates who choose their positions in order to maximize their chances of winning the election (or, equivalently, choose to run/are selected to run in the general election based on these considerations; see section S1 D). Denoting the probability that a vote from someone with opinion $x$ will go to the first candidate minus the probability that it will go to the second by $p_x(y_1, y_2)$ (where $y_1$ and $y_2$ are the opinions of the first and second candidates, respectively), we assume there exists some function $u_x$ such that

$$p_x(y_1,y_2) = u_x(y_1) - u_x(y_2) \tag{7}$$

This assumption yields an affine linear *utility difference model* [11]; possible election outcomes are given by (Theorem 4 of [11]):

$$\underset{y}{\text{argmax}} \int_{-\infty}^{\infty} u_x(y)f(x)dx \tag{8}$$

Essentially, this model assumes that the position that maximizes a candidate's margin of victory does not depend on the position of the other candidate. Although this model may not be an accurate description of individual behavior and although it ignores factors such as primaries, it can nonetheless describe the properties of real-world elections, since representation and instability depend only on the election outcome as a function of electorate opinions and not on the mechanism through which the outcome arises. For instance, this model cannot describe the deterministic voting underlying the median voter theorem but can nonetheless exactly capture the collective behavior that such voting gives rise to, namely the election of the median opinion for $u_x(y) = -|y-x|$. (Mean voting is captured by $u_x(y) = -(y-x)^2$ and a system between median and mean voting is captured by $u_x(y) = -\sqrt{(y-x)^2+b^2}$ for varying values of $b$—see section S1 E.)

Voters with opinions that are very far from both candidates are more likely to abstain from voting, a phenomenon known as *alienation* [13–15]. Since not voting for either candidate is equivalent to preferring them equally, a utility function $u_x(y)$ that captures alienation will be almost flat for large $|y-x|$. One such utility function is

$$u_x(y) = u(y-x) = e^{-\frac{(y-x)^2}{2a^2}} \tag{9}$$

where $a$ is a positive constant. Representation is then given by (section S1 F)

$$r(f,x) \propto -\frac{1}{N}u''(y^*-x) = \frac{1}{Na^4}(a^2-(y^*-x)^2)e^{-\frac{(y^*-x)^2}{2a^2}} \tag{10}$$

where $N$ is the number of constituents. Opinions far from the election outcome ($|x-y^*| > a$) are negatively represented ($r(f,x) < 0$, see fig. 2): the election outcome is inversely sensitive to changes in those opinions. For example, given a center-left candidate and a candidate to the right, as left-wing individuals move farther left, they become less likely to vote for the center-left candidate (often choosing instead not to vote), which increases the probability that the candidate on the right will win. In response, the electoral equilibrium of future elections may shift rightward, as candidates no longer vie for these left-wing votes, or instability may arise, as future left-leaning candidates chase these votes, widening the distance between themselves and the right-leaning candidates. Mathematically, negative representation is avoided only when $u(y-x)$ is concave, an assumption that is quite limiting but nonetheless often used in the social choice literature to simplify the mathematics (if $u$ is concave, then so is $\int_{-\infty}^{\infty} u(y-x)f(x)dx$, which guarantees a single local maximum $y^*$ that is also the global maximum). From an individual's perspective, negative representation is a natural outcome: of course the election outcome will move away from those who decide not to vote. But on a system level, negative representation indicates a perversion in the election's aggregation of constituents' opinions.

When combined with a polarized electorate, negative representation leads to instability (we proved the converse under more general conditions; see section S1 G for further discussion and mathematical details). For $u_x$ defined by eq. (9) and an opinion distribution $f(x)$ consisting of two (potentially unequally weighted) normal distributions centered at $\pm\Delta$,

$$f(x) = w_1 e^{-\frac{(x+\Delta)^2}{2\sigma^2}} + w_2 e^{-\frac{(x-\Delta)^2}{2\sigma^2}}, \tag{11}$$

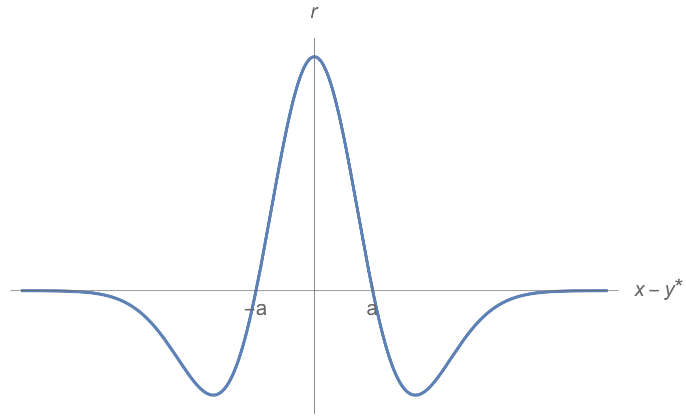


FIG. 2. When some voters abstain due to alienation, opinions far from the election outcome may be negatively represented. This graph depicts the representation ($r$) of opinions ($x$) as a function of their distance from the election outcome ($y^*$) for voting behavior given by eq. (9).

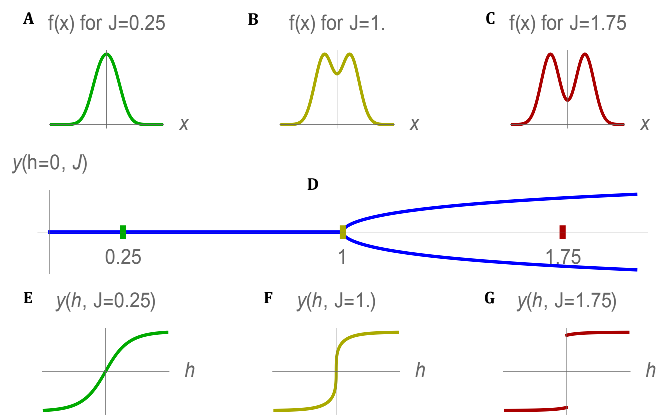


FIG. 3. (A through C) With increasing polarization ($J$) of the voter opinion distribution, (D) the electoral system undergoes a phase transition from possessing a single stable outcome to possessing two possible unstable outcomes. (E) In the stable regime, the outcome smoothly responds to changes in the relative sizes of the two subpopulations (changes in $h$). (F) At the phase transition ($J = 1$), the outcome is continuous but not differentiable in the relative sizes of the two subpopulations. (G) In the unstable regime, the outcome discontinuously jumps. These figures were created using eqs. (11) and (12) for $a = \sigma = 1$ (for which $J = \Delta^2/2$).

the outcome $y$ is given by the following condition:

$$y/\Delta = \tanh(Jy/\Delta + h) \tag{12}$$

where $J = \Delta^2/(a^2+\sigma^2)$ and $h = \frac{1}{2}\ln\frac{w_2}{w_1}$. For $w_1 = w_2$ (i.e. for equally sized subpopulations), $h = 0$, and the election is stable for $J \le 1$ and unstable for $J > 1$. In the stable regime, $y = 0$. In the unstable regime, there are two possible outcomes described by $\pm|y^*|$, and an arbitrarily small change in

