SPECIAL FEATURE

COMPUTER SCIENCES

POLITICAL SCIENCES

Polarization and tipping points

Michael W. Macy[a,b,1], Manqing Ma[c,d,1], Daniel R. Tabin[c,d], Jianxi Gao[c,d], and Boleslaw K. Szymanski[c,d,e,2]

[a]Department of Information Science, Cornell University, Ithaca, NY 14853; [b]Department of Sociology, Cornell University, Ithaca, NY 14850; [c]Department of Computer Science, Rensselaer Polytechnic Institute, Troy, NY 12180; [d]Network Science and Technology Center, Rensselaer Polytechnic Institute, Troy, NY 12180; and [e]Institute of Computer Technologies, Społeczna Akademia Nauk, 90-113 Łódź, Poland

Edited by Helen V. Milner, Princeton University, Princeton, NJ, and approved July 2, 2021 (received for review March 1, 2021)

Research has documented increasing partisan division and extremist positions that are more pronounced among political elites than among voters. Attention has now begun to focus on how polarization might be attenuated. We use a general model of opinion change to see if the self-reinforcing dynamics of influence and homophily may be characterized by tipping points that make reversibility problematic. The model applies to a legislative body or other small, densely connected organization, but does not assume country-specific institutional arrangements that would obscure the identification of fundamental regularities in the phase transitions. Agents in the model have initially random locations in a multidimensional issue space consisting of membership in one of two equal-sized parties and positions on 10 issues. Agents then update their issue positions by moving closer to nearby neighbors and farther from those with whom they disagree, depending on the agents' tolerance of disagreement and strength of party identification compared to their ideological commitment to the issues. We conducted computational experiments in which we manipulated agents' tolerance for disagreement and strength of party identification. Importantly, we also introduced exogenous shocks corresponding to events that create a shared interest against a common threat (e.g., a global pandemic). Phase diagrams of political polarization reveal difficult-to-predict transitions that can be irreversible due to asymmetric hysteresis trajectories. We conclude that future empirical research needs to pay much closer attention to the identification of tipping points and the effectiveness of possible countermeasures.

polarization | phase transition | tipping points | hysteresis dynamics

Democratic societies thrive on disagreement, debate, and intense competition among multiple interest groups (1). Nevertheless, there is growing recognition that political division can become a liability for democratic governance when multiple lines of controversy become aligned with partisan identities (2). Political scientists refer to this crystallization of opinion as "constraint" (3) and point to two principal reasons for concern: partisan division and extremism. First, the alignment of substantively unrelated issues (e.g., capital punishment, reproductive rights, and gun control) attenuates intraparty differences, subdues political cacophony, and rearranges an ideological mosaic into two diametrically opposed political camps (4). Second, issue alignment and factional bifurcation allow differences of opinion to reinforce one another such that moderate voices become muffled and the distribution of opinion becomes increasingly bimodal (5). Analyses of roll-call voting show that political elites have adopted more extreme positions on core political issues in recent decades (6), while polarization among voters has become mainly affective rather than ideological (7). This combination of partisan division and political extremism can eviscerate the "cross-cutting cleavages" on which pluralistic diversity depends, thereby undermining the capacity for compromise required for effective democratic governance.

In this paper, we point to a third danger, one that has received too little attention in the growing literature on political and cultural polarization: the existence of a tipping point beyond which the activation of shared interests can no longer bring warring factions together, even in the face of a common threat. Our interest in this problem is motivated by a series of crises that might be expected to activate a broad political identity and unified response: the Great Recession, Russian electoral interference, impending climate catastrophe, a global pandemic, and, most recently, the January 6 attack on the US Congress. It is not surprising that these events prompted a call to arms. What is surprising is the direction in which the arms were pointed.

Our goal is not to explain polarization, about which there is an extensive literature and a lively debate. Instead, we focus narrowly on a phase analysis of polarization, a problem that has largely escaped attention in previous research. We use a general model of opinion dynamics to demonstrate the existence of tipping points, at which even an external threat may be insufficient to reverse the self-reinforcing dynamics of political polarization. Polarization reaches a tipping point when the rate of increase suddenly accelerates and when the process displays a phase change characterized by asymmetric hysteresis loops. The existence of a tipping point in a self-reinforcing dynamic is neither inevitable nor especially counterintuitive. However, the existence of multiple tipping points, one when polarization is increasing and another when it is decreasing, cannot be assumed. Moreover, if the threshold on the downward trajectory falls below the threshold going up, the dynamics can be hard to reverse. This study is motivated by the need to call greater attention to that possibility.

Model

Our model is preceded by the Szymanski model presented in Lu et al. (8). The Szymanski model applies to polarization among

Significance

Our study was motivated by a highly disturbing puzzle. Confronted with a deadly global pandemic that threatened not only massive loss of life but also the collapse of our medical system and economy, why were we unable to put partisan divisions aside and unite in a common cause, similar to the national mobilization in the Great Depression and the Second World War? We used a computational model to search for an answer in the phase transitions of political polarization. The model reveals asymmetric hysteresis trajectories with tipping points that are hard to predict and that make polarization extremely difficult to reverse once the level exceeds a critical value.

Author contributions: M.W.M., M.M., J.G., and B.K.S. designed research; M.W.M., M.M., D.R.T., and B.K.S. performed research; M.W.M., M.M., D.R.T., J.G., and B.K.S. analyzed data; M.W.M., M.M., D.R.T., J.G., and B.K.S. wrote the paper; M.W.M., M.M., and B.K.S. designed simulation experiments; M.W.M., M.M., and D.R.T. implemented software.

The authors declare no competing interest.

This article is a PNAS Direct Submission.

[1] M.W.M. and M.M. contributed equally to this work.

[2] To whom correspondence may be addressed. Email: szymab@rpi.edu.

This article contains supporting information online at https://www.pnas.org/lookup/suppl/doi:10.1073/pnas.2102144118/-/DCSupplemental.

Published December 6, 2021.

legislators in a two-party political system. Legislators reveal their positions on issues through votes on bills. The Rice Index (9) of voting behavior measures partisan polarization. The dynamics are governed by differential equations that assume legislators optimize their chances of reelection by aligning their votes with the needs of their constituents, parties, and donors, whose influence acts as an invisible hand (10), balancing legislators' and parties' commitment to partisan and bipartisan strategies. The authors fitted the model parameters to training data and demonstrated its strong predictive power on voting records of the last 30 US Congresses.

