PNAS RESEARCH ARTICLE | SOCIAL SCIENCES

How digital media drive affective polarization through partisan sorting

Petter Törnberg[a,b,1]

Edited by Helen Milner, Princeton University, Princeton, NJ; received April 25, 2022; accepted August 30, 2022

Politics has in recent decades entered an era of intense polarization. Explanations have implicated digital media, with the so-called echo chamber remaining a dominant causal hypothesis despite growing challenge by empirical evidence. This paper suggests that this mounting evidence provides not only reason to reject the echo chamber hypothesis but also the foundation for an alternative causal mechanism. To propose such a mechanism, the paper draws on the literatures on affective polarization, digital media, and opinion dynamics. From the affective polarization literature, we follow the move from seeing polarization as diverging issue positions to rooted in sorting: an alignment of differences which is effectively dividing the electorate into two increasingly homogeneous megaparties. To explain the rise in sorting, the paper draws on opinion dynamics and digital media research to present a model which essentially turns the echo chamber on its head: it is not isolation from opposing views that drives polarization but precisely the fact that digital media bring us to interact outside our local bubble. When individuals interact locally, the outcome is a stable plural patchwork of cross-cutting conflicts. By encouraging nonlocal interaction, digital media drive an alignment of conflicts along partisan lines, thus effacing the counterbalancing effects of local heterogeneity. The result is polarization, even if individual interaction leads to convergence. The model thus suggests that digital media polarize through partisan sorting, creating a maelstrom in which more and more identities, beliefs, and cultural preferences become drawn into an all-encompassing societal division.

polarization | sorting | social cohesion | agent-based modeling | opinion dynamics

Significance

Recent years have seen a rapid rise of affective polarization, characterized by intense negative feelings between partisan groups. This represents a severe societal risk, threatening democratic institutions and constituting a metacrisis, reducing our capacity to respond to pressing societal challenges such as climate change, pandemics, or rising inequality. This paper provides a causal mechanism to explain this rise in polarization, by identifying how digital media may drive a sorting of differences, which has been linked to a breakdown of social cohesion and rising affective polarization. By outlining a potential causal link between digital media and affective polarization, the paper suggests ways of designing digital media so as to reduce their negative consequences.

According to a recent study, nearly half of Americans expect a civil war in the coming few years, and one in five now believes that political violence is justified (1). Such numbers are a stark expression of the unprecedented levels of political polarization currently facing not only the United States but many countries around the world. The threat that such polarization poses is not only gridlocked policymaking and loss of trust in democratic institutions but was in the American context perhaps most viscerally illustrated in the January 6 storming of the US Capitol.

Scholarship seeking to explain the rising polarization has centrally implicated the digitalization of media and communication systems (2–4). However, while studies have identified a link between digital media and rising polarization (2, 5, 6), the causal mechanism at play has been subject to significant debate (7). "Selective exposure" has long been a dominant hypothesis, suggesting that polarization on digital media is driven by individuals isolating themselves into so-called "echo chambers"—homogeneous clusters protected from opposing individuals and perspectives—which are said to lead to the divergence of opinions toward more extreme positions (8–13). However, empirical evidence has been accumulating against this hypothesis, leading some researchers to suggest that the echo chamber may be an intellectual cul-de-sac (14), instead calling for alternative explanations. To propose an alternative causal mechanism for a possible link between digital media and rising polarization, this paper connects the adjacent literatures on affective polarization, digital media, and opinion dynamics.

From the literature on affective polarization, we draw a new understanding of the nature of polarization. The concept of affective polarization stems from the observation that opposing partisans have grown to "dislike, even loathe" each other (15). To explain this rise of partisan animosity, scholars have pointed to partisanship emerging as an important social identity under which societal divisions and conflicts are coming to align—a process we refer to as sorting (15–18). The American electorate appears to be coalescing into two increasingly homogeneous parties (18, 19), which are absorbing a growing number of political and ideological divisions, possibly even extending to leisure activities, consumption, aesthetic taste, and personal morality—expanding politics into a broader "culture war" (20–22). Politics is thus spreading to a larger number of political issues, while the dimensionality of the political issue space is simultaneously decreasing, as partisanship

Author affiliations: [a]Amsterdam Institute for Social Science Research, University of Amsterdam, 1018 WV, Amsterdam, The Netherlands; and [b]Institute of Geography, University of Neuchâtel, 2000, Neuchâtel, Switzerland

Author contributions: P.T. wrote the paper.

The author declares no competing interest.

This article is a PNAS Direct Submission.

[1]Email: p.tornberg@uva.nl.

This article contains supporting information online at http://www.pnas.org/lookup/suppl/doi:10.1073/pnas.2207159119/-/DCSupplemental.

Published October 10, 2022.

induces party-based sorting which makes individuals' opinions so strongly correlated with their political ideology that there are, effectively, only one or two issue dimensions (23). The effects of such partisan alignment are described in the long tradition of political theory on social cohesion, which suggests that stable and cohesive society are characterized not by lack of conflict but by their plural conflicts balancing each other out (24). Social conflict is sustainable as long as there are multiple and nonoverlapping lines of disagreement: we may differ on our views on one issue but agree on another; we may vote differently, but if we support the same football team or go to the same church, there remains space for interpersonal respect. The recent rise in polarization is thus expressive of a gradual breakdown of this cohesive glue, driven by a gradual alignment of social, economic, geographic, and ideological differences and conflicts (19). According to these scholars, sorting is thus central to understanding the rise in polarization and the growing sense of difference, distrust, and disdain between opposing partisans, suggesting that the causal link between digital media and polarization runs not through divergence of opinions but rather via the dynamics of partisan sorting (19).

From the literature on digital media, we will draw from the accumulating empirical evidence against the echo chamber hypothesis to form the empirical foundation for an alternative causal mechanism. This literature suggests that digital media do not isolate us from opposing ideas; au contraire, they bring us to interact with individuals outside our local bubble, and they throw us into a political war, in which we are forced to take sides (25–27). As we will see, such a shift in interaction patterns lies at the core of the rise of partisan sorting.

From the literature on opinion dynamics, we draw tools and methods needed for studying the relationship between digital media and polarization: these systems form a complex whole, riddled with feedback effects and emergent dynamics of the type that have previously been the purview of the hard sciences of complexity theory and physics (28–32). Opinion dynamics provides a wealth of models which examine the feedback effects and dynamics through which shifts in interaction structures can bring about unexpected outcomes, enabling us to show how the observed shift in interaction brought about by digital media can drive a rise in partisan sorting through an emergent causal mechanism.

Combining insights from these literatures, the paper thus presents a computational model that isolates an emergent mechanism through which digital media may drive partisan sorting. The suggested mechanism essentially turns the echo chamber on its head: it is not that isolation drives divergence of opinions but that the diverse and nonlocal interactions of digital media drive plural conflicts to align along partisan lines. By connecting individuals outside their local networks, digital media drive a global alignment of conflicts by effacing the counterbalancing effects of local cultures. The model thus suggests that digital media can intensify affective polarization by contributing to a runaway process in which more and more issues become drawn into a single growing social and cultural divide, in turn driving a breakdown of social cohesion.

We turn now to outlining, in turn, the three adjacent literatures on which the model draws.