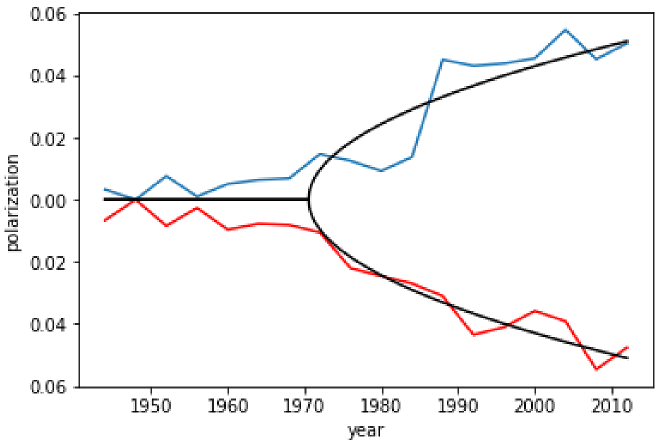


FIG. 4. The polarization of the Democratic (blue) and Republican (red) parties between 1944 and 2012, as measured by the fraction of polarizing words in the party platforms relative to a baseline. Party platforms are released once every four years. During the 1970s, there appears to be a divergence, which may correspond to a phase transition of the electoral dynamics into instability. The prediction of our model (black, cf. fig. 3D) is superimposed; $R^2$ values are 0.86 for the Democratic party and 0.89 for the Republican party. See section S1 I for further details.

$h$ can cause $y$ to swing between its positive and negative values, leaving one or the other subpopulations with most of its opinions negatively represented. If one subpopulation is significantly larger, the smaller subpopulation will consistently lose, but if there is a close balance between the two subpopulations, then small variations determine which subpopulation wins. The variations might include changes in population, nuances in the candidates' personalities, changes in the rules (simple majority versus electoral college system, for example), voting restrictions, and the effectiveness of turnout operations. From election cycle to election cycle, the outcome can swing between the two subpopulations, with the majority of the opinions in the losing subpopulation being negatively represented (section S1 G). This instability that arises in a polarized electorate was qualitatively predicted in 1957 by Downs [5], who claimed that "democracy does not lead to effective, stable government when the electorate is polarized."

This voting model (eq. (12)) is precisely equivalent to a mean-field Ising model of a ferromagnet [16], in which each spin (magnetic dipole) interacts with every other spin, highlighting the hidden dependencies that can arise from a system in which voters' behaviors are superficially independent (section S1 H). The seemingly independent voters are coupled through collectively choosing a candidate. Because of these effective interactions between voters, the system can exhibit emergent behavior in which there can be discontinuities in the election outcome despite the continuous behavior of the voters. Such emergent discontinuities are a common feature of complex systems [17–20].

Empirically, the opinions of the United States population, and in particular the opinions of those most likely to vote, have been polarizing over the past few decades [21]. Thus, we might expect that over time, election outcomes have undergone a phase transition, and indeed this appears to be the case (fig. 4). Changes to the electoral process, such as reforms in the early 1970s as to how presidential party nominees are chosen, together with increasing polarization, have likely driven this phase transition into instability (and therefore negative representation). That the policies of legislators in politically homogeneous districts are more strongly correlated with the preferences of the district's median voter than the policies of legislators in heterogeneous districts [22] lends further empirical support.

Our results imply that electoral outcomes can be polarized and unstable, a reflection of negative representation. Therefore, the impact of electoral reforms on this instability should be considered. For instance, alienation, which occurs when not everyone votes, can cause this departure from median voting. Low turnout, therefore, results not only in fewer voices being heard, but also in negative representation and instability, which in turn can breed anger, resentment, political dysfunction, and a loss of faith in democracy itself.


[1] Christopher Mele and Annie Correal. "'Not our president': Protests spread after Donald Trump's election," *New York Times*; nytimes.com/2016/11/10/us/trump-election-protests.html.

[2] Matt Motyl. "If he wins, I'm moving to Canada": Ideological migration threats following the 2012 us presidential election. *Analyses of Social Issues and Public Policy*, 14(1):123–136, 2014.

[3] Harold Hotelling. Stability in competition. *The Economic Journal*, 39(153):41–57, 1929.

[4] Duncan Black. On the rationale of group decision-making. *Journal of Political Economy*, 56(1):23–34, 1948.

[5] Anthony Downs. An economic theory of political action in a democracy. *Journal of Political Economy*, 65(2):135–150, 1957.

[6] George Rabinowitz and Stuart E Macdonald. A directional theory of issue voting. *American Political Science Review*, 83(1):93–121, 1989.

[7] Stefan Napel and Mika Widgrén. Power measurement as sensitivity analysis: a unified approach. *Journal of Theoretical Politics*, 16(4):517–538, 2004.

[8] Guillermo Owen and Lloyd S Shapley. Optimal location of candidates in ideological space. *International Journal of Game Theory*, 18(3):339–356, 1989.

[9] Peter Coughlin and Shmuel Nitzan. Electoral outcomes with probabilistic voting and Nash social welfare maxima. *Journal of Public Economics*, 15(1):113–121, 1981.

[10] James M Enelow and Melvin J Hinich. A general probabilistic spatial theory of elections. *Public Choice*, 61(2):101–113, 1989.

[11] Jeffrey S Banks and John Duggan. Probabilistic voting in the spatial model of elections: The theory of office-motivated candidates. In D. Austen-Smith and J. Duggan, editors, *Social Choice and Strategic Decisions*, pages 15–56. Springer, 2005.