Following Lu et al. (8), we model partisan polarization at the elite level, not in the general population. (By endowing a social network with local structure, our model becomes applicable to much larger populations with lower density of interaction, but we leave that investigation to future research.) However, unlike Lu et al. (8), the model in this study is not intended as a tool for empirical prediction. Instead, we use a theoretical model to look for tipping points in the system-level dynamics that emerge out of a set of assumptions about interpersonal interaction. Although a legislative body is the prototypical application, unlike Lu et al. (8), our model is not based on the Congress of the United States or any other country, and there are no legislature-specific institutional assumptions that would prevent the model from being applied to any small, densely connected organization. By design, the model omits distinctive features that would limit its applicability to a particular organizational setting, such as pressure from lobbyists or constituents, internal organizational structure that constrains access to other members, or power differentials between organizational leaders who can sanction noncompliant members. The model avoids specific institutional practices, customs, or formal rules that might constrain behavior. These complications are abstracted away so that we can investigate general properties of tipping points in the polarization of opinion that might arise in diverse organizational settings. A model that does not correspond to any particular empirical instantiation can thereby also be said to capture lawful regularities that may be broadly relevant.

The model in this study also differs from Lu et al. (8) in that it is agent-based rather than equation-based. An agent-based model can be conceptualized as a "population of models," in contrast with an equation-based "model of a population" (11). In the latter, features of the system interact (e.g., the size, density, clustering, and polarization of the population). In an agent-based model, the agents interact, such that the properties of the system emerge out of the dynamics of their interactions. Like equation-based models, agent-based models can be empirically calibrated for predictive accuracy. Alternatively, the models can also be highly abstract, based on a set of simple, but plausible, behavioral assumptions, similar to game-theoretic models used to explore the complexity of cooperation (12) or tipping points in neighborhood segregation (13). Both approaches—empirical prediction and theoretical exploration—are useful for the study of polarization, but the approach here is theoretical. As in game-theoretic applications, the goal is to generate hypotheses, leaving their testing to follow-up empirical studies.

Attribute Set. Each agent has an attribute set **M** consisting of the agent's party identity and the agent's positions on issues. Although legislators occasionally change parties, we assume that party identity is a fixed binary attribute. For simplicity, we assume there are only two parties, each with 50 members (like the current US Senate), but the model generalizes to multiparty systems with a governing (majority) coalition and a minority opposition. (Robustness tests show that the two-party results generalize to multiparty systems; see *SI Appendix*, Figs. S5–S7 for details.)

Unlike party identification, positions on issues are continuous and susceptible to change. An agent's position on an issue can range continuously from $+1$ (extreme support) to -1 (extreme opposition). We limit the number of salient issues to 10, which is more than sufficient to ensure that polarization must arise out of cross-cutting cleavages; i.e., two agents who agree on some issues might nevertheless disagree on others. (See *SI Appendix*, Fig. S1 for robustness tests for the number of dimensions.) Were the model limited to only one issue dimension, such cross-cutting divisions could not occur. In contrast, if two agents happen to disagree on one issue, agreement on a second issue will greatly increase the similarity between the agents, and the resulting increase in attraction could then lead one of the two agents to switch sides on the issue on which the agents disagreed. Although a single issue is insufficient, there is little to be gained by having more than 10 issues. If two agents happen to disagree on 10 issues, agreement on an 11th issue will have relatively little effect on the overall similarity between their opinion profiles. In short, allowing only one issue would make polarization trivially easy, but as more dimensions are added to the model, the distribution of opinion on each additional dimension has a declining effect on the polarization dynamics.

Influence. We incorporate two core assumptions that inform previous computational models of opinion dynamics: influence and homophily (14–17). Influence refers to repositioning of an agent on issues in response to neighbors to whom the agent is attracted. Homophily refers to an agent's attraction to a network neighbor that increases with the similarity of their attribute sets. Influence and homophily can create a self-reinforcing dynamic, in which agreement strengthens influence, which then attenuates remaining disagreement.

Nearly all previous models of opinion dynamics shared the assumption that influence is positive, thereby strengthening agreement. However, Schelling (13) famously showed how stable division (e.g., residential segregation) emerges when influence is negative, thereby exacerbating differences, as when in-group neighbors are influenced to move out in response to out-group members who move in. In this context, negative influence is not to be confused with influence to adopt a peer's negatively valued trait (e.g., smoking or vandalism). Rather, negative influence refers to the tendency for an in-group member to differentiate from an out-group neighbor. Applied to opinion dynamics, negative influence (or "distancing") has also been shown to generate polarization into opposing ideological camps (18–20), an empirical pattern that is dramatically illustrated by historical data on partisan differences of opinion regarding global warming (*SI Appendix*, Fig. S10). We model positive and negative influence by assuming that an agent adjusts its positions on issues to be closer to a neighbor to whom the agent is attracted and to differentiate from a neighbor from whom the agent is repelled (4).

Homophily. Axelrod (14) showed how homophily can create local communities demarcated by impermeable cultural boundaries that preclude influence between members of different communities that have no shared attributes. Although Axelrod assumed discrete attributes, the dynamics were generalized to continuous attributes in models of "bounded confidence" (15, 16). Even with discrete attributes, cultural differentiation in the Axelrod model has been shown to eventually collapse into monoculture due to cultural drift (21).

Importantly, collapse into monoculture does not occur if homophily (the principle that "likes attract") is combined with xenophobia, in which "opposites repel" (4). As with bounded confidence models (15, 16), our agents have a threshold for tolerance of disagreement. The threshold defines a reference distance that an agent compares with the distance to a neighbor when deciding whether to move closer to or farther from this neighbor. As the reference distance approaches one, the agent becomes highly intolerant, such that even small disagreements on the issues are unacceptable. As the reference distance decreases, the agent becomes more open-minded and willing to listen to and compromise, even with those whose opinions on the issues differ sharply.

Formally, agents' responses to their neighbors depend on their proximity to their neighbors in an **M**-dimensional attribute space that includes party identity and positions on issues. The closer an agent is to its neighbor across all dimensions, the more likely the neighbor's influence will be positive rather than negative. An agent i's distance D_{ij} to a neighbor j is the weighted average of their Euclidean distances on the 10 issues and their distance as party members. Agents from the same party have zero partisan distance, and agents from different parties have distance one. The agents also have continuous Euclidean distances over the 10 issues, rescaled to a minimum of zero (the agents are identical on all issues) to one (the agents maximally disagree on all issues). The overall distance between two agents is then the weighted average of their partisan distance and ideological distances:

$$D_{ij} = D_{ij}^{party}\beta + D_{ij}^{issues}(1-\beta), \qquad \textbf{[1]}$$

where the weight $\beta \in [0, 1]$ is a continuous exogenous parameter that corresponds to party identification.

In sum, our model borrows from the classic models of Schelling (negative influence) and Axelrod (homophily), along with models of bounded confidence (threshold distances). Our contribution is not in proposing a new model of opinion dynamics or a new explanation of polarization. Instead, we narrowly focus on a single question: Is there a tipping point beyond which polarization becomes difficult or impossible to reverse, even in response to exogenous shocks that pose a shared threat?