Affective Polarization and the Alignment of Difference

While the political divide has remained relatively stable in the United States in terms of issue positions (33), many other indicators point to a significant increase in partisan polarization and animosity in the mass public (18, 34): Democrats and Republicans are growing increasingly cold to each other, coming to dislike or even disdain their partisan opponents (15). The mismatch between stable levels of issue disagreement and a growing emotional sense of difference has brought some political scientists to shift their understanding of polarization to highlight the role of emotion and identity, captured under notions such as affective (15), sectarian (5), or social polarization (18). Political scientists tend to measure affective polarization as the gap between individuals' positive feelings toward their own political party and negative feelings toward the opposing party. Many scholars link these negative feelings to shifts in identity, with partisanship emerging as an important social identity under which many societal divisions and conflicts are coming to align (15–18). These scholars point to a long political science tradition which suggests that integrated societies are characterized not so much by an absence of conflict as by their conflicts being cross-cutting: particular individuals or social groups will be allies in some circumstances and opponents in others (e.g., refs. 24, 35–40). Social conflict is sustainable as long as there are multiple and nonoverlapping lines of disagreement, as the "segmental participation in a multiplicity of conflicts constitutes a balancing mechanism within the structure" (ref. 24, p. 154). When disagreements are cross-cutting, pluralistic disagreement can channel social conflict toward mutual tolerance (41), thus maintaining social cohesion (42) and political forbearance (43). Correspondingly, when cleavages and conflicts come to align—what we will refer to as sorting—the effects are synergetic in terms of intensifying prejudice and conflict between opposing political groups (18, 44). According to scholars such as Mason (18), sorting results in the broadening and deepening in society of a sense of fundamental difference and a mutual questioning—or even denial—of the other side's legitimacy. This is not merely a matter of courteous political discourse: scholars have found that civil war is 12 times less probable in societies where ethnicity is cross-cut by another social identity such as class, geography, or religion (ref. 45; see also refs. 46, 47).

Such a process of sorting lies at the core of the recent rise in affective polarization in the United States, according to these scholars. While some decades ago, the United States was characterized by social cohesion enabled by cross-cutting social divisions over party, ideology, religion, class, race, and geography, these identities have since started moving into alignment (19, 48, 49). The American electorate has thus come to be organized in two increasingly homogeneous parties, with a variety of social, cultural, economic, geographic, and ideological cleavages falling in line with the partisan divide (19, 50, 51). Mason (18) argues that the result is two megaparties, with each party coming to represent not only policy positions but also a growing list of preferences and identities. Partisanship comes to absorb "otherwise unrelated divisions, emasculating cross-cutting cleavages, and dividing society and politics into two separate, opposing, and unyielding blocks" (ref. 52, p. 8) and turns the diverse social identities of a plural society into a singular megaidentity divide (48).

While polarization has traditionally been understood as divergence in issue position, this literature thus shifts our understanding by describing polarization as linked to a process in which the political divide coming to encompass more and more issues, preferences, and identities (53). DellaPosta (21) describes this process as an "oil spill" model of polarization: it is not that partisan positions have become more radical but rather that partisanship has

become more encompassing in terms of political positions. Some suggest that this oil spill is spreading even to issues that would appear unrelated to politics (21, 54): popular accounts of lifestyle politics and culture wars imply that the growing political and ideological divisions extend also to leisure activities, consumption, and aesthetic taste (55–59). Where we live, what car we drive, the color of our skin, and what sports we watch all come to speak to a common identity in "a process whereby the normal multiplicity of differences in the society increasingly align along a single dimension, cross-cutting differences become reinforcing, and people increasingly perceive and describe politics and society in terms of 'us' versus 'them'" (ref. 60, p. 18). The degree to which political divisions have come to dominate cultural life, however, remains subject to academic debate, with a recent study finding that while some lifestyle preferences are politically polarized, most remain shared across the political divide (54).

In summary, the literature on affective polarization brings a shift of understanding of polarization, with some scholars suggesting that the recent rise in partisan animosity may be linked to a process of sorting: conflicts and differences coming to align under partisanship, creating a growing societal division. This suggests the need to explain the observed rise in partisan sorting. Following the suggested implication of digital media, we will now turn to drawing on the empirical literature on the relationship between digital media and polarization to identify a potential mechanism driving the rise in sorting.

Media and Polarization

We will now turn to draw the foundations of an alternative hypothesis for the link between digital media and rising polarization from the accumulating evidence against the so-called selective exposure hypothesis. The selective exposure hypothesis seeks to explain polarization through the divergence of issue positions, starting from the suggestion that high-choice media environments enable individuals to avoid the discomfort of having their worldviews challenged by opposing opinions and perspectives (3, 61), choosing instead to self-segregate into homogeneous clusters—so-called echo chambers (8–13). The resulting lack of exposure to competing perspectives is, in turn, said to lead to more extreme issue positions as interacting with opposing viewpoints is seen as central for moderating opinions; as Sunstein (ref. 13, p. 4) puts it, "homogeneity can be breeding grounds for unjustified extremism."

However, while selective exposure and echo chambers remain dominant explanations for polarization and have been subject to significant study and modeling efforts (28), they have been increasingly questioned by empirical findings. Studies on social media have found limited evidence for the described homogeneous clusters on most social media platforms: while users may be unlikely to rebroadcast content from the opposing side, they are often exposed to it and may even respond (11, 62–69). Research using internet traffic data has even suggested that the online news audience is in fact less ideologically segregated than in-person interactions with family, friends, coworkers, and political discussants (70–72). This evidence has been taken to suggest that in contrast to the assumptions of the selective exposure hypothesis, social media is in fact characterized by intense interaction across the political divide (73). According to such findings, digital media does not appear to lock people into isolated echo chambers but rather intensifies interaction with diverse actors and ideas from outside one's local bubble.

Empirical findings have also challenged the second component of the selective exposure mechanism: the idea that homogeneous groups lead to extreme opinions, while interaction with opposing ideas or individuals leads to political moderation. For instance, Garrett et al. (74) found survey evidence suggesting that exposure to out-party news sources will in fact intensify polarization. Bail et al. (75) similarly found in a field experiment on Twitter that individuals who were exposed to opposing views became more polarized, not less (see also ref. 25). This lack of moderating effect from exposure to opposing viewpoints may be explained by work in psychology suggesting that our reception of a message is shaped by how we view the messenger: information received from those whom we dislike or view as different from us may carry little to no influence. In short, we seek to be more like those who are already like us. While there is substantial empirical evidence that this is the case, there is less agreement on what psychological mechanism drives this effect. Prior (3) suggests selective processing, which implies that our judgment of new information depends on our identities and interests, allowing us to counterargue or disregard opposing arguments or ideas through mechanisms such as confirmation bias, motivated reasoning, identity-protective cognition, or biased argument processing (76–79). Druckman and McGrath (80) instead suggest a Bayesian model in which the differences in outcome stem from differences in what individuals consider to be credible evidence (see also refs. 81, 82).

Regardless of the underlying mechanism at play, the field has gathered substantial evidence for observation that individuals tend be less influenced by ideas from the opposing side. Some earlier studies even suggested that messages from the opposing side can be rejected so strongly as to lead individuals to shift their position in the opposite direction—what is referred to as "negative influence" or the "backfire effect" (83). Empirical research has, however, found mixed evidence for the existence of such an effect, with many studies failing to reproduce it (84–89) and the initial research identifying the phenomenon having been criticized on methodological grounds (90–92). This has brought researchers to conclude that negative influence effect is elusive, with limited relevance for politics outside the laboratory (88, 90, 93–95): while we may be less influenced by our political opponents, we are unlikely to be negatively influenced by them.

That psychological mechanisms such as selective processing may be at play when we interact with opposing messages on digital media finds some support in empirical studies showing that while media consumption may be largely bipartisan, media trust is highly polarized (96, 97). In other words, partisans consume media from across the partisan divide but do not trust in what opposing media channels tell them. Research on social media data tells a similar story: while partisan social media users do tend to interact across the ideological divide (62), this interaction is not characterized by rational arguments and deliberation but tends to be contentious and conflictual, suggesting that users may not be engaging in good faith attempts at seeing things through other perspectives (98–100). While partisans interact and consume information from across the ideological aisle, this does not necessarily come with openness to new ideas and perspectives.