[12] Peter J Coughlin. *Probabilistic Voting Theory*. Cambridge University Press, 1992.
[13] Melvin J Hinich. Some evidence on non-voting models in the spatial theory of electoral competition. *Public Choice*, 33(2):83–102, 1978.
[14] Priscilla L Southwell. The politics of alienation: Nonvoting and support for third-party candidates among 18–30-year-olds. *The Social Science Journal*, 40(1):99–107, 2003.
[15] James Adams, Jay Dow, and Samuel Merrill. The political consequences of alienation-based and indifference-based voter abstention: Applications to presidential elections. *Political Behavior*, 28(1):65–86, 2006.
[16] Leo P Kadanoff. More is the same; phase transitions and mean field theories. *Journal of Statistical Physics*, 137(5-6):777, 2009.
[17] Robert M May, Simon A Levin, and George Sugihara. Complex systems: Ecology for bankers. *Nature*, 451(7181):893, 2008.
[18] Marten Scheffer, Jordi Bascompte, William A Brock, Victor Brovkin, Stephen R Carpenter, Vasilis Dakos, Hermann Held, Egbert H Van Nes, Max Rietkerk, and George Sugihara. Early-warning signals for critical transitions. *Nature*, 461(7260):53, 2009.
[19] Thierry Mora and William Bialek. Are biological systems poised at criticality? *Journal of Statistical Physics*, 144(2):268–302, 2011.
[20] Jean-Philippe Bouchaud. Crises and collective socio-economic phenomena: simple models and challenges. *Journal of Statistical Physics*, 151(3-4):567–606, 2013.
[21] "Political polarization in the american public," *Pew Research Center*, Washington, D.C.; people-press.org/2014/06/12/political-polarization-in-the-american-public/.
[22] Elisabeth R Gerber and Jeffrey B Lewis. Beyond the median: Voter preferences, district heterogeneity, and political representation. *Journal of Political Economy*, 112(6):1364–1383, 2004.
[23] Soren Jordan, Clayton M Webb, and B. Dan Wood. The president, polarization and the party platforms. *The Forum*, 12(1):169–189, 2014.
[24] Keith T Poole and Howard Rosenthal. A spatial model for legislative roll call analysis. *American Journal of Political Science*, 29(2):357–384, 1985.
[25] Mehran Kardar. *Statistical Physics of Fields*. Cambridge University Press, 2007.
[26] James M Enelow and Melvin J Hinich. *The Spatial Theory of Voting: An Introduction*. Cambridge University Press, 1984.
[27] Dan S Felsenthal and Moshé Machover. A priori voting power: what is it all about? *Political Studies Review*, 2(1):1–23, 2004.
[28] Lionel S Penrose. The elementary statistics of majority voting. *Journal of the Royal Statistical Society*, 109(1):53–57, 1946.
[29] Lloyd S Shapley and Martin Shubik. A method for evaluating the distribution of power in a committee system. *The American Political Science Review*, 48(3):787–792, 1954.
[30] Andrew Gelman, Jonathan N Katz, and Francis Tuerlinckx. The mathematics and statistics of voting power. *Statistical Science*, 17(4):420–435, 2002.
[31] Stefano Benati and Giuseppe V Marzetti. Probabilistic spatial power indexes. *Social Choice and Welfare*, 40(2):391–410, 2013.

### Acknowledgments

This material is based upon work supported by the National Science Foundation Graduate Research Fellowship under Grant No. 1122374 and the Hertz Foundation. We would also like to thank B. Dan Wood for sharing the data from his paper on the polarization of party platforms [23].

## S1. METHODS

### A. Distributions and functional derivatives

In this section, we give a brief introduction to the mathematics behind Dirac delta functions and functional derivatives. The Dirac delta function $\delta(x)$ is not technically a function, but is rather a *generalized function*, also known as a *distribution*. A distribution may not be well-defined if evaluated at a particular point (e.g. $\delta(x)$ is not defined for $x = 0$), but is instead defined through the integral of its product with ordinary functions.[1] For instance, the Dirac delta function is defined by eq. (13):

$$\int_{-\infty}^{\infty} \delta(x) g(x) dx \equiv g(0) \tag{13}$$

for all continuous functions $g : \mathbb{R} \to \mathbb{R}$. We note that $\delta(x)$ can be approximated by an arbitrarily narrow gaussian, which has a total area under its curve of 1. (The more narrow the gaussian, the taller it must be, so that the product of its height and width remains constant.) Formally, $\delta(x) \sim \frac{1}{\sqrt{\pi\epsilon}} e^{-x^2/\epsilon}$ for small $\epsilon$, in the sense that

$$\lim_{\epsilon \to 0} \int_{-\infty}^{\infty} \frac{1}{\sqrt{\pi\epsilon}} e^{-x^2/\epsilon} g(x) dx = g(0) \tag{14}$$

for continuous $g$.

For a set of $N$ voters with opinions $\{x_1, x_2, ..., x_N\}$, the distribution of voter opinions is $f(x) = \sum_{i=1}^{N} \delta(x - x_i)$, which is the only distribution satisfying the property that $\int_a^b f(x) dx$ is the number of voters with opinions in the interval $[a, b]$. However, it is often useful to choose $f(x)$ to be a smooth function that approximates $\sum_{i=1}^{N} \delta(x - x_i)$, in the sense that the difference between $\int_a^b f(x) dx$ and the number of voters with opinions in the interval $[a, b]$ is no greater than 1 for all $a, b$. One way to achieve this smoothing is to replace the delta functions by their gaussian approximations (see fig. S1). For large enough $N$, the error of up to 1 voter opinion will generally not be significant. Whether or not $f(x)$ is chosen to be smooth does not matter for the results of the text, although for the results that rely on the assumption that the number of voters is large, the mathematics are simpler if $f(x)$ is assumed to be a function rather than a distribution. For instance, the expression for representation in the case of median voting involves evaluating $f$ at its median, an operation which is not well-defined if $f$ is a sum of Dirac delta functions.

[1] Technically, a distribution is a map from the set of smooth functions with compact support to $\mathbb{R}$.

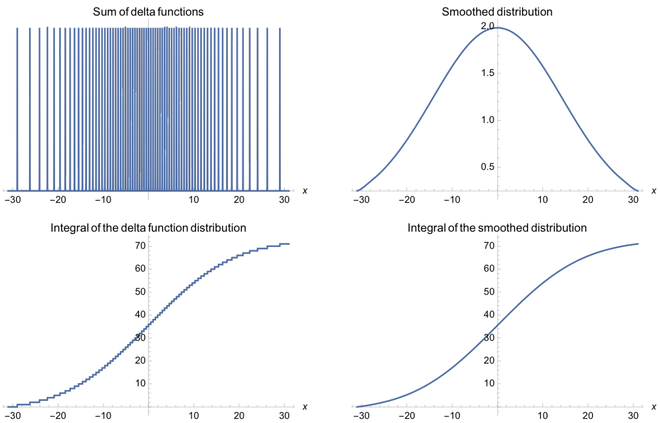


FIG. S1. By replacing the Dirac delta functions with the approximation $\delta(x) \sim \frac{1}{\sqrt{2\pi}} e^{-x^2/2}$, a smooth distribution is obtained from the sum of delta functions. Note that the integrals of these distributions (taken from the lower bound of the domain of the graphs to $x$) are very similar, despite the striking differences between the distributions themselves.

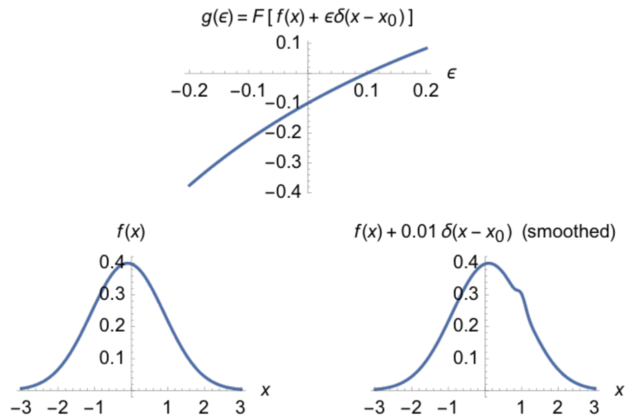


FIG. S2. These graphs depict the output of the functional $F$ as its input distribution changes due to a small increase in opinion at $x_0 = 1$. Here, $F[\cdot] = \text{mean}[\cdot]$ and $f(x)$ is a normal distribution centered at $-0.1$. The delta function is smoothed using $\delta(x) \sim \frac{3}{\sqrt{\pi}} e^{-9x^2}$