Exogenous Shocks. Following Condie and Condie (22), we model exogenous shocks as potentially unifying events, such as a global pandemic, economic collapse, attack by a foreign adversary or terrorist (whether foreign or domestic), or impending climatic catastrophe. The shock is implemented by creating an additional issue on which all agents strongly agree. (Our model also predicts other responses to exogenous shocks, such as changes in party affiliation, but we leave these possibilities for future investigations and focus here on bipartisan unity in the face of a common threat.) The responses to shocks are driven by the levels of polarization at the time they arise. The shock occurs at a level of polarization that is manipulated as a control parameter. When the shock occurs at time step t_s, every agent i adds a new issue $\mathbf{M}+1$ and sets it to $Z_{i,t_s} = 1$, regardless of party identity or positions on the other 10 issues. An exogenous weight parameter γ controls the importance of the shock relative to the other 11 attributes in calculating D_{ij}. Agents then update their positions on the shock issue in the same way they update positions on other issues through the positive and negative influence of other agents.

Model Execution. At time $t = 0$, each agent i is assigned an initial vector $Z_{i,0}$ of $\mathbf{M} = 11$ dimensions. The first dimension is binary and static, corresponding to party identity. The other 10 dimensions are continuous and change over time, corresponding to positions on issues. Agents are initialized with a randomly assigned party identity (either -1 or 1), while location on each issue is randomly drawn from a normal distribution with mean zero and SD 0.25, with the range truncated where necessary to the unit interval in absolute value. Thus, at time $t = 0$, the parties have no preexisting ideological differences, and an agent's position on a given issue cannot be predicted from the agent's party identity or position on any of the other dimensions. The random start is not meant to correspond to the historical origins of political parties in the real world. On the contrary, newly formed parties often emerge with extreme polarization levels, as noted by Madison in *The Federalist Papers*. However, previous theoretical research cautions against starting the model in a converged state, since doing so might obscure the possibility that the system dynamics cannot reach that equilibrium. Instead, van Strien et al. (23) and Brändle et al. (24) demonstrate that repeating initialization with random starting positions in agent-based models is a reliable way of finding stability points. In addition to the random start, we model hysteresis loops by starting in a completely polarized configuration to see if polarization eventually collapses as the control parameters decline. (See *SI Appendix*, Figs. S1 and S5–S7 for robustness tests for the number of dimensions and parties.)

Following initialization at time step $t = 0$, an agent i is randomly chosen to update its state at the next time step $t + 1$ by the influence of a neighbor j randomly chosen without regard to party; i.e., members are equally likely to be influenced by members of the other party. The agent then updates its state $Z_{i,t}$ to a new state $Z_{i,t+1}$ in response to the state of its neighbor, $Z_{j,t}$, as follows. First, i assesses its proximity to j relative to a continuous intolerance parameter $\alpha \in [0, 1]$ that is identical for all agents and represents the minimum distance for tolerance of disagreement. If $\alpha = 0$, every distance D_{ij} will be tolerated, and all influence will be positive; thus, i will move closer to j on each issue. If $\alpha = 1$, no distance will be tolerated, and all influence will be negative; thus, i will move away from j on each issue.

Formally, the probability of positive influence (P_+) is a cumulative logistic function of the distance D_{ij} between i and j with steepness m as follows:

$$P_+ = \frac{1}{1 + e^{m\left(D_{ij}-(1-\alpha)\right)}}. \qquad \textbf{[2]}$$

We set $m = 10$, but the results are robust as the function approaches a deterministic threshold with $m \gg 10$. However, as m approaches one, the function becomes linear, which introduces sufficient noise to attenuate (but not eliminate) polarization. (See *SI Appendix*, Fig. S2 for tests of the robustness of the results to different values of m.)

If j's influence is positive, $Z_{id,t+1}$ is a weighted average of $Z_{id,t}$ and $Z_{jd,t}$ on each dynamic dimension d:

$$Z_{id,t+1} = Z_{id,t} + (L - Z_{id,t})|D_{ij} - c|ran, \qquad \textbf{[3]}$$

where ran is a uniform random real number in the unit interval, d is the dimension that is updated, $c = 1$, and $L = Z_{jd,t}$. The expression $L - Z_{id,t}$ in Eq. **3** defines the distance i must move to reach j on d. The expression $|D_{ij} - c|$ means that the fraction of that distance that i will move decreases with the overall distance between i and j across all **M** dimensions (including d). Simply put, the smaller the disagreement on a given issue, the easier it is to resolve, in accordance with the principle of homophily (or likes attract).

If j's influence is negative, then $c = 0$ and i moves away from j in the direction that increases the distance from i to j on each dynamic dimension d. The value of L depends on values of $Z_{id,t}, Z_{jd,t}$. If $Z_{id,t} > Z_{jd,t}$ or $Z_{id,t} = Z_{jd,t} < 0$, then $L = 1$. If $Z_{id,t} < Z_{jd,t}$ or $Z_{id,t} = Z_{jd,t} > 0$, then $L = -1$. If $Z_{id,t} = Z_{jd,t} = 0$, and L is chosen randomly to be 1 with probability 0.5 and -1 otherwise. As the distance from i to j increases along each dynamic dimension d, the maximum possible distance from i to L decreases; i.e., the distance that i needs to move to be as far away from j as possible decreases, while the fraction of that distance that i moves away from j increases with the overall distance D_{ij}. In short, negative influence is the mirror image of positive influence: The larger the disagreement between neighbors, the harder it is to resolve, in accordance with the principle of xenophobia (or opposites repel).

After updating all 10 issues, another i, j pair is randomly drawn, and the time step increments by one. The process iterates until it either stabilizes in a steady state or reaches an arbitrary time limit.

The model records two measures of polarization at each time step: partisan polarization and extremism. Partisan polarization is measured as the expected difference on a randomly chosen issue between randomly chosen members of the two political parties. The range is from zero (no difference on any issue) to one (maximum possible differences on every issue). Extremism is measured as the expected SD for a randomly chosen issue (i.e., the mean of the SDs over all issues).

Parameters. We manipulate four global parameters hypothesized to exhibit tipping dynamics:

- α is the minimum distance for tolerance of disagreement, used to investigate the effects of intolerance toward those with dissimilar issue positions and party identity, as α increases from zero (complete tolerance) to one (complete intolerance).
- β is the importance of party identity relative to positions on issues in D_{ij}, used to investigate the effects of party identification as β increases from zero (party is ignored) to one (issues are ignored).
- γ is the strength of the shock relative to all other dimensions in D_{ij}, used to find the magnitude of the shock needed to reverse polarization as γ increases from zero (shock is ignored) to one (all else is ignored).
- σ is the level of extremism at which an exogenous shock occurs, used to investigate the reversibility of polarization as σ increases from zero (no polarization) to one (complete polarization).