In summary, the empirical literature suggests that digitalization does not appear to lead to a reduction of interaction across political divide, but quite the opposite: it confronts us with diverse individuals, perspectives, and viewpoints, often in contentious ways. At the same time, such exposure does not seem to lead to political moderation: individuals are readily influenced by those

they view as already similar to themselves, while disregarding messages from those they view as the outgroup. The suggestion of this paper is that these two points of empirical evidence may not only be reason to reject the echo chamber hypothesis (26), but they also provide the foundation for an alternative hypothesis for media polarization. To propose such a mechanism, we will now turn to the opinion dynamics literature, which has long used a complex systems perspective to study how mass interaction can result in unexpected emergent mechanisms (101–103).

Sorting and Opinion Dynamics

Axelrod's (104) study of cultural dissemination presented an early and highly influential model within opinion dynamics, showing that cultural diversity can emerge even when individuals display microlevel convergence. In this model, agents with a number of attributes are interacting locally on a lattice grid, with the strength of influence between two agents being a function of the similarity across their attributes. The model shows that differentiated local cultures can emerge through only positive social influence, as individuals converge locally. While this model became highly influential in the opinion dynamics literature, it was seen as failing to explain polarization as the distance between the cultures remains fixed (92). Later studies showed that the averaging of opinions means that positions will never leave the initial range (105, 106). As Flache and Macy (107) show, the diversity of Axelrod's model is furthermore unstable, with even a small amount of noise leading to the model converging on a common culture. Based on this, they argue that to explain polarization, it is necessary to assume the existence of negative influence, i.e., that interaction between individuals may lead to divergence rather than convergence of opinions.

Accordingly, later polarization models brought in negative influence to explain polarization. Focusing on polarization as the divergence over single attributes, models found that negative influence can lead to bipolarization—that is, division into two divergent groups—even when there are initially no agents with extreme opinions (86, 108, 109). The individuals with the strongest initial views are pushed to further intensify their positions, gradually developing two opposing extremes which, in turn, push moderate individuals to over time adopt also extreme opinions. While most of these negative influence models focus on divergence of single opinions, DellaPosta et al. (58) focused on the sorting of cultural preferences, by adding negative social influence to an Axelrod-like model. The model did not seek to explain the shift in the level of political sorting but provided important clues to its dynamics by showing that network autocorrelation can result in lifestyle politics through a path-dependent process in which small and contingent demographic and socioeconomic differences amplify over time.

However, while modeling work found that microlevel negative influence can produce polarization, it became less plausible as an explanation as empirical evidence against the existence of negative influence accumulated (83–89). As research concluded that the effect is elusive and unlikely to be relevant outside the laboratory (88, 90, 93–95), this has brought a renewed focus in the modeling literature on the central puzzle of if and how polarization can take place without assuming the existence of negative influence (92, 110). Banisch and Olbrich (110) and Mäs and Flache (92) provide attempts at identifying such conditions, by assuming that individuals do not necessarily converge but can become more extreme when interacting with likeminded others. However, these models lean on similarly questioned assumptions of selective exposure—i.e., that individual primarily interact with likeminded others—and no model based on microconvergence has as of yet been proposed.

In summary, the task at hand is to propose a polarization model that does not lean on assumptions that have been questioned by empirical findings, such as negative influence, selective exposure, or interaction homophily. For the opinion dynamics literature, a central puzzle is whether and how microconvergence can lead to macropolarization. The suggestion of this paper in seeking to solve this puzzle is to follow the affective polarization literature's shift in focus from divergence to sorting. While convergence may not be capable of producing divergence, we will show that it can lead to sorting—i.e., to the type of partisan alignment which in the literature has been linked to a rise in affective polarization. We will now turn to developing a model which draws on the empirical observations on digital media to propose an emergent mechanism through which social media may drive affective polarization through partisan sorting.

Model Description

The model presented here builds on the traditions of Axelrod (104), DellaPosta et al. (58), and Mäs and Flache (92) but seeks to show how a shift in the structure of social interaction can lead to the sorting observed by the affective polarization literature. While the model can be applied on any graph structure, we will primarily focus on the structure used in Axelrod's (104) classic model, in which the model's w^2 agents, A, are located on a two-dimensional torus lattice of width and height w, with neighbors are defined by their Moore neighborhood (i.e., their eight adjacent and diagonal neighbors). As in Axelrod's (104) model, each agent i has a set of n dynamic attributes, D_i, that take on an integer between 1 and m, corresponding to opinions, lifestyle preferences, beliefs, attitudes, or behaviors that may spread through influence. Each agent i also has a single attribute, S_i, corresponding to partisan affiliation, which held fixed throughout the model run (see also ref. 58). S_i takes on an integer between 1 and k. For simplicity and in line with previous research, we use $k = 2$ to focus on bipolarization—that is, polarization between two groups. However, the model can easily be expanded to polypolarization (*SI Appendix* includes an examination of the dynamics with $k > 2$). We here treat partisan affiliation as fixed, assuming that it will change on a slower timescale than the other attributes. All dimensions are initially uniformly randomized.

In each step of the model, a uniformly random agent is chosen and asynchronously updated (111). The agent has a set of interlocutors, I, which are the local lattice Moore neighbors, with a fraction γ (in [0,1]) replaced by nodes sampled randomly from the entire network. The parameter γ thus represents a simple way of capturing the empirically observed effect of digital media: digital media means that individuals are brought into interaction with others outside their local network. While the selective exposure hypothesis would suggest that these individuals would be selected with a bias toward self-similarity, we will here let the choice be unbiased as we are seeking to examine whether sorting can appear without such empirically questioned assumptions. (Were such an effect to be added, it would likely intensify the observed effect; however, the exploration of this is left to future research.)

Following the idea that we are more influenced by those similar to us, the node is then influenced as a function of the similarity with its interlocutors. As in Axelrod's (104) model, the absolute similarity is defined as the fraction of attributes that

are shared between two given nodes, but in this model, the static attribute is weighted by a parameter c relative to the dynamic attributes (in the simulations below $c = 4$). The parameter c thus represents how much more important partisanship is relative to other attributes.

We first define $\delta^{S}_{ij,t} = \begin{cases} 1 & if\ S_i = S_j \\ 0 & otherwise \end{cases}$ and $\delta^{D}_{ij,l,t} = \begin{cases} 1 & if\ D_{i,l,t} = D_{j,l,t} \\ 0 & otherwise \end{cases}$, allowing us to define the absolute similarity $\delta_{ij,t}$ between node i and j at time t as

$$\delta_{ij,t} = \frac{c\delta^{S}_{ij} + \sum_{l\in 1,2,\ldots,n}\delta^{D}_{ij,l,t}}{c+n}.$$

Two observations should be made here. First, the importance of the fixed attribute relative to the dynamic attributes also depends on the number of dynamic attributes, as each attribute is weighed separately. Second, since there are only two possible values for the fixed attributes, while there are m possible values for the dynamic attributes, the probability that two nodes share a specific attribute is lower than the probability that they share their fixed attribute. This captures the intuition that while there are many possible cultural or political preferences, there tend to be fewer partisan groups around which these may align.

The strength of influence between two agents depends on how socially similar they are in comparison with the other interlocutors. The social similarity is in other words treated as relative: that is, how similar we see ourselves to a particular individual depends on the larger social context (for instance, while the differences between Trotskyists and Stalinists may appear as a large chasm for members of the Party, they may for outsiders appear as a matter of relatively trivial detail). One of the interactors is thus chosen using an urn model, as in DellaPosta et al. (58), with the relative similarity defined as (92)

$$w_{ij,t} = \frac{{\delta_{ij,t}}^{h}}{\sum_{l\in I}{\delta_{il,t}}^{h}}.$$

The parameter h thus determines the steepness of the relationship between absolute similarity and influence, i.e., the level of influence homophily: the higher the value of h, the less the agent is influenced by those deemed different. When influenced, the agent takes one of the dynamic attributes from the other agent for which the two have different values (again in line with ref. 104).