A functional $F$ is a map from a space of functions or distributions to $\mathbb{R}$. In analogy to an ordinary derivative, we can define the functional derivative $\frac{\delta F}{\delta f(x)}$ as a function of $x$ that satisfies the following equation for all $\delta f$:

$$\lim_{\epsilon \to 0} \frac{F[f + \epsilon \delta f] - F[f]}{\epsilon} = \int_{-\infty}^{\infty} \frac{\delta F}{\delta f(x)} \delta f(x) dx \qquad (15)$$

Just as the ordinary derivative of a function $g(x)$ will in general depend on $x$, the functional $\frac{\delta F}{\delta f(x)}$ will in general depend on $f$. The expression $\frac{\delta F}{\delta f(x)}$ refers to the value obtained when the functional derivative of $F$ is evaluated at $x$, just as $\frac{dg(x)}{dx}$ refers to the value obtained when the derivative of $g$ is evaluated at $x$. By substituting $\delta f(x) = \delta(x - x_0)$ into eq. (15), a simpler formulation for the functional derivative can be obtained:

$$\frac{\delta F}{\delta f(x_0)} \equiv \lim_{\epsilon \to 0} \frac{F[f(x) + \epsilon \delta(x - x_0)] - F[f(x)]}{\epsilon} \qquad (16)$$

Thus, we can express a functional derivative in terms of ordinary derivatives: if, for a particular $f(x)$ and $x_0$, we define $g(\epsilon) \equiv F[f(x) + \epsilon \delta(x - x_0)]$, then $\frac{\delta F}{\delta f(x_0)} = g'(0)$. This function $g(\epsilon)$ is shown in fig. S2.

### B. Translational invariance

In the main text, we make the assumption of translational invariance (eq. (1)). Technically, translational invariance is defined only in relation to a particular metric. Thus, the assumption of translational invariance can be relaxed without invalidating the paper's results. For the proof that negative representation implies instability, all that is required is that the election be continuous (rather than invariant) under translations or, mathematically, that for $f_c(x) = f(x + c)$, $y[f_c]$ be continuous in $c$ (note that this property is independent of the metric). The proof given in the manuscript then follows in the limit $\epsilon \to 0$. For the proof that the total representation sums to 1 (eq. (5)), there need only exist some metric on the opinion space under which the election is translationally invariant. The total representation will then equal 1 assuming the election is translationally invariant under the metric used to define representation.

### C. Representation in the large-population limit

In this section, we derive properties of our representation measure when the number of voters $N$ is large. In the limit of a large population ($N >> 1$), the change $\delta f$ in the opinion distribution arising from an individual opinion will be small compared to the opinion distribution as a whole, and so we expand $\delta y$ to first order in $\delta f$:

$$\delta y = \int_{-\infty}^{\infty} \delta f(z) \frac{\delta y}{\delta f(z)} dz \qquad (17)$$

Note that eq. (17) does not apply to cases in which $y[f]$ is not differentiable, e.g. when the election is unstable and small changes in the opinion distribution can have an outsized impact; thus we do not use results derived from these equations when analyzing instability.

We now derive eq. (4). Note that when an individual opinion changes from $x$ to $x' = x + c$, the opinion distribution changes by

$$\delta f(z) = \delta(z - x - c) - \delta(z - x) \qquad (18)$$

where $\delta(z)$ is the Dirac delta function. Representation (eq. (2)) is then obtained in terms of functional derivatives of the election by substituting eq. (18) into eq. (17):

$$r_c(f,x) = \frac{1}{c}\delta y = \frac{1}{c}(\frac{\delta y}{\delta f(x+c)} - \frac{\delta y}{\delta f(x)}) \quad (19)$$

which, combined with eq. (3), yields eq. (4) (reproduced below).

$$r(f,x) = \frac{d}{dx}\frac{\delta y}{\delta f(x)} \quad (20)$$

By the fundamental theorem of calculus, we see from eq. (19) that $r_c(f,x)$ is the average of $r(f,x)$ over $[x, x+c]$.

We now prove eq. (5) ($\int_{-\infty}^{\infty} f(x)r(f,x)dx = 1$). For small $\epsilon$, $f(x+\epsilon) = f(x) + \epsilon\frac{d}{dx}f(x)$, and thus $y[f(x+\epsilon)] = y[f(x)+\epsilon\frac{d}{dx}f(x)] = y[f(x)] + \epsilon\int_{-\infty}^{\infty}\frac{df(x)}{dx}\frac{\delta y}{\delta f(x)}dx$. Since $f$ has compact support, which follows from $f$ being an approximation of the opinions of a finite number of voters, integrating by parts yields $y[f(x+\epsilon)] = y[f(x)] - \epsilon\int_{-\infty}^{\infty} f(x)r(f,x)dx$, which, combined with eq. (1), yields eq. (5). This proof assumes that $f$ is differentiable for illustrative purposes, but the result will generally hold whenever $\int_{-\infty}^{\infty} f(x)r(f,x)dx$ is well-defined, if $f$ is treated as a distribution (generalized function).

### D. Nash equilibria of the electoral game

Consider a two-candidate election with endogenous candidacy: candidate positions (or, equivalently, candidates) are chosen in order to maximize the probability of victory. In this framework, the winner of the election will have adopted an unbeatable position $y^*$, provided such a position exists (i.e. a candidate with position $y^*$ will have at least a 50% chance of winning against a candidate with any other position). Formally, $(y^*, y^*)$ is a *Nash equilibrium*, since no candidate can improve her chances by changing her position. Because the voting game is symmetric, if $(y_1, y_2)$ is a Nash equilibrium, then so are $(y_1, y_1)$ and $(y_2, y_2)$; see, for instance, section 2.3 of [9]. Thus, if there is a unique Nash equilibrium, it must be of the form $(y^*, y^*)$.

### E. Utility difference model examples

Here, we give examples of the utility difference model given by eq. (8) describing median voting, mean voting, and an election between median and mean voting. For $u_x(y) = -a^2(y-x)^2$ where $a$ is a constant, the mean opinion is selected, since for a random variable $X$, $\mathbb{E}[(X-\mu)^2]$ is minimized for $\mu = \mathbb{E}[X]$. (Note that $y$ and the support of $f$ must be confined to an interval of length at most $a^{-1}$ so that no probabilities are greater than 1.) For $u_x(y) = -a|y-x|$ (where again, $y$ and the support of $f$ must be confined to an interval of length at most $a^{-1}$), the median opinion is selected (since the median minimizes $\mathbb{E}[|X-m|]$), although by a different mechanism than the deterministic voting assumptions of the Median Voter Theorem. Both of these functions can be viewed as limiting cases of the hyperbolic $u_x(y) = -\sqrt{a^2(y-x)^2+b^2}$, with $b << a$ approximating median voting and $b >> a$ approximating mean voting. Under mean voting, for a voter either to the right of both candidates or to the left of both candidates, the farther away this voter is, the stronger the voter's preference between the two candidates. Under median voting, the strength of this voter's preference for one candidate over the other is independent of how far the voter is from both candidates. For the intermediate case, the strength of this voter's preference gets stronger up to a point and then levels off as the voter moves farther away from both candidates. However, in actual elections, voters with opinions that are very far from both candidates are more likely to abstain from voting (or vote for a third-party candidate), which is why eq. (9) may be more realistic.