Although intolerance and party identification are agent-level attributes that can vary locally within an organization, our interest is in comparisons across organizations, not across individuals within an organization. We therefore assume every agent in the entire population has an identical value on these parameters and observe how the level of polarization changes as this value differs across organizations. In addition, when the parameters take limiting expected values, the population variance must go to zero. By holding the variance constant at zero, we isolate the effect of manipulating the parameter level from changes in parameter variance. (We tested the robustness of the results for midrange parameter settings using Gaussian distributions and found little change.) Lastly, the logistic probability function in Eq. **2** means that two agents can have opposite responses to a neighbor, even when all else is identical.

Intolerance and party identification are also modeled as independent control parameters. The correlation between these parameters has varied historically. For example, in the 1960s, party identification was relatively low in the United States, and some of the most intense conflicts happened within parties (e.g., divisions over civil rights, Vietnam, and environmental protection). More recently, intense partisanship has been accompanied by intolerance of disagreement, especially disagreement from within one's own party. We use a computational experiment to identify the independent effects of two processes that are rarely independent in natural settings.

Results

We analyze tipping points in polarization as we vary the levels of party identity, tolerance of disagreement, and the timing and importance of exogenous shocks. Fig. 1 frames the central question that motivates our study: Can we reverse polarization by reducing the level of the factors that cause polarization to increase? The answer is "yes" if the phase transition is linear or sigmoidal (Fig. 1 *A* and *B*). The answer is "maybe" if the transition is a first-order knife-edge transition with a sudden regime shift (Fig. 1*C*) or a hysteresis loop with two different thresholds for forward and backward transitions (Fig. 1*D*). The answer is "no" if the transition involves an asymmetric hysteresis trajectory with bifurcation (Fig. 1*E*).

Fig. 1*C* illustrates the intuition that if polarization increases as the control parameter increases, then we might expect that polarization also decreases as the control parameter decreases, even when correlated vectors are characterized by a phase transition (as in Fig. 1*C*). In short, Fig. 1*C* provides the context for calling attention to the possibility (illustrated in Fig. 1 *D* and *E*) that polarization might instead *fail* to decrease. The key difference between Fig. 1 *D* and *C* is the existence of two critical points: C_P, at which polarization starts to rise uncontrollably, and C_R, at which recovery from high polarization is possible. If $C_P > C_R$ (i.e., the loop runs counter-clockwise), reducing the level of the control parameter to C_P would not be sufficient to reduce the level of polarization; the control parameter needs to drop all the way down to C_R. Hence, the greater the asymmetry in the critical values, the farther the control parameter must be reduced below C_P before there will be any effect on polarization.

Fig. 1*E* shows an asymmetric hysteresis loop that is irreversible. The important difference from Fig. 1*D* is the

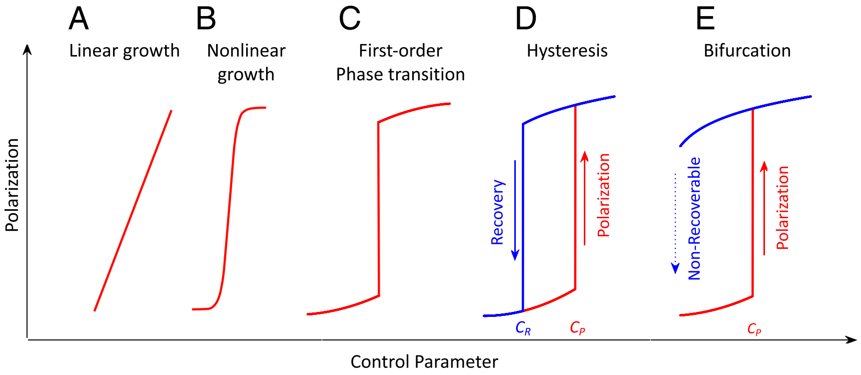

Fig. 1. Schematic of phase transitions. *A* and *B* show continuous phase transitions that lack a tipping point. *C* is a discontinuous knife-edge (first-order) transition that is hard to predict but usually reversible. *D* illustrates a phase transition with a sudden catastrophic shift between alternative stable states in a hysteretic loop with two different thresholds for forward transition to polarization (C_P) and backward transition to recovery (C_R). A reduction in the control parameter to below C_P is insufficient for recovery but a further reduction could eventually allow recovery to occur. In *E*, recovery is no longer possible, no matter how low the control parameter might go.

disappearance of C_R. That is because the "gap" between the critical points in Fig. 1*E* is even more extreme than it is in Fig. 1*D*, and C_R is below the minimal value of the control parameter. In short, in Fig. 1*D*, a reduction in the control parameter to below C_R is insufficient for recovery, but a further reduction could eventually allow the system to recover. In Fig. 1*E*, recovery is no longer possible, no matter how low the control parameter might go.

Figs. 2–4 reveal a hysteresis phase transition as we increase the levels of party identity (Fig. 2), intolerance of disagreement (Fig. 3), and the strength of an exogenous shock (Fig. 4). Fig. 2 shows the phase transition for party identity using two values of intolerance ($\alpha = 0.1$ and $\alpha = 0.3$; we examine additional parameter combinations in Fig. 5). Polarization is measured as partisan differences (Fig. 2 *A* and *C*) and extremism (Fig. 2 *B* and *D*). (The phase transition is illustrated with a single realization of the model; see *SI Appendix*, Fig. S3 for nearly identical results based on the mean over 20 realizations.) The recovery trajectory with $\alpha = 0.1$ (Fig. 2*A*) has a much smaller critical value (C_R) than the forward trajectory. This means that a reduction in the strength of party identity will have no effect on reducing the level of polarization unless the strength of party identity declines far below C_P, the critical value for polarization to increase. The explanation is intuitive: Once the two parties differ on the issues, polarization no longer requires a strong party identity (relative to the importance of issues) in assessing the distance to other agents (D_{ij}).

With slightly greater intolerance ($\alpha = 0.3$), the dynamics become even more troubling. In Fig. 2 *C* and *D*, party identity does not need to be strong in order to trigger a phase transition to high polarization. Moreover, the hysteresis trajectory becomes asymmetric, indicating bifurcation and the irreversibility of polarization, even if the strength of party identity were to drop to zero. In sum, Fig. 2 shows that, even with minimal intolerance, the differences between the two parties will be difficult to reduce by lowering the strength of party identity. With only moderate intolerance, it becomes impossible.