Drawing on the affective polarization literature, we here focus on level of sorting between the groups, denoted by ψ. We define sorting by the probability that a given attribute is shared between two individuals of the same groups, minus the probability that a given attribute is shared between two individuals of different groups. This captures the level of sorting on the group level: ψ will be highest (1.0) if the groups are internally completely homogeneous, while not sharing any attributes between the groups. If agents between the group share all attributes, ψ will be zero. The expected value of ψ, i.e., its value when all attributes are randomized, is 0. We can calculate ψ by first defining the fraction of shared attributes between two nodes i and j as

$$d_{ij,t} = \frac{\sum_{l\in 1,2,\ldots,n}\delta^{D}_{ij,t}}{\mathrm{n}}.$$

We then operationalize ψ by calculating the fraction of shared attributes for all agents with the same partisan belonging, minus the fraction of shared attributes for all agents with different partisan belonging. If we define the set of similarities between all same-party agents at time t $\mathrm{Q}_{same,\,t} = \{\mathrm{d}_{ij,t} | \mathrm{i,j} \in \mathrm{A};\ \mathrm{i} \neq \mathrm{j};\ S_i = S_j\}$ and the set of similarities between different-party agents $\mathrm{Q}_{diff,\,t} = \{\mathrm{d}_{ij,t} | \mathrm{i,j} \in \mathrm{A};\ S_i \neq S_j\}$, we get

$$\psi_t = \frac{\sum \mathrm{Q}_{same,\,t}}{|\mathrm{Q}_{same,\,t}|} - \frac{\sum \mathrm{Q}_{diff,\,t}}{|\mathrm{Q}_{diff,\,t}|}.$$

The model converges when either all nodes are the same or when the influence between nodes that are different is so low that no influence occurs, or when it does occur, it is likely to quickly be unmade by influence from another node. The probabilistic and dynamic nature of this convergence makes it challenging to prove analytically. Unless otherwise specified, the simulations here therefore use an empirically chosen metric for when convergence has occurred, verified on a large number of cases and with substantial margin: the simulation ends either when no update has occurred during 10 runs across the full number of agents or to prevent runs becoming infinitely stuck in a dynamic equilibrium, after 1,000,000 iterations.

Results

We focus on the question of how observed shifts in the patterns of social interaction resulting from digital media impact the level of partisan sorting, that is, how ψ varies as a function of γ. We begin by running the model for a fixed number of steps with three different values of γ to examine how ψ varies over time, using a high level of influence homophily ($h = 8$). Fig. 1 shows that the evolution of ψ indeed depends on γ. When γ is high, ψ quickly reaches a high level; when γ is 0.5, ψ rises more slowly and settles on a lower value; and with $\gamma = 0$, ψ remains fixed at a low level.

To examine the relationship between ψ and γ systematically, we carry out 8,000 simulation runs for three values of h (1, 2, 5), varying γ and examining the final value of ψ as the model reaches a stable state. Fig. 2 shows the result. At low levels of influence homophily ($h = 1$; Fig. 2*A*), the result is no partisan sorting, regardless of γ. At high levels of influence homophily

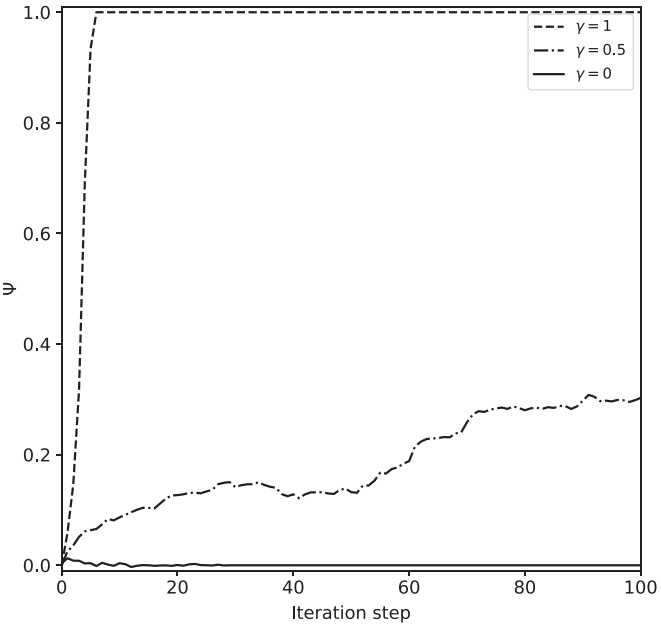

Fig. 1. Three example runs over time, for three parameters of γ. ψ is measured every 5,000 steps, over 500,000 time steps. Parameters are $h = 8$, $n = 10$, $|A| = 196$, $c = 4$, $m = 10$. As the figure shows, the run with $\gamma = 1$ quickly reaches a high ψ-value. The medium-γ population gradually increases to a ψ of 0.3. The run with completely local interaction remains near the expected ψ value of 0.

($h = 8$; Fig. 2*C*), the result is a rapid increase in sorting at a certain threshold of nonlocal interaction. This is a striking result, showing that if individuals are more influenced by those who are similar to them, interaction outside of local bubbles can result in increased partisan sorting, in turn associated to affective polarization.

When influence homophily is in between low and high ($h = 4$; Fig. 2*B*), the system exhibits bistability: for higher values of γ, the system can end up in either a nonsorted or highly sorted end state, depending on initial conditions and stochastic dynamics. This result can best be understood by considering that with 10 flexible attributes and 10 possible states, the probability that the two groups will share at least one attribute is around 65% (since $1 - \left(\frac{m-1}{m}\right)^n = 1 - \left(\frac{9}{10}\right)^{10} \approx 0.65$). This can also be seen in Fig. 2*C*: the level of sorting does not always reach 1, as the groups by chance come to share some attributes. As each shared attribute increases the strength of influence between the groups, the result is a certain probability of convergence which is a function of c and h. This suggests that if we do not assume the existence of negative influence, the diversity of cultural preferences—and thus the likelihood of two groups by chance sharing some preferences—will play a role in determining the level of social sorting.

Examining interaction in a two-dimensional torus lattice has the two benefits of capturing something of the spatiality of offline sociality—different locations in the graphs abstractly corresponding to geographical regions—and of allowing intuitive visual representation. However, it is in other ways a poor representation of real-world social networks, which tend to display highly uneven degree distributions, clustering, and social cliques. It is thus relevant to examine whether the dynamics of the models holds also for other interaction structures. Fig. 3 *A*–*C* shows the model applied to three common social network models: a random regular network, a scale-free network, and connected caveman network. As can be seen, the relationship between γ and ψ remains in all these structures: more nonlocal interaction tends to lead to increased partisan sorting. It is worth noting that highly clique-based networks, such as the connected caveman network (Fig. 3*C*), require higher levels of γ before partisan sorting rises. We now turn to investigating the underlying mechanism of these observed effects.