### F. Representation in the utility difference model

In this section, we calculate representation for the utility difference model given by eq. (8). We derive the first part of eq. (10), and we then calculate representation for the examples given in section S1 E. To do so, we must assume there is a single possible election outcome $y^*$ (see section S1 D). Then, from eq. (8),

$$y^* = \underset{\tilde{y}}{\operatorname{argmax}} \int_{-\infty}^{\infty} u_x(\tilde{y})f(x)dx \quad (21)$$

which implies

$$0 = \int_{-\infty}^{\infty} u'_x(y^*)f(x)dx \quad (22)$$

Note that eq. (21) satisfies $y[f(x)] = y[\lambda f(x)]$ for any positive constant $\lambda$ (scale invariance), and let $\tilde{f}(x) = f(x)/N$ where $N$ is the size of the electorate, so that $\int_{-\infty}^{\infty}\tilde{f}(x)dx = 1$. Considering the change that arises from the addition of a single individual with opinion $x_0$ to the population, we define $\delta y^*$ by

$$y^* + \delta y^* = y[f(x)+\delta(x-x_0)] = y[\tilde{f}(x)+\epsilon\delta(x-x_0)] \quad (23)$$

where $\epsilon = 1/N$. Then, from eqs. (22) and (23),

$$0 = \int_{-\infty}^{\infty} u'_x(y^*+\delta y^*)(\tilde{f}(x)+\epsilon\delta(x-x_0))dx \quad (24)$$

Because $\int_{-\infty}^{\infty} u_x(x)f(x)dx$ is differentiable in $f$ and has a single maximum in $y$,

$$\delta y^* = \epsilon\frac{-u'_{x_0}(y^*)}{\int_{-\infty}^{\infty} u''_x(y^*)\tilde{f}(x)dx} + O(\epsilon^2) \quad (25)$$

Noting that the denominator is independent of $x_0$, of order 1 (i.e. independent of $N$), and negative (otherwise, $y^*$ would be a minimum rather than a maximum),

$$\frac{\delta y}{\delta \tilde{f}(x_0)} = \lim_{\epsilon \to 0} \frac{\delta y^*}{\epsilon} \propto u'_x(y^*) \tag{26}$$

So, using eq. (4),

$$r(f,x) = \frac{d}{dx}\frac{\delta y}{\delta f(x)} = \frac{1}{N}\frac{\delta y}{\delta \tilde{f}(x)} \propto \frac{1}{N}\frac{d}{dx}u'_x(y^*) \tag{27}$$

If $u_x(y) = u(y-x)$, as it must be for some function $u$ if the election is translationally invariant as in eq. (1), then

$$r(f,x) \propto -\frac{1}{N}u''(y^* - x) \tag{28}$$

Eq. 28 provides a direct link between voter preferences and the representation of opinions. (If needed, the constant of proportionality can be determined through eq. (5).) For $u(y-x) = -(y-x)^2$ and $u(y-x) = -|y-x|$, we can quickly derive the results for median and mean voting, respectively. For $u(y-x) = b - \sqrt{(y-x)^2 + b^2}$, which yields an outcome between that of median and mean voting,

$$r(f,x) \propto ((\frac{x - y[f]}{b})^2 + 1)^{-3/2} \tag{29}$$

resulting in the representation of opinions being concentrated around the election outcome, but not infinitely concentrated as it is for median voting. For $u$ that are not concave, there will exist some $x$ such that $u''(x-y) > 0$, and representation will be negative for those opinions (see eq. (28)).

### G. Instability in the utility difference model

Here, we explore the conditions under which instability can arise in the model described by eq. (8), and we elaborate on the concrete model of instability with outcomes given by eq. (12). For the model given by eq. (8), $r(f,x) = \frac{d}{dx}\frac{\delta y}{\delta f(x)}$ is shown to be well-defined as long as $y[f]$ is single-valued, i.e. eq. (8) has a single maximum (section S1 F). Thus, instability can occur only when there are multiple maxima.[2] For an opinion distribution in which eq. (8) has two maxima $y_1^*$ and $y_2^*$, if each position is taken by one candidate, then each candidate has a 50% chance of winning the election. However, an arbitrarily small change in the opinion distribution can favor one outcome over the other, giving either $y_1^*$ or $y_2^*$ a chance of winning that is arbitrarily close to 100% in the large-population limit.[3]

The existence of multiple maxima in $y$ implies that $\int_{-\infty}^{\infty} u(x-y)f(x)dx$ is not concave (ignoring the degenerate case in which $\int_{-\infty}^{\infty} u(x-y)f(x)dx$ is constant over some interval). Thus, instability can arise only in the case of non-concave $u$, which is precisely the same condition under which negative representation occurs. That instability can arise only in the presence of negative representation should not surprise us, since it was proven under more general conditions in the main text. For this class of models, we also find that negative representation implies that there exist distributions of opinions for which instability arises, i.e. if $u$ is not concave, then there exists an $f(x)$ such that eq. (8) has multiple maxima. To see why this is true, consider an opinion distribution $f$ such that $f(x) = f(-x)$. Then, if there is a single maximum of eq. (8), it must lie at $y^* = 0$. In order for $y^* = 0$ to be a maximum, we must have $\int_{-\infty}^{\infty} f(x)u''(x)dx = 2\int_0^{\infty} f(x)u''(x)dx \leq 0$ (where the equality follows from the symmetry of $u$ and $f$). But since $u$ is twice continuously differentiable and not concave, there exists an $f$ such that $f(x) = f(-x)$ and $\int_0^{\infty} f(x)u''(x)dx > 0$. For such an $f$, $y^* = 0$ is not a maximum, thus contradicting our assumption that there was a unique maximum.

To provide an example of how, for non-concave $u$, the election is unstable for certain opinion distributions, we consider the $u$ used for the example of negative representation: $u(y-x) = \exp\left[-\frac{(y-x)^2}{2a^2}\right]$ for some positive constant $a$ (eq. (9)). We take the distribution of voter opinions to be a sum of two (potentially unequally weighted) normal distributions of equal variance:

$$f(x) = \sum_{\alpha=1,2} w_\alpha e^{-\frac{(x-\mu_\alpha)^2}{2\sigma^2}} \tag{30}$$

Then, from eq. (8), the outcome of the election is then given by

$$\underset{y}{\operatorname{argmax}} \int_{-\infty}^{\infty} f(x) e^{-\frac{(y-x)^2}{2a^2}} dx = \underset{y}{\operatorname{argmax}} \sum_{\alpha=1,2} w_\alpha e^{-\frac{(y-\mu_\alpha)^2}{2(a^2+\sigma^2)}} \tag{31}$$

Without loss of generality, we can assume that $\mu_2 = -\mu_1 \equiv \Delta \geq 0$. Defining the normalized election outcome $\hat{y} \equiv y/\Delta$, we solve eq. (31) to get the following condition:

$$\hat{y} = \tanh(J\hat{y} + h) \tag{32}$$

where $J = \Delta^2/(a^2+\sigma^2)$ and $h = \frac{1}{2}\ln\frac{w_2}{w_1}$. For $w_1 = w_2$ (i.e.