Fig. 3 is similar to Fig. 2, except that the dynamics are driven by intolerance. Fig. 3 *A* and *B* show the phase transition for two values of party identity, $\beta = 0.3$ and $\beta = 0.5$. Fig. 3 *A* and *C* measure polarization as partisan differences and Fig. 3 *B* and *D* as extremism. The results show that, above the critical point for polarization, a subsequent decrease in political intolerance will not bring the parties back together (Fig. 3*A*) or reduce extremism (Fig. 3*B*) until the level of intolerance has dropped far below the level at which polarization increased. Fig. 3 *C* and *D* show that an increase in party identity shrinks the width of the hysteresis loop, but as we observed in Fig. 2, this does not mean polarization becomes more easily reversed. On the contrary, an increase in party identity reduces C_R even further, approaching the point of bifurcation, at which polarization cannot be reversed.

Fig. 4 reports the phase analysis for an external shock about which all agents initially agree. In all three panels $\alpha = \beta = 0.5$. When the shock occurs, there is a "tug of war" between the ability of the shock to spread agreement to the other issues vs. the other issues spreading disagreement to the shock. Fig. 4*A* shows a clockwise hysteresis loop as the strength of the shock increases. There is a critical level of γ (C_R) above which a shock nearly always leads to recovery, even if the shock occurs at or near full polarization. There is a second critical level of γ ($C_P < C_R$), below which a shock almost never leads to recovery, even if the shock occurs at or near zero polarization. Between these two critical points, a shock may or may not lead to recovery, depending on the level of polarization at which the shock occurs. Fig. 4 *B* and *C* illustrate bifurcation in the polarization trajectories for shocks that occur within the critical region depicted

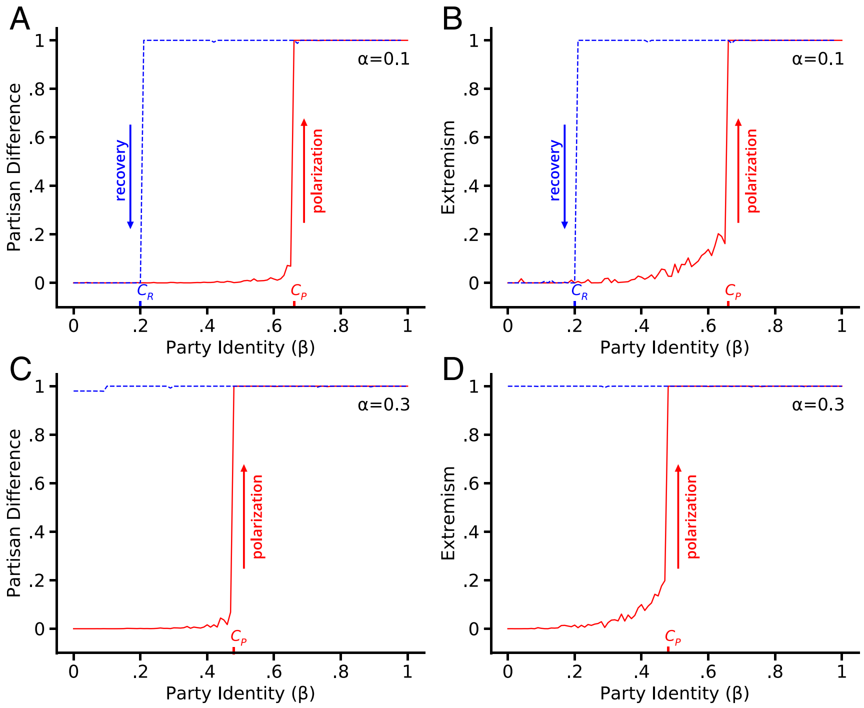

Fig. 2. Tipping points in the level of party identity. Polarization is measured as partisan difference (*A* and *C*) and extremism (*B* and *D*). *A* and *B* show a higher critical value in the polarizing trajectory than in the recovery trajectory. This means that there would be little effect on partisan divisions (*A*) or extremism (*B*) should the strength of party identity drop below C_P, the critical value at which rising polarization suddenly explodes. C_P drops sharply when the level of intolerance increases from α = 0.1 to α = 0.3 (*C* and *D*); i.e., party identity does not need to be strong in order to trigger a phase transition to high polarization. Moreover, the hysteresis trajectory becomes asymmetric, indicating bifurcation and the irreversibility of polarization, even if the strength of party identity were to drop to zero.

https://doi.org/10.1073/pnas.2102144118

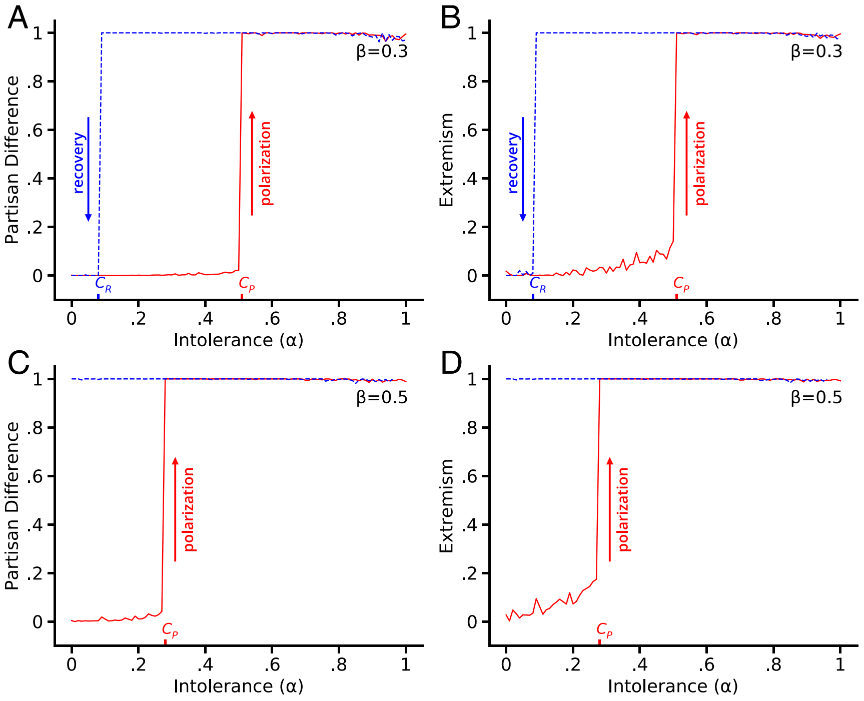

Fig. 3. Tipping points in the level of intolerance. The tipping point for intolerance is similar to the dynamics in Fig. 2. *A* and *B* show the phase transition for two values of party identity, β = 0.3 and β = 0.5. Beyond a critical point C_P, a subsequent increase in political tolerance will not bring the parties back together (*A*) or reduce extremism (*B*), unless the level of intolerance has dropped far below the level at which polarization increased. *C* and *D* show that an increase in party identity shrinks the size of the hysteresis loop by reducing C_R even further, approaching the point of bifurcation at which polarization cannot be reversed.