To understand the underlying cause for nonlocal interactions leading to sorting, we examine a set of simulations in more detail. The end result of a model run can be illustrated visually on a grid by treating the agents' dynamic attributes as a base-m number, thus allowing us to represent each configuration as a unique color. Two agents thus have the same color if they share the same dynamic attributes. Fig. 4 shows two runs on the opposite sides of the x axis of Fig. 2*C*: one with only local interaction ($\gamma = 0$) and one with completely nonlocal interaction ($\gamma = 1$). As Fig. 4 reveals, local interaction leads to a local convergence on two separate sets of dynamic attributes. This is similar to Axelrod's (104) finding that positive influence can lead to differentiation as agents converge locally. However, the central difference here is that the local convergence tends to occur within groups with shared static attributes. In other words, the local convergence occurs along partisan lines. This can be understood as the emergence of local political cultures, in which political affiliation is associated to a certain set of preferences and opinions only within a given geographical area. This outcome does not correspond to high levels of sorting between the parties, since the local convergence means that there is limited global sorting along the fixed attribute: partisan sorting remains low as the different local cultures counterbalance one another on the global arena.

When γ is high, however, the agents tend to converge globally on two separate sets of common attributes along the fixed attribute. This means that the system becomes highly sorted, with two different and internally coherent groups. The dynamic at play is in other words that increasing nonlocal interaction can produce a shift in the scale of sorting, bringing identities to align on the system level. This in turn leads to partisan sorting of identities, creating a large cultural cleavage between the groups. This furthermore suggests that the reason for networks with strong cliques being more resilient to sorting (Fig. 3*C*) is that the local political cultures emerging in those cliques are relatively stable, thus requiring higher levels of nonlocal interaction to align.

To capture these shifting relationships between the groups, we can carry out a dimension reduction of the distances between each individual in the model, thus allowing us to visually represent the relationship between each individual in the

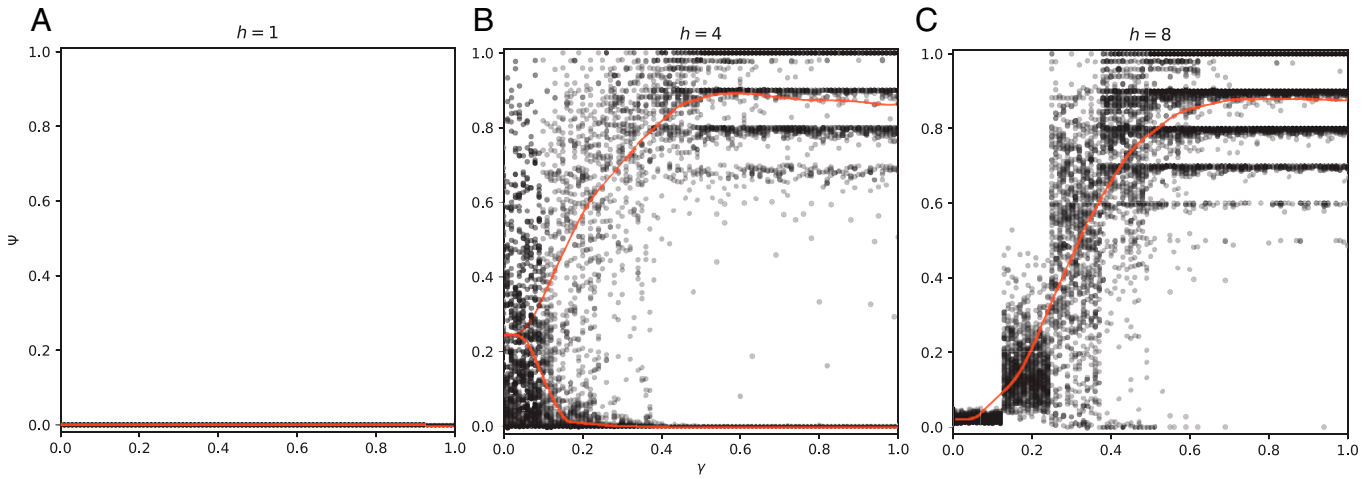

Fig. 2. The level of polarization (ψ) as a function of the level of nonlocal interaction (γ) when the level of influence homophily is (*A*) low ($h = 1$), (*B*) medium ($h = 4$), and (*C*) high ($h = 8$). Each dot represents the final ψ of a single run. The red line shows the LOWESS (Locally Weighted Scatterplot Smoothing) curves for the distributions. Parameters are $|A| = 196$, $c = 4$, $m = 10$, $n = 10$. As the figure shows, the level of polarization increases sharply when γ passes a certain threshold. This suggests that increasing the number of interactions with individuals outside the local network can lead to bipolarization, even given individual convergence on the microlevel. When influence homophily is in between high and low ($h = 4$), the system exhibits bistability: for higher values of γ, the system can end up in low-polarization or high-polarization states, depending on initial conditions and stochastic dynamics.

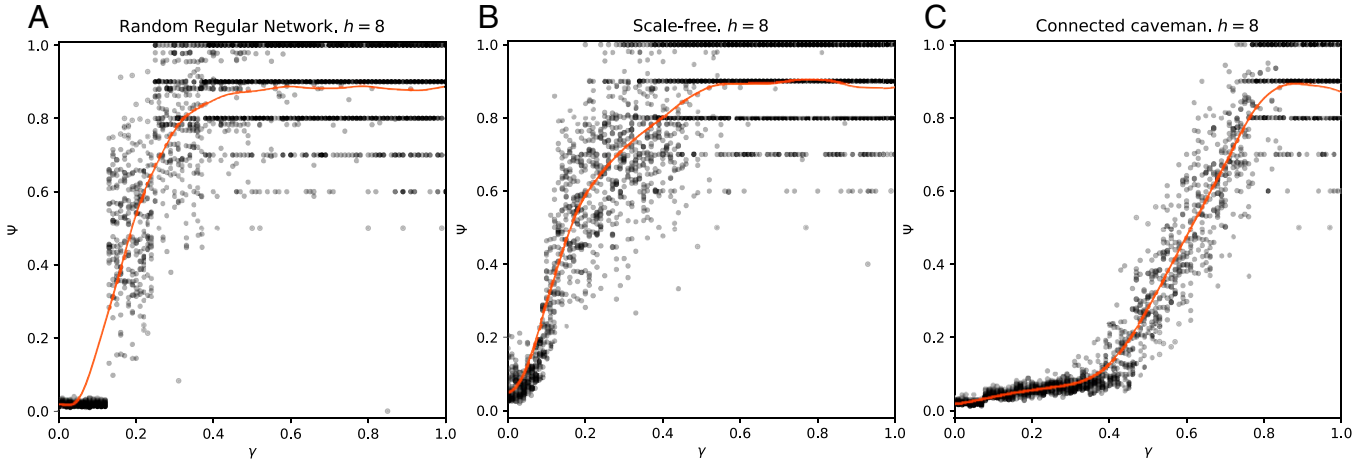

Fig. 3. The model applied on three common network structures, running with otherwise the same parameters as for Fig. 2*C*, with 1,600 runs over the parameter space. (*A*) A random regular graph, generated through the algorithm described in Steger and Wormald (112). Parameters used were 196 nodes and 8 edges. (*B*) A scale-free network, characterized by highly uneven degree distribution with some dominant nodes, generated with the algorithm described in ref. 113, based on Barabási–Albert growth model but adding that each random edge is followed by a chance of making an edge to one of its neighbors, thus capturing clustering. Parameters used were 196 nodes, 8 edges, and 0.01 probability of adding a triangle to any given edge. (*C*) The connected caveman, a highly cliqued graph structure used also in DellaPosta et al. (58), with 14 cliques of 14 nodes and 0.05 probability for reconnecting a given edge. As can be seen, strikingly, all network structures show the same general relationship between γ and ψ but with differences in the immediacy of the rise of sorting as a function of nonlocal interactions. In particular, the clique-dominated network in *C* requires a higher value of γ to become sorted.

groups over time. To do so, we apply principal component analysis (PCA) on the matrix of distances between all agents in each time step, thus reducing this matrix to a one-dimensional list. We then estimate the probability density function of this list, plotting the result on a ridgeline plot to illustrate the evolution of the distributions over time. Fig. 5 shows the result. Fig. 5*A* illustrates how local convergence leads to a broad and overlapping distributions of the groups, while Fig. 5*B* shows how nonlocal interaction leads to the two groups converging on separate points, expressing partisan sorting.