---

[2] For a single maximum $y^*$ with $\int_{-\infty}^{\infty} u''(x-y^*)f(x)dx = 0$, the functional derivative of $y$ is not defined, but, as can be shown in a higher order analysis, there is no instability. In particular, for $\delta y^*$ defined by eq. (23), we can derive

$$\frac{1}{6}(\delta y^*)^3 = \frac{-\epsilon u'_{x_0}(y^*)}{\int_{-\infty}^{\infty} \tilde{f}(x) u''''_x(y^*)dx} + O(\epsilon^{4/3})$$

in place of eq. (25), which yields $\delta y^* \propto \epsilon^{1/3} u'_{x_0}(y^*)^{1/3} + O(\epsilon^{2/3})$. Note that

$$r(f,x_0) = \frac{d}{dx}\Big|_{x=x_0} \delta y^*$$

is well-defined for any given $\epsilon = 1/N$—although it is not given by eq. (4) since $\frac{\delta y}{\delta f(x)}$ does not exist—thus, there is no instability.

[3] For a finite population, the outcome of the election in this model is not deterministic, and the probability of a given candidate winning is continuous with respect to the opinion distribution. This discontinuity in y[f] arising from multiple Nash equilibria in the large-population limit is analogous to a first-order phase transition.

for equally sized subpopulations), $h = 0$, and the election is stable with $\hat{y} = 0$ for $J \leq 1$ and unstable for $J > 1$. In the unstable regime, there are two possible outcomes described by $\pm|\hat{y}^*|$, and an arbitrarily small change in $h$ can cause $\hat{y}$ to swing between its positive and negative value. In this regime, the majority of one of the subpopulations will be negatively represented: for $y^* < 0$, over half of the subpopulation centered at $\Delta$ will have opinions $x$ with $x - y^* > \Delta - y^* > a$ ($\Delta > a$ in the unstable regime), and, from eq. (10), representation is negative for these opinions. Likewise, for $y^* > 0$ in the unstable regime, over half of the subpopulation centered at $-\Delta$ will be negatively represented.

### H. Connection with the mean-field Ising model

We map the voting model that gives rise to eq. (12) (described in detail in section S1 G) onto a mean-field Ising model [16]. To begin constructing this map, note that the left-hand side of eq. (31) gives the limiting value of $y$ as $N \to \infty$ when $y$ is drawn from a probability distribution corresponding to the partition function[4]

$$Z = \int_{-\infty}^{\infty} \left[ \int_{-\infty}^{\infty} e^{-\frac{(y-x)^2}{2a^2}} f(x) dx \right]^N dy \qquad (33)$$

For large $N$, having $N$ voters with deterministic opinions $x_i$ distributed according to $f(x)$ is equivalent to having $N$ voters with probabilistic i.i.d. opinions with PDF $f(x)$. We can therefore view the partition function as describing the voting population, with each voter interacting with the election outcome $y$.

Because $y$ is a gaussian random variable, we integrate over $y$, which yields, up to a multiplicative constant,

$$Z = \int_{-\infty}^{\infty} e^{-\sum_i \frac{(x_i - \bar{x})^2}{2a^2}} \Pi_{i=1}^N (f(x_i) dx_i) = \int_{-\infty}^{\infty} e^{-\frac{1}{2N} \sum_{i,j} \frac{(x_i - x_j)^2}{2a^2}} \Pi_{i=1}^N (f(x_i) dx_i) \qquad (34)$$

This equation describes $N$ interacting probabilistic "spins," with each spin weighted by the opinion distribution $f(x)$, with an energy penalty proportional to its mean-square distance from all of the other spins. For $f(x)$ given by eq. (11), the behavior is exactly that of a mean-field Ising model (with an external magnetic field for $w_1 \neq w_2$); in general, for bimodal symmetric $f(x)$ we expect a phase transition in which the system will spontaneously break the symmetry between the peaks as the peaks move farther apart. In the stable/disordered phase, both of the peaks of $f(x)$ are sampled by the "spins;" in the unstable/ordered phase, however, only one of the two peaks is sampled, and therefore the other peak is not represented. Despite the fact that each individual votes independently from everyone else, voters are coupled through their collectively choosing a candidate, which, as shown above, is equivalent to each voter being coupled to every other voter. In the limit of weak interactions, i.e. $a \to \infty$, we recover mean voting, since in this limit, $u(y-x) = \exp\left[-\frac{(y-x)^2}{2a^2}\right] \approx 1 - \frac{1}{2a^2}(y-x)^2$. (Quadratic utility functions yield mean voting—see section S1 E.) Thus, mean voting is a way of "independently" aggregating opinions.

For a general $u(y - x_i) = \exp[-V(y - x_i)]$ (since $0 \leq u(x-y) \leq 1$, we can always write $u$ in this form with $V(x) \geq 0$), we note that eq. (8) yields an election outcome equivalent to the limit of $y$ as $N \to \infty$ with $y$ drawn from

$$Z = \int_{-\infty}^{\infty} \left[ \int_{-\infty}^{\infty} e^{-V(y-x)} f(x) dx \right]^N dy = \int_{-\infty}^{\infty} e^{-\sum_i V(y - x_i)} \Pi_{i=1}^N (f(x_i) dx_i) \qquad (35)$$

For quadratic $V$, we saw above that we could exactly integrate over the election outcome $y$ to yield pairwise quadratic interactions between the $x_i$ variables, but for general $V$, such an integration will yield many interaction terms of higher than quadratic order between these $x_i$. Although such integration

[4] If the states of a physical system described by the variables $m_1, ..., m_N$ have a ratio of energy to temperature of $f(m_1, ..., m_N)$, the probability density of the system being in any one configuration is given by $\frac{1}{Z} e^{-f(m_1,...,m_N)}$ where the normalization $Z$ is known as the *partition function* and is given by

$$Z = \int e^{-f(m_1,...,m_N)} dm_1 dm_2 ... dm_N = \int e^{-f(m_1,...,m_N)} \Pi_{i=1}^N (dm_i)$$

where each $m_i$ is integrated over its entire domain.

cannot be carried out precisely, we expect this interacting system to undergo a phase transition for bimodal $f(x)$ if the expansion of $V$ produces sufficiently strong interactions. Thus, the system's behavior should be similar to the exactly solvable case in which $V$ is quadratic.