in Fig. 4*A* (γ = 0.45). The black lines show the trajectories of the 10 preexisting issues (excluding the shock), and orange shows the trajectories of disagreement regarding the shock. The shock occurs at a level of polarization (σ = 0.65) at which the probability to polarize reaches 0.5. Fig. 4*B* shows how a shock can reverse a polarizing trajectory, while Fig. 4*C* shows how the recovery can also fail, even though all parameters are identical to those in Fig. 4*B*. Although the probability of polarization increases linearly with the level of polarization at which the shock occurs, there is no change in the bifurcation of the trajectories. (See *SI Appendix*, Fig. S8 for changes in the probability of recovery with changes in the level of polarization at which the shock occurs and for shocks with higher and lower salience; see *SI Appendix*, Fig. S4 for populations with higher and lower levels of party identification and intolerance.)

Fig. 5 generalizes the results in Figs. 2 and 3 to the entire range of party identity and intolerance. The red surface shows the forward trajectory as polarization increases, and the blue surface shows the recovery. The critical points (where the trajectory experiences a sharp change) are indicated in green along the cliff edge. The void between the red and blue regions corresponds to the hysteresis loops in Fig. 3 (in Fig. 5 *A* and *B*) and in Fig. 2 (in Fig. 5 *C* and *D*). The width of the loops decreases from front to back, due to the larger decrease in the critical values for polarization compared to the decrease in critical values for recovery.

In all four panels, polarization becomes increasingly hard to reverse as party identity and intolerance increase. It is not surprising that polarization increases with the strength of party identity and intolerance. Intuitively, we would then expect polarization to diminish should party identity and intolerance decline. Fig. 5 shows that this is not necessarily the case. Across much of the parameter space, polarization tends to persist, even after the conditions that caused it are abated.

Discussion

We use a general model of opinion dynamics to demonstrate the existence of tipping points, at which even an external threat, such as a global pandemic, economic collapse, foreign adversary, climate change, or a violent assault on Congress, may be insufficient to reverse the self-reinforcing dynamics of partisan polarization. Polarization reaches a tipping point when the rate of increase suddenly accelerates and when the process displays a phase change characterized by an asymmetric hysteresis loop. The model applies to small, densely connected organizations like a legislative body, but abstracts away empirical particularities, such as organizational structure and institutional arrangements, in order to investigate phase transitions in the self-reinforcing dynamics of influence and homophily. Our modeling strategy most closely resembles Schelling's (13) seminal demonstration of a tipping point in neighborhood segregation using a simple checkerboard. The core assumption in our model is the self-reinforcing dynamics of influence and homophily: Influence leads individuals to adopt the opinions of those to whom they are attracted and to differentiate from an out-group, while homophily sorts the population into mutually antagonistic clusters of like-minded combatants.

We measure polarization as the levels of extremism and partisan division. Extremism in the distribution of opinion indicates the erosion of common ground on which people can agree. Partisan division indicates the alignment of issues and the disappearance of the "cross-cutting cleavages" that pluralists have long regarded as the bedrock of a stable democracy (25).

Tipping points in political polarization are important for two reasons. First, the existence and location of phase transitions can be hard to predict, obscuring the risk of an impending collapse of common ground. Second, phase transitions can preclude the ability to recover the previous state, as occurred with the split of Czechoslovakia in 1993. We look for tipping points by

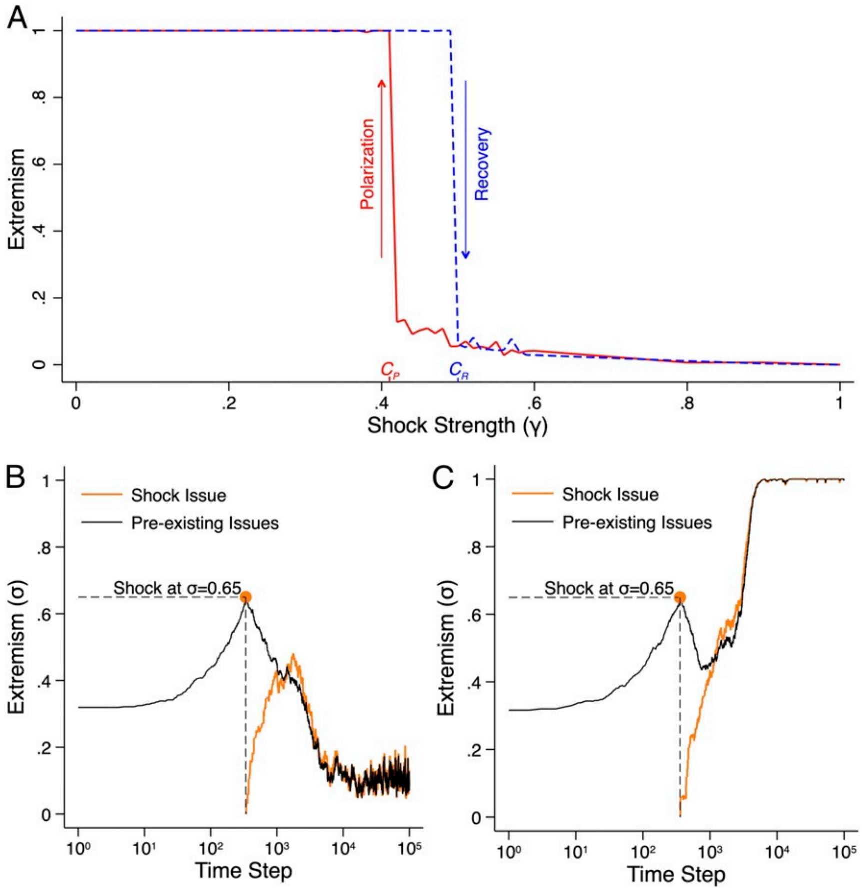

Fig. 4. Tipping points in the effects of an exogenous shock. In all three panels, $\alpha = \beta = 0.5$. *A* shows a clockwise hysteresis loop traversed by the system with the increasing strength of an external shock, about which all agents initially agree. *B* and *C* illustrate bifurcation in the polarization trajectories for shocks that occur within the critical region depicted in *A* ($\gamma = 0.45$). The black lines show the trajectories of the 10 preexisting issues (excluding the shock), and orange shows the trajectories of disagreement regarding the shock. The shock occurs at a level of polarization ($\sigma = 0.65$) at which the probability to polarize reaches 0.5. *B* shows how a shock can reverse a polarizing trajectory, while *C* shows how the recovery can also fail, even though all parameters are identical to those in *B*. See *SI Appendix*, Fig. S9 for the effects of an exogenous shock on partisan difference.

manipulating three exogenous parameters: party identification, intolerance of those who disagree, and external shocks that introduce a new issue on which a population is united. Phase diagrams reveal a difficult-to-predict transition that can be irreversible if recovery follows an asymmetric hysteresis trajectory. Above a threshold level of polarization, remediation may be unable to offset the self-reinforcing dynamics of increasing division, such that polarization becomes difficult or impossible to reverse.