Fig. 6 shows the outcome when the level of homophily of influence is low ($h = 1$). As Fig. 6 illustrates, the result is that convergence does not occur along partisan lines. When nonlocal interaction is low ($\gamma = 0$), the local cultures converge on a common set of attributes, shared by both parties [thus essentially reproducing the dynamics of Axelrod's model (104)]. When γ is high, the entire system converges on the global optimum of complete convergence between all agents.

Discussion

The paper has provided a minimal model capturing an emergent mechanism through which digital media may drive affective polarization, not by isolating us in echo chambers that shield us from other viewpoints and positions but precisely by connecting us to views and positions outside our local bubble. An analysis of the model suggests that this ostensibly paradoxical effect is the result of the dynamics of network autocorrelation resulting from individuals being more influenced by those to whom they are already are similar, leading to convergence to

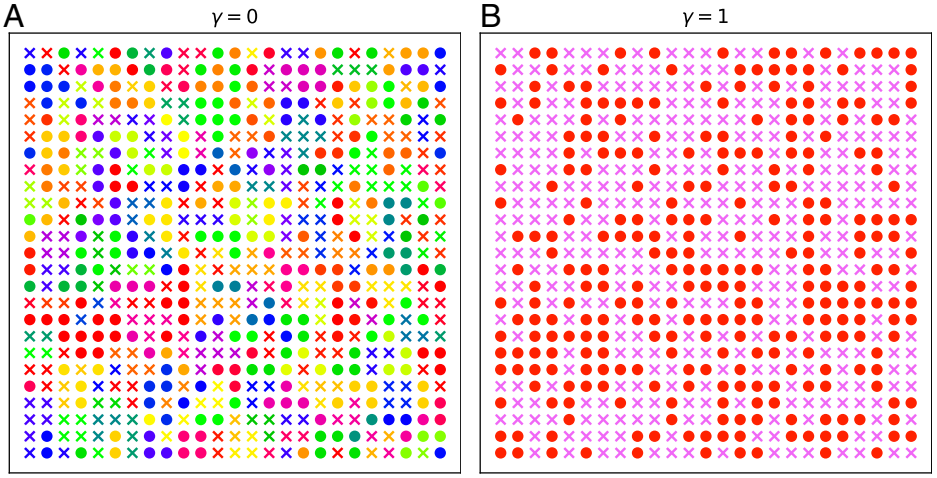

Fig. 4. The final state of two simulations at opposite ends of the *x* axis of Fig. 2*C*. (*A*) $\gamma = 0$ and (*B*) $\gamma = 1$. The agents are represented as crosses or circles depending on their static attribute. The colors are chosen by treating the list of dynamic attributes as a base-*m* number and then normalizing the result (i.e., $\sum_{l=1\ldots n} D_{i,l} m^{l-1}/m^n$), using this fraction to select a color from matplotlib's cm_prism. This allows us to represent each configuration of dynamic attributes as a unique color, with agents with different static attributes having the same color if they have the same dynamic attributes. Parameters are $|A| = 625$, $h = 8$, $c = 4$, $m = 10$, $n = 10$. As can be seen, when γ is low, the agents converge locally within their groups, resulting in a low level of global sorting as local differences cancel out. When γ is high, the groups converge internally, leading to high level of sorting.

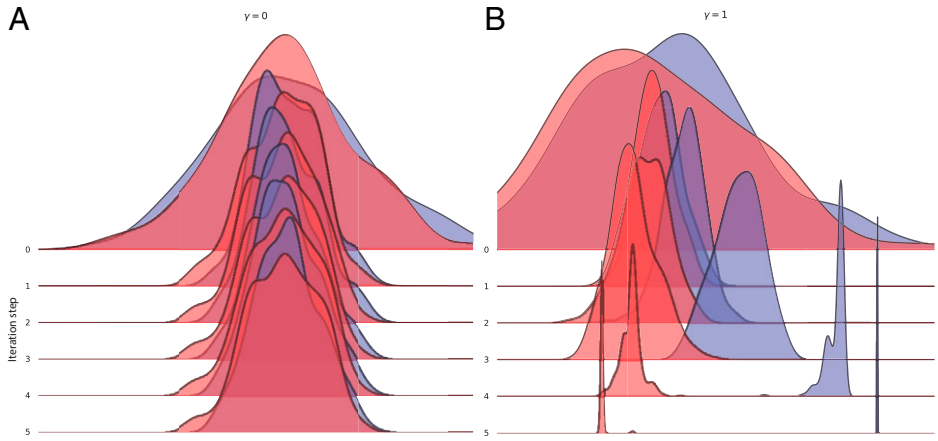

Fig. 5. (*A* and *B*) The same model runs as Fig. 4 *A* and *B*, respectively. The lines on the *y* axis represent time steps in the run, with each line representing 100,000 time steps. These ridgeline plots (also called joy plots) are a way of giving a sense of the distributions over time. The plots show the kernel density estimation of the probability density function by time step (*y* axis), applied on the one-dimensional PCA of the matrix of distances between each node in the model. The distributions of the two static groups are plotted separately as blue and red. The PCA reduces the dimensionality of the distances between each node, then representing them as a one-dimensional distribution. The figures reveal how the groups effectively bipolarize when the level of nonlocal interaction is high, while they separate into local cultures which counterbalance one another when nonlocal interaction is low.

shared cultures within partisan groups. If interaction takes place locally in geographical space or social networks, the process of sorting takes place locally, leading to local alignment of differences. This means that there will be limited sorting on the group level as the various local political cultures cancel each other out: some preferences are politicized in one region but not in others, and some are associated to one political one side in one region but the opposite side in another. The local diversity thus comes to function as a check on political polarization as politics is fractured into multiple local identities. When political cultures are internally diverse across space or across social groups, politics becomes rife with cross-cutting incentives, which leads to relatively high levels of social cohesion.

The rise of digital media, however, acts to destabilize this counterforce. By connecting individuals with others from outside their local social bubbles, digital media pressure local political cultures to align globally. Over time, the system comes to sort on the global scale, with a single political culture becoming systemwide. The effects are a dimensionality reduction, in which conflicts align along a single partisan divide (23). This means that geographical differences no longer act to counterbalance partisanship: local political cultures align, resulting in partisan sorting, in turn bringing stacked incentives, intensified political conflicts, and higher levels of affective polarization (48).

Such national-level partisan alignment of conflicts is particularly harmful in political systems such as that of the United States, whose constitution is founded on the assumption that geographical heterogeneity will counterbalance national-level partisanship. The US House and Senate were intended to represent not two parties but the nation's districts and states, allowing regional interests to moderate partisan excesses and leading Madison to refer to federalism as democracy's "double security." As Bednar (114) argues, such federalism can effectively provide a source for cross-cutting cleavages, thus functioning as a safeguard and counterweight to the national government. However, affective polarization can undermine this system as loyalties to the parties become stronger than to the state or region (115).

The dynamics of the model captures patterns identified in the empirical literature on the historic effects of the nationalization of politics in the United States (115, 116). According to this literature, partisan identity was for a long time cross-cut by geographical divides, meaning that a Mississippi Republican may have more in common with a Mississippi Democrat than with a Massachusetts Republican. The nationalization of politics brought the national alignment of political positions, resulting in the national party emerging as the main division, bringing about alignment of cleavages and conflicts along a single national line, thus intensifying polarization. This means that Republicans in Mississippi relate to their political wins and losses as Republicans rather than as Mississippi Republicans, thus dissipating the moderating effects and double security of federalism. The suggestion of the model is that the nonlocal interaction afforded by digital media may contribute to a corresponding process of deepening partisan alignment, thus further intensifying the sense of social distance between partisan groups. The historical parallel furthermore suggests that the dynamics identified by the model may be broader in scope than just the effects of the digitalization of media, capturing the effects of a broader range of processes which are bringing disparate groups into contact.