### I. Empirical data

To determine if the stability of U.S. presidential elections has changed over time, we used data from Jordan *et al.* [23] on the polarization in the party platforms. Jordan *et al.* used a combination of machine learning and human judgment to determine which of the frequently used words in the party platforms were polarizing and then determined the number of polarizing words (classified by political issue dimensions such as economic, foreign, etc.) in the Republican and Democratic platforms from 1944 to 2012. From this data, we calculated the total number of polarizing words as a percentage of all words in the platforms. We chose the percentage of polarizing words in the party platforms as a measure of political polarization over other measures of ideology—such as NOMINATE scores [24], which measure ideological purity based on agreement with other politicians—because we wanted an external, content-based measure of divergence in opinion rather than a measure of ideology that depends only on the positions that politicians take relative to one another. To construct fig. 4, we plotted by year the fraction of polarizing words in the Democratic platforms and the negative of the fraction of polarizing words in the Republican platforms. To correct for any time-independent bias affecting the number of polarizing words in the party platforms—for instance, which words Jordan *et al.* designated as polarizing—we subtracted from the data for each party separately the fraction of polarizing words from the year of least polarization for that party (which was 1948 for both parties). Thus, the data shown are the changes in the fraction of polarizing words relative to their baseline value (0.0258 for Democratic platforms and 0.0693 for Republican platforms).

As was noted in the manuscript and explained in section S1 H, our voting model (fig. 3) is equivalent to a mean-field Ising model. Real-world elections are unlikely to follow this model exactly, and even if they did, there is no reason to believe that there would be a simple relationship between polarization (for which $J$ is a dimensionless measure) and time. Nonetheless, if U.S. presidential elections underwent a phase transition in the same *universality class* as the mean-field Ising model, then in the vicinity of the phase transition, polarization would increase in proportion to the square root of the time from the transition, regardless of the precise relationship between time and polarization. In much the same way, magnetization increases near a ferromagnetic phase transition in proportion to $(T - T_c)^{\beta}$, where $T$ is temperature, $T_c$ is the temperature at which the phase transition occurs, and $\beta$ is known as a *critical exponent*, which depends only on the universality class to which the phase transition belongs [25]. Inspired by this universality, we fit the polarization of both parties to the piecewise function

$$f(x) = \begin{cases} 0 & x \leq x_0 \\ A\sqrt{x - x_0} & x > x_0 \end{cases} \tag{36}$$

where $x$ is the year and $f(x)$ is the fraction of polarizing words in that year's platform relative to the baseline value (see above), and $A$ and $x_0$ are free parameters, corresponding to the amplitude of the polarization and the year that it begins, respectively. We found that $x_0 = 1970.54$ and $A = 0.0079196$ minimize the total sum of square errors, yielding $R^2$ values of $0.86$ for the Democratic party and $0.89$ for the Republican party. If the two parties are considered together, $R^2 = 0.87$.

## S2. SUPPLEMENTARY TEXT

### A. Multidimensional opinion space

For the sake of simplicity, this paper focuses on systems with a one-dimensional opinion space, but the concepts developed in this paper can naturally be extended to a multidimensional opinion space, where the opinions of the electorate and candidates lie in $\mathbb{R}^n$, as in [9–11, 26]. This extension will be briefly outlined here. The definition of representation is generalized by replacing eq. (2) with

$$r_{\vec{c},\mu\nu}(f, x) = \frac{\delta y_\mu c_\nu}{|c|^2} \tag{37}$$

For a scalar measure, we use

$$\mathrm{tr}[r_{\vec{c}}] = \frac{\vec{c} \cdot \delta\vec{y}}{|c|^2} \tag{38}$$

When there exists an $r_{\mu\nu}(f, x)$ such that

$$\delta y_\mu = \sum_\nu r_{\mu\nu}(f, x) c_\nu + O(c^2) \tag{39}$$

for all $\vec{c}$, this $r_{\mu\nu}(f, x)$ can be used in place of eq. (3) as a representation independent of $\vec{c}$. In the large-population limit, eq. (19) is then replaced (using Einstein-summation notation) by the path-independent integral

$$\mathrm{tr}[r_{\vec{c}}(f, x)] = \frac{c_\mu}{|c|^2}\left(\frac{\delta y_\mu}{\delta f(\vec{x} + \vec{c})} - \frac{\delta y_\mu}{\delta f(\vec{x})}\right) = \frac{c_\mu}{|c|^2}\int_{\vec{x}}^{\vec{x}+\vec{c}} r_{\mu\nu}(f, x')dx'_\nu \tag{40}$$

where $r(f,x)$ (which satisfies eq. (39)) is a matrix defined by

$$r_{\mu\nu}(f,x) = \frac{\partial}{\partial x_\nu}\frac{\delta y_\mu}{\delta f(x)} \tag{41}$$

The differential representation in a direction given by the unit vector $\hat{v}$ is then given by $\hat{v}_\mu r_{\mu\nu}(f,x)\hat{v}_\nu$, which yields the same results as eq. 7 of [7] in the limit of a continuum of voters. The trace $\mathrm{tr}[r]$ gives a rotationally invariant scalar measure.

The representation normalization condition corresponding to eq. (5) is $\int f(x) r_{\mu\nu}(f,x)dx = \delta_{\mu\nu}$ where the integral is taken over $\mathbb{R}^n$.

In the multidimensional case, instability also implies a failure in representation. Analagously to eq. (6), instability is characterized by

$$\lim_{\vec{c}\to 0} c_\nu r_{\vec{c},\mu\nu} \neq 0 \tag{42}$$

Generally, instability implies either that $\lim_{\vec{c}\to 0} |c|\,\mathrm{tr}[r_{\vec{c}}] \neq 0$, in which case negative representation (defined by $\mathrm{tr}[r_{\vec{c}}] < 0$) follows in the same way as the one-dimensional case, or that an infinitesimal change in opinion causes a finite orthogonal change in the outcome of the election. In this case, by considering further infinitesimal changes in opinion parallel to the first change in election outcome, and assuming that the magnitude of the change in the outcome of the election cannot grow without bound, one either gets negative representation directly ($\mathrm{tr}[r_{\vec{c}}] < 0$ for some $\vec{c}$)—or $\mathrm{tr}[r_{\vec{c}}] > 1$, from which negative representation follows as it does in the one-dimensional case.

### B. The Owen-Shapley index as a special case

In this section we show that there exist functionals $y[f(x)]$ for which our representation measure (eq. (4)) reproduces the values of both the deterministic and probabilistic Owen-Shapley voting power indices. Thus, these voting power indices can be thought of as special cases of our measure. We give a brief background on the voting power literature and then consider the case of a one-dimensional opinion space, followed by a generalization to the case of a multidimensional opinion space for which the Owen-Shapley index was primarily designed.

When nothing is known about the preferences of voters, their political power has traditionally been measured by *a priori voting power* [27], which reflects the probability that a given individual or entity will cast the deciding vote and is usually measured by either the *Penrose index* [28] or the *Shapley-Shubik index* [29]. More precisely, to calculate a voter's *a priori* voting power, consider a random division of the rest of the voters into two camps. Then the probability that the excluded voter will get his way regardless of which camp he joins is his voting power; the Penrose index and the Shapley-Shubik index differ only in the way in which they randomly choose a division. While the Penrose index assumes that each voter randomly chooses one side or the other, the Shapley-Shubik index re-weights the probabilities so that each ordering of voters is equally likely.[5] These indices provide useful and counterintuitive results when the voters possess differing numbers of votes, as in the European Union. For instance, under the 1958 voting rules for the European Economic Community, Luxembourg, despite having had one vote, had no voting power, since there were no possible divisions of the other five countries such that Luxembourg's vote would be decisive [30]. But in elections in which each voter has one vote, all measures of *a priori* voting power result in each voter having an equal amount of power. A measure of voting power that takes voter preferences into account is needed to determine how various opinions are differently represented. Many preference-based measures have been proposed—for instance, the spatial Shapley-Shubik index, also known as the Owen-Shapley index [8]—but, as we will see, these measures implicitly assume that people vote in a particular way.