The tipping dynamics of party identity are revealing. Intuitively, it might appear that a decline in issue importance relative to party membership would lead to interparty disagreement, but that is not what we observe. Instead, party identity becomes an organizing template that aligns issue positions into a polarized pattern. The level of partisan division undergoes a knife-edge transition at a critical point in the level of party identification. Moreover, once issues have become aligned with a party, the polarization persists unchanged, even if party identification were to all but disappear. Political intolerance displays similar phase dynamics, with a hard-to-predict critical point beyond which polarization becomes unlikely to reverse, even with an infusion of open-minded pragmatism.

The response of the system to an external shock is particularly alarming. The shock rallies both sides to a common cause on a highly salient issue. Above a critical level of salience (i.e., the strength of the shock relative to preexisting issues), unity in the face of a common threat brings people together, increasing the strength of mutual attraction and attenuating disagreement on previously contentious issues. The shock thus reverses what had been a steadily rising level of partisan and ideological division. However, the opposite happens if the level of salience falls below a tipping point. Instead of coming together, deep divisions on other issues can metastasize to the external shock, depending on the level of polarization when the shock occurs. The implications of our findings go beyond the demonstration that polarization can attenuate the unifying response to external shocks. Future research might also examine the effects of a shock on party realignment, as happened in the 1930s New Deal realignment following the economic collapse of 1929 (26). If parties are internally divided in their response to the shock, our model predicts a dynamic in which factions within opposing parties could become attracted to one another and repelled by members of their own party. In short, the results suggest hypotheses about the timing, length, and impact of shocks that present an agenda for future empirical research that is of both theoretical and practical importance.

We see indications of empirical validation in recent events. Prior to 2019, one might have assumed that a global pandemic would bring together those who disagreed on issues for which hot-button righteous indignation was a luxury that could no longer be afforded. Instead, mask wearing became a partisan crest that identified friend and foe on a partisan battlefield. Similar dynamics can be observed in the two impeachment trials of

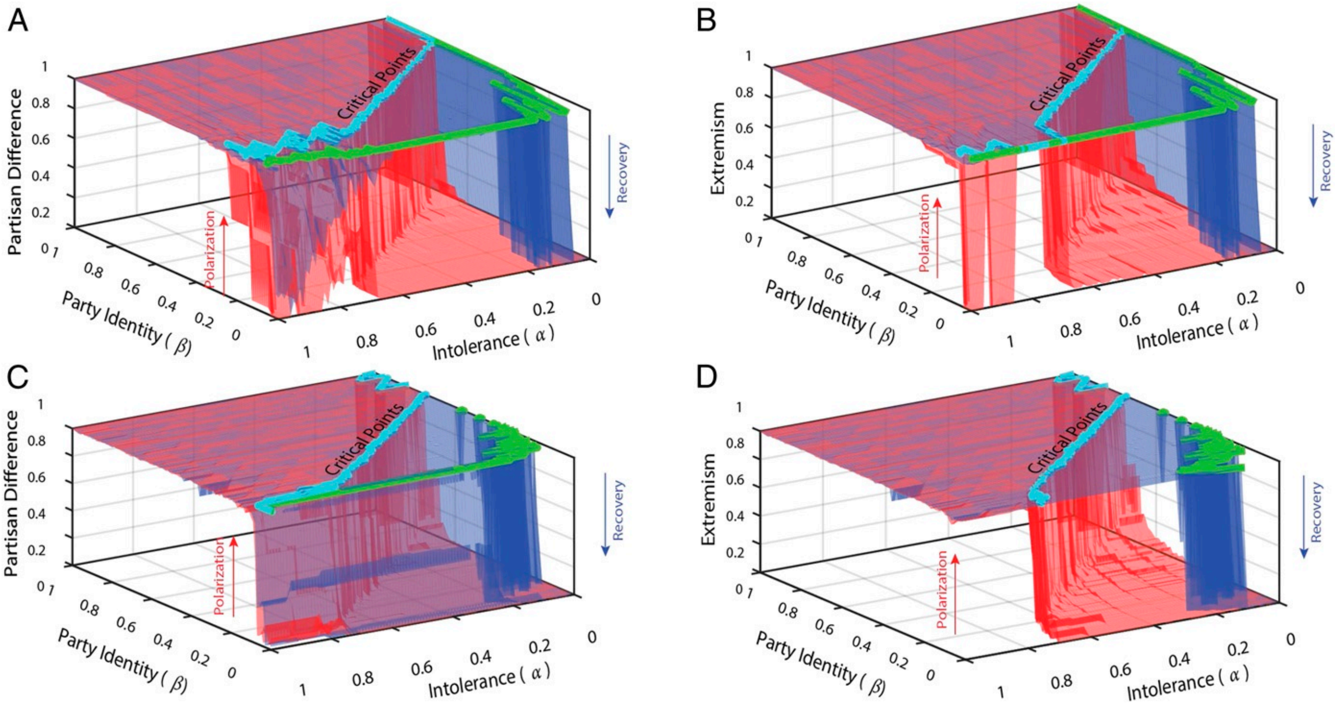

Fig. 5. Robustness tests over the entire range of party identity and intolerance. The red surface shows the forward trajectory as polarization increases and the blue surface shows the recovery. The critical points (where the trajectory experiences a sharp change) are indicated in green along the cliff edge. The void between the red and blue regions corresponds to the hysteresis loops in Fig. 3 (in *A* and *B*) and Fig. 2 (in *C* and *D*). The critical values fluctuate widely for very small α and β. The width of the loops decreases along with the increase of the control parameters due to the larger decrease in the critical values for polarization compared to the decrease in critical values for recovery. In all four panels, polarization becomes increasingly hard to reverse as party identity and intolerance increase.

former President Donald Trump. In the first trial, evidence of collusion with a foreign government failed to exert the expected unifying effect. In the second case, an attack on the US Capitol initially elicited bipartisan outrage, followed by a reversal of position among Republican leaders in the weeks leading up to the Senate trial.