Conclusions

Political theorists have long argued that stable and cohesive society are characterized not by lack of conflict but by different conflicts balancing each other out; each conflict is cross-cut by others, creating a cohesive web of plural cleavages. According to recent literatures on affective polarization, the contemporary wave of polarization is expressive of a gradual tearing of this web, with conflicts coming to align along a single division (19, 48). As partisanship comes to encompass more and more political positions, values, and cultural preferences (21), the result is a form of polarization characterized by difference, distrust, and distain for one's political opponent—that is, affective polarization. However, while casting new light on the nature of polarization, this literature raises the central question of why this sorting process is taking place, with scholars pointing to the possible role of new media and communication technology.

In the media literature, the fundamental mechanism for explaining the impact of new media technology on politics has been selective exposure, captured in notions such as echo chambers or filter bubbles: media technologies are said to polarize by allowing us to isolate ourselves with likeminded others, thus avoiding the discomfort of being exposed to views and ideas from other groups. Since interacting with opposing viewpoints

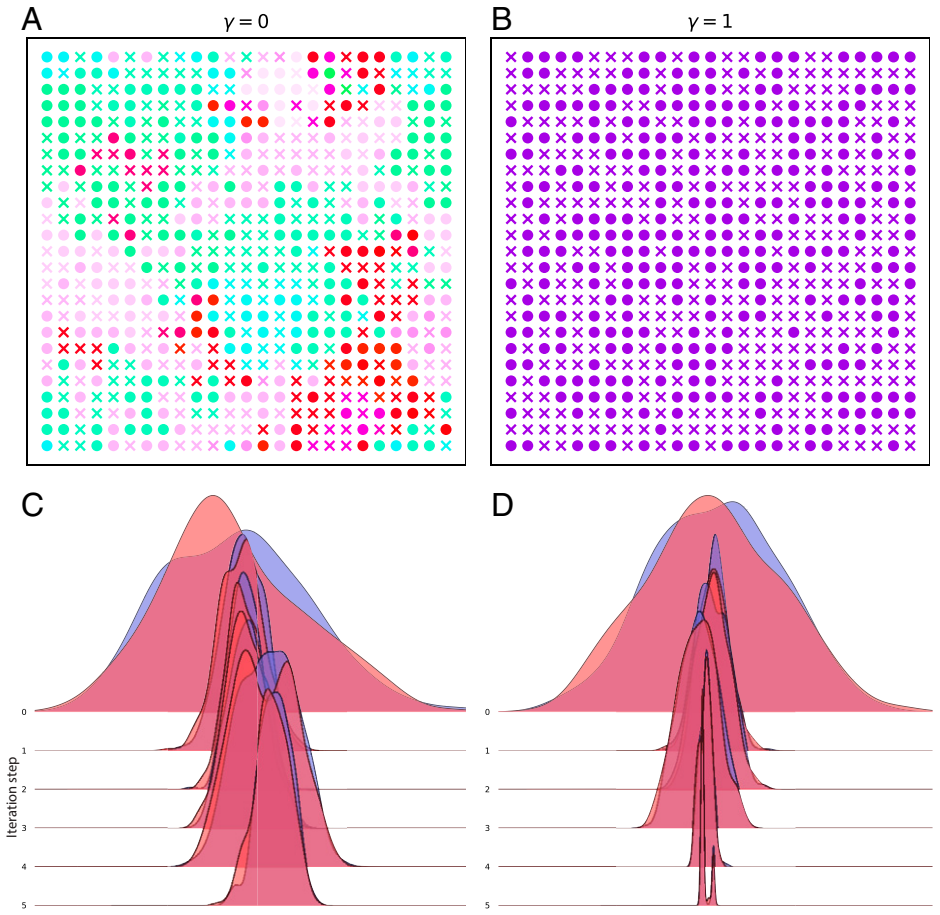

Fig. 6. Ridge plots illustrating the dynamics of the model with low levels of influence homophily ($h = 1$). (*A* and *B*) The final state of the two simulations and (*C* and *D*) their dynamics over time, with each line representing 100,000 time steps. *A* and *C* have low nonlocal interaction ($\gamma = 0$), and *B* and *D* have high nonlocal interaction ($\gamma = 1$). Parameters are $|A| = 625$, $n = 10$, $m = 10$, $h = 1$, $c = 4$. As can be seen, the runs without nonlocal interaction converge locally, forming local cultures with members of both groups. This is in line with Axelrod's original model of cultural diversity. The run with a high level of nonlocal interaction converges globally on a single set of attributes.

is thought to be central to moderating opinions, the result is said to be more extreme issue positions (13). However, this hypothesis has become increasingly questioned as two empirical findings question the two fundamental assumptions of this mechanism: first, results show that digital media is in fact characterized by substantial interaction across partisan lines (26); second, such interaction with opposing views has been shown to not necessarily reduce polarization, as psychological mechanisms allow individuals to disregard messages from individuals whom they deem different (75). These findings capture the intuitive observation that while digital media are rife with contentious debate, these rarely lead to individuals moderating their positions, let alone being convinced by opposing arguments. While we may interact and consume information from across the ideological divide, such bipartisan exchange is not necessarily an expression of good faith attempts at seeing things through another perspective. The suggestion of this paper has been that these two points of empirical evidence may not only be reason to reject the echo chamber hypothesis (26), but that they also provide the foundation for an alternative emergent causal mechanism underlying the polarizing effects of digital media.

To identify this alternative emergent mechanism, the paper drew on the opinion dynamics literature, which has long been focused on the complex feedback dynamics that link patterns of social interaction to emergence of polarization. For this literature, a long-standing puzzle has been if and how microlevel convergence can lead to macrolevel polarization. The paper brought to this literature the idea from the affective polarization literature that alignment of difference can produce a sense of social distance, thus shifting the focus from divergence to sorting in explaining the dynamics of polarization.

The paper presented a simple model which connects these three adjacent literatures, contributing to answering puzzles in each literature by combining their methods and theoretical insights. The model described an emergent causal mechanism through which more interaction outside one's local bubble can bring about the type of partisan sorting that has been linked to affective polarization. For the affective polarization literature, the model thus presents a causal mechanism for partisan sorting. For the media literature, the model showed that the combination between stronger influence between similar individuals and the empirically observed increase in interaction with others can provide an alternative to selective exposure for explaining the link between media and polarization. For the opinion dynamics literature, the model showed that if we follow the affective polarization literature's focus on sorting as a driver of affective polarization, polarization can occur even if individual interaction leads to convergence.

The model presented in this paper thus brings an important shift in how to think of the role of media in politics, by essentially turning the echo chamber hypothesis on its head: it is not lack of exposure to competing ideas that lead to polarization but precisely that digital media brings us to interact outside of our local bubble. When individuals interact in clusters, the

result tends to be local convergence, resulting in a stable plural patchwork of cross-cutting conflicts. However, when interaction takes place across space, the tendency is for groups to converge along the lines of partisan identity. The result is the crystallization of conflicting identities and the intensification of polarization, driven by a process in which sorting begets sorting and polarization begets polarization. These dynamics thus suggest a feedback loop between partisan sorting and affective polarization: sorting causes partisans to "dislike, even loathe" one another (15), in turn reducing their mutual social influence which further intensifies the process of sorting. Digital media may in this way disturb the balancing mechanism (24) of plural societies, by pushing conflicts and cleavages to align, creating a maelstrom in which additional identities, beliefs, and cultural belonging become sucked into a growing and all-encompassing societal division, which threatens the very foundation of social cohesion.