In one dimension, the deterministic Owen-Shapley index [8] allows for only two possible orderings of the voters (left to right or right to left), for which the median voter is pivotal in both, thus yielding the same concentration of power that our representation measure (eq. (3)) yields in the case of median voting. Benati and Marzetti [31] note that this extreme concentration of power is due to the deterministic nature of the Owen-Shapley model, which assigns zero probability to almost all orderings. They propose a generalized election model in which voters' opinions have both a deterministic and a random component. In the one-dimensional case, their treatment is equivalent to denoting the probabilistic opinion $X_i$ of voter $i$ by

$$X_i = x_i + \epsilon_i \tag{43}$$

where the $x_i$ are deterministic and the $\epsilon_i$ are independent random variables with a continuous probability density function $f_\epsilon(\epsilon_i)$. Denoting the distribution of the $x_i$ over the population by $f(x_i)$ (note that $f$ will be a sum of delta functions for a finite population) and choosing an election in which people vote for the candidate closest to their probabilistic opinions $X_i$,[6] the Nash equilibrium for the two candidates' opinions is the median of the distribution $f_X(X) \equiv (f * f_\epsilon)(x) = \int_{-\infty}^{\infty} f(x) f_\epsilon(X-x)dx$, i.e.

$$y[f] = \mathrm{median}[f_X] = \mathrm{median}[f * f_\epsilon] \equiv m \tag{44}$$

The continuity of $f_X$ follows from that of $f_\epsilon$, and we have made the additional assumption that $f_X(m) \neq 0$; otherwise, there is no unique Nash equilibrium.

We then calculate

[5] There are some fundamental differences in the motivation behind the indices[27], but mathematically, they are rather similar, though neither is without drawbacks: while the assumptions behind the Penrose index are simpler, in general the sum of all voters' Penrose indices will not equal 1,

$$\frac{\delta y[f]}{\delta f(x)} = \int_{-\infty}^{\infty} \frac{\delta\,\mathrm{median}[f_X]}{\delta f_X(z)} \frac{\delta f_X(z)}{\delta f(x)} dz = \int_{-\infty}^{\infty} \frac{\mathrm{sign}(z-m)}{2 f_X(m)} f_\epsilon(z-x) dz = \int_{-\infty}^{\infty} \frac{\mathrm{sign}(z+x-m)}{2 f_X(m)} f_\epsilon(z) dz \tag{45}$$

which yields the representation measure (eq. (4)) for opinion $i$:

$$r(f, x_i) = \frac{d}{dx_i} \frac{\delta y[f]}{\delta f(x_i)} = \int_{-\infty}^{\infty} \frac{2\delta(z+x_i-m)}{2 f_X(m)} f_\epsilon(z) dz = \frac{1}{f_X(m)} f_\epsilon(m-x_i) \tag{46}$$

The rightmost side of eq. (46) is the probability that voter $i$ is the median—i.e. pivotal—voter; thus, $r(f, x_i)$ is equal to the generalized Shapley-Shubik index (SSI) for voter $i$.

The Owen-Shapley index was developed primarily for multidimensional opinion spaces in $\mathbb{R}^n$. Owen and Shapley [8] consider a randomly drawn unit vector $\hat{v} \in \mathbb{R}^n$ and then order individuals by defining $i < j$ if $\vec{X}_i \cdot \hat{v} < \vec{X}_j \cdot \hat{v}$. (The $\vec{X}_i$ are deterministic but can easily be modified to be partially probabilistic as in [31].) The SSI of $i$ is again the probability that $i$ is the median of the resulting ordering. To see how this power index is a special case of our multidimensional representation measure (eq. (41)), consider the following method of choosing a candidate, given a set of voters with (potentially probabilistic) opinions $X_i \in \mathbb{R}^n$:

1) Randomly choose an orthonormal basis $\{\hat{v}_1, ..., \hat{v}_n\}$ for $\mathbb{R}^n$.

2) Let $m_\alpha$ be the median of the probability distribution function for $\hat{v}_\alpha \cdot \vec{X}$, where $\vec{X}$ is randomly drawn from the voter opinions $X_i$.

3) The election outcome is then given by $\vec{y}[f] = \sum_{\alpha=1}^{n} m_\alpha \hat{v}_\alpha$.

Note that $\vec{y}$ is now a random variable, and so the right-hand side of eq. (41) must be replaced by its expectation value, i.e. $r_{\mu\nu}(f, \vec{x}) = \mathbb{E}[\frac{\partial}{\partial x_\nu} \frac{\delta y_\mu}{\delta f(\vec{x})}]$. From eq. (46) (with $\hat{v} \cdot \vec{y}$, $\hat{v} \cdot \vec{x}_i$, and $\hat{v} \cdot \vec{\epsilon_i}$ substituting for $y$, $x_i$, and $\epsilon_i$), $\hat{v}_\mu r_{\mu\nu}(f, \vec{x}_i) \hat{v}_\nu$ is equal to the probability that $i$ will be the median voter along $\hat{v}$. Therefore, $\mathrm{tr}[r(f, \vec{x}_i)]$ is equal to the expected number of basis vectors along which $i$ will be the median, and so $\mathrm{tr}[r(f, \vec{x}_i)]$ is equivalent to $n$ times the SSI.[7]

The agreement between $r(f, x)$ and the SSI follows from the fact that $\frac{d}{d(\hat{v} \cdot \vec{x}_i)} \frac{\delta\,\mathrm{median}[f_{\hat{v} \cdot \vec{X}}]}{\delta f(\vec{x}_i)}$ measures the probability that $i$ is the median voter along $\hat{v}$ for this class of voting models. In this sense, the Owen-Shapley index of power implicitly assumes an election in which some sort of median is chosen. This model is appropriate when voters vote deterministically (although the options presented for them to vote on may be random). But such deterministic voting assumes that voters distinguish between very small differences in policy with 100% certainty, and it also assumes that there is no chance that a voter abstains. While these assumptions may hold for assemblies of elected officials (and in particular to the EU, where these measures are most commonly applied), they tend to fail for mass elections, in which a voter may sometimes vote for the candidate farther from her opinion and sometimes may choose not to vote at all.

while the sum of all voters' Shapley-Shubik indices will.

[6] This election takes the form of the *Random Utility Model* mentioned in section 2.2 of [11].

[7] Note that had we treated the election as deterministic by defining the outcome $y'$ by $y' = \mathbb{E}[y]$ and then defined representation by $\frac{\partial}{\partial x_\nu} \frac{\delta y'_\mu}{\delta f(\vec{x})} = \frac{\partial}{\partial x_\nu} \frac{\delta \mathbb{E}[y_\mu]}{\delta f(\vec{x})}$, we would have arrived at a different answer, since the probability distribution over which the expected value is taken depends on $f(x)$.