Nevertheless, we make no claims about the model's predictive accuracy. The model is highly abstract and remains to be empirically calibrated and tested. Nor can we assume that tipping points would obtain in radically different applications, such as affective polarization, voter polarization, or media polarization. For example, social media and cable news have been credited with replacing the "median voter" with partisan "echo chambers" and "filter bubbles," while other studies have concluded that media polarization is more of a symptom than a cause of ideological division (27); thus, it remains an open question whether future algorithmic changes in Facebook's News Feed could reverse the process. Future research is needed to investigate the possibility that tipping dynamics generalize to other models of polarization.

In closing, our study should be viewed as a small, but important, first step. The sources of political and ideological polarization have been widely investigated, but relatively little attention has been directed to the possibility that the causal mechanisms are characterized by irreversible tipping points. The lack of attention does not reflect the low importance of the problem. The historical lesson from climate research may be instructive. As with incremental global warming (28), the dynamics of reversibility cannot be revealed with observational data tracking changes over time in the level of polarization. Instead, climatologists have relied on increasingly sophisticated and empirically calibrated computational models to show how the self-reinforcing dynamics of global warming can be reversed through the reduction of carbon and methane emissions only up to a critical threshold, beyond which civilization as we know it may be doomed. We extend the concern with an environmental tipping point to the study of polarization. The need for empirical calibration in our model calls for increased investment in the study of irreversible phase change, while our findings call for urgency in mobilizing remediation efforts before it is too late.

Data Availability. Simulated data have been deposited in GitHub (29).

ACKNOWLEDGMENTS. M.W.M. was supported by NSF Division of Social and Economic Sciences Grant1756822. M.M. and B.K.S. were supported by Army Research Office Grant W911NF-16-1-0524. J.G. was supported by NSF Grant 2047488; and the Rensselaer–IBM AI Research Collaboration.

1. D. B. Truman, *The Governmental Process* (Alfred A. Knopf, New York, 1951).
2. D. Baldassarri, A. Gelman, Partisans without constraint: Political polarization and trends in American public opinion. *AJS* **114**, 408–446 (2008).
3. P. E. Converse, "The nature of belief systems in mass publics" in *Ideology and Discontent*, D. E. Apter, Ed. (Free Press, New York, 1964), pp. 206–261.
4. D. DellaPosta, Y. Shi, M. Macy, Why do liberals drink lattes? *AJS* **120**, 1473–1511 (2015).
5. P. Dreyer, J. Bauer, Does voter polarisation induce party extremism? The moderating role of abstention. *West Eur. Polit.* **42**, 824–847 (2019).
6. N. McCarty, K. T. Poole, H. Rosenthal, *Polarized America: The Dance of Ideology and Unequal Riches* (MIT Press, Cambridge, MA, 2006).
7. S. Iyengar, Y. Lelkes, M. Levendusky, N. Malhotra, S. J. Westwood, The origins and consequences of affective polarization in the united states. *Annu. Rev. Polit. Sci.* **22**, 129–146 (2019).
8. X. Lu, J. Gao, B. K. Szymanski, The evolution of polarization in the legislative branch of government. *J. R. Soc. Interface* **16**, 20190010 (2019).
9. S. A. Rice, *Quantitative Methods in Politics* (Alfred A. Knopf, New York, 1928).
10. A. Smith, *The Theory of Moral Sentiments* (Penguin, New York, 2010).
11. M. Macy, R. Willer, From factors to actors: Computational sociology and agent-based modeling. *Annu. Rev. Sociol.* **28**, 143–166 (2002).

12. R. Axelrod, *The Complexity of Cooperation: Agent-Based Models of Competition and Collaboration* (Princeton University Press, Princeton, NJ, 1997).
13. T. Schelling, Dynamic models of segregation. *J. Math. Sociol.* **1**, 143–186 (1971).
14. R. Axelrod, The dissemination of culture: A model with local convergence and global polarization. *J. Conflict Resolut.* **41**, 203–226 (1997).
15. G. Deffuant, F. Amblard, G. Weisbuch, T. Faure, How can extremism prevail? A study based on the relative agreement interaction model. *J. Artif. Soc. Soc. Simul.* **5**, 1 (2002).
16. R. Hegselmann, U. Krause, Opinion dynamics and bounded confidence models, analysis and simulation. *J. Artif. Soc. Soc. Simul.* **5**, 2 (2002).
17. A. Flache, M. Macy, Local convergence and global diversity: From interpersonal to social influence. *J. Conflict Resolut.* **55**, 970–995 (2007).
18. M. W. Macy, J. A. Kitts, A. Flache, S. Benard, "Polarization in dynamic networks: A hopfield model of emergent structure" in *Dynamic Social Network Modelling and Analysis*, R. Brieger, K. Carley, P. Pattison, Eds. (The National Academies Press, Washington, DC, 2003), pp. 162–173.
19. N. P. Mark, Culture and competition: Homophily and distancing explanations for cultural niches. *Am. Sociol. Rev.* **68**, 319–345 (2003).
20. D. Baldassarri, P. Bearman, Dynamics of political polarization. *Am. Sociol. Rev.* **72**, 784–811 (2007).
21. K. Klemm, V. M. Eguíluz, R. Toral, M. S. Miguel, Global culture: A noise-induced transition in finite systems. *Phys. Rev. E Stat. Nonlin. Soft Matter Phys.* **67**, 045101 (2003).
22. S. A. Condie, C. M. Condie, Stochastic events can explain sustained clustering and polarisation of opinions in social networks. *Sci. Rep.* **11**, 1355 (2021).
23. M. J. van Strien, S. H. Huber, J. M. Anderies, AGrêt-Regamey, Resilience in social-ecological systems: Identifying stable and unstable equilibria with agent-based models. *Ecol. Soc.* **24**, 8 (2019).
24. J. M. Brändle, G. Langendijk, S. Peter, S. H. Brunner, R. Huber, Sensitivity analysis of a land-use change model with and without agents to assess land abandonment and long-term re-forestation in a swiss mountain region. *Land (Basel)* **4**, 475–512 (2015).
25. R. Dahl, *Pluralist Democracy in the United States* (Rand McNally, Chicago, IL, 1967).
26. E. C. Ladd, C. D. Hadley, *Transformations of the American Party System* (Norton, New York, 1978).
27. K. Arceneaux, M. Johnson, "More a symptom than a cause: Polarization and partisan news media in America" in *American Gridlock: The Sources, Character, and Impact of Political Polarization* (Cambridge University Press New York, NY, 2015), pp. 309–336.
28. J. C. Rocha, G. Peterson, Ö. Bodin, S. Levin, Cascading regime shifts within and across scales. *Science* **362**, 1379–1383 (2018).
29. M. Ma, Polarization experiment simulation. GitHub. https://github.com/mongooma/polarizationTipping. Deposited 25 July 2021.