The model thus suggests rethinking digital media as not merely arenas for rational deliberation and political debate but as spaces for social identity formation and for symbolic displays of solidarity with allies and difference from outgroups (27). Digital media do not isolate us from opposing ideas; au contraire, they throw us into a national political war, in which we are forced to take sides. The suggestion is, in short, that polarization on digital media is driven by conflict rather than isolation (117), affording a form of politics rooted in identity rather than opinion (4). Digital media intensify polarization not as echo chambers but as a sorting machine, fueling a runaway social process that destabilizes plural societies by drawing more and more issues into a single expanding social and cultural divide. This suggests that the attempts of media platforms to reduce polarization by acting against echo chambers—algorithmically increasing exposure to opposing ideas—may backfire, instead resulting in intensified polarization and conflict.

Data, Materials, and Software Availability. The full Python code of the model has been deposited in GitHub, and is available from: https://github.com/cssmodels/tornberg2022pnas (118).

ACKNOWLEDGMENTS. I thank the three anonymous reviewers who provided invaluable comments on an earlier draft of the manuscript. I also acknowledge funding from the research program VENI, financed by the Dutch Research Council (NWO), project VI.Veni.201S.006.

1. G. J. Wintemute *et al.*, Views of American democracy and society and support for political violence: First report from a nationwide population-representative survey. medRxiv [Preprint] (2022). https://www.medrxiv.org/content/10.1101/2022.07.15.22277693 (Accessed 17 September 2022).
2. Y. Lelkes, G. Sood, S. Iyengar, The hostile audience: The effect of access to broadband internet on partisan affect. *Am. J. Pol. Sci.* **61**, 5–20 (2017).
3. M. Prior, Media and political polarization. *Annu. Rev. Polit. Sci.* **16**, 101–127 (2013).
4. P. Törnberg, C. Andersson, K. Lindgren, S. Banisch, Modeling the emergence of affective polarization in the social media society. *PLoS One* **16**, e0258259 (2021).
5. E. J. Finkel *et al.*, Political sectarianism in America. *Science* **370**, 533–536 (2020).
6. H. Allcott, L. Braghieri, S. Eichmeyer, M. Gentzkow, The welfare effects of social media. *Am. Econ. Rev.* **110**, 629–676 (2020).
7. J. Druckman, J. Levy, *Affective Polarization in the American Public* (Northwestern Institute for Policy Research, 2021).
8. M. Conover, J. Ratkiewicz, M. Francisco, Political polarization on twitter. *ICWSM* **133**, 89–96 (2011).
9. E. Dubois, G. Blank, The echo chamber is overstated: The moderating effect of political interest and diverse media. *Inform. Commun. Soc.* **21**, 729–745 (2018).
10. M. E. Del Valle, R. B. Bravo, Echo chambers in parliamentary Twitter networks: The Catalan case. *Int. J. Commun.* **12**, 21 (2018).
11. R. Karlsen, K. Steen-Johnsen, D. Wollebaek, B. Enjolras, Echo chamber and trench warfare dynamics in online debates. *Eur. J. Commun.* **32**, 257–273 (2017).
12. C. R. Sunstein, *Echo Chambers: Bush v. Gore, Impeachment, and Beyond* (Princeton University Press, Princeton, NJ, 2001).
13. C. R. Sunstein, The law of group polarization. University of Chicago Law School, John M. Olin Law & Economics Working Paper (1999). https://doi.org/10.1111/1467-9760.00148. Accessed 17 September 2022.
14. A. Bruns, *Are Filter Bubbles Real?* (John Wiley & Sons, 2019).
15. S. Iyengar, G. Sood, Y. Lelkes, Affect, not ideology: A social identity perspective on polarization. *Public Opin. Q.* **76**, 405–431 (2012).
16. S. Iyengar, S. J. Westwood, Fear and loathing across party lines: New evidence on group polarization. *Am. J. Pol. Sci.* **59**, 690–707 (2015).
17. L. Mason, "I disrespectfully agree": The differential effects of partisan sorting on social and issue polarization. *Am. J. Pol. Sci.* **59**, 128–145 (2015).
18. L. Mason, A cross-cutting calm: How social sorting drives affective polarization. *Public Opin. Q.* **80**, 351–377 (2016).
19. M. Levendusky, *The Partisan Sort: How Liberals Became Democrats and Conservatives Became Republicans* (University of Chicago Press, 2009).
20. M. Hetherington, J. Weiler, *Prius Or Pickup?: How the Answers to Four Simple Questions Explain America's Great Divide* (Houghton Mifflin, 2018).
21. D. DellaPosta, Pluralistic collapse: The "oil spill" model of mass opinion polarization. *Am. Sociol. Rev.* **85**, 507–536 (2020).
22. J. R. Brown, R. D. Enos, The measurement of partisan sorting for 180 million voters. *Nat. Hum. Behav.* **5**, 998–1008 (2021).
23. M. Kawakatsu, Y. Lelkes, S. A. Levin, C. E. Tarnita, Interindividual cooperation mediated by partisanship complicates Madison's cure for "mischiefs of faction". *Proc. Natl. Acad. Sci. U.S.A.* **118**, e2102148118 (2021).
24. L. Coser, *The Functions of Social Conflict* (The Free Press, New York, NY, 1956).
25. C. A. Bail, *Breaking the Social Media Prism: How to Make Our Platforms Less Polarizing* (Princeton University Press, 2021).
26. A. Guess, B. Nyhan, B. Lyons, J. Reifler, Avoiding the echo chamber about echo chambers. *Knight Foundation* **2**, 1–25 (2018).
27. P. Törnberg, J. Uitermark, Tweeting ourselves to death: The cultural logic of digital capitalism. *Media Cult. Soc.* **44**, 1–20 (2022).
28. P. Törnberg, Echo chambers and viral misinformation: Modeling fake news as complex contagion. *PLoS One* **13**, e0203958 (2018).
29. J. B. Bak-Coleman *et al.*, Stewardship of global collective behavior. *Proc. Natl. Acad. Sci. U.S.A.* **118**, e2025764118 (2021).
30. S. A. Levin, H. V. Milner, C. Perrings, The dynamics of political polarization. *Proc. Natl. Acad. Sci. U.S.A.* **118**, e2116950118 (2021).
31. N. E. Leonard, K. Lipsitz, A. Bizyaeva, A. Franci, Y. Lelkes, The nonlinear feedback dynamics of asymmetric political polarization. *Proc. Natl. Acad. Sci. U.S.A.* **118**, e2102149118 (2021).
32. C. K. Tokita, A. M. Guess, C. E. Tarnita, Polarized information ecosystems can reorganize social networks via information cascades. *Proc. Natl. Acad. Sci. U.S.A.* **118**, e2102147118 (2021).
33. M. P. Fiorina, *Unstable Majorities: Polarization, Party Sorting, and Political Stalemate* (Hoover Press, 2017).
34. Pew Research Center, Partisanship and Political Animosity in 2016: Highly Negative Views of the Opposing Party and Its Members (2016). https://www.pewresearch.org/politics/2016/06/22/partisanship-and-political-animosity-in-2016/. Accessed 17 September 2022.
35. D. B. Truman, The governmental process: Political interests and public opinion (Alfred A. Knopf, New York, NY, 1951).
36. R. A. Dahl, The behavioral approach in political science: Epitaph for a monument to a successful protest. *Am. Polit. Sci. Rev.* **55**, 763–772 (1961).
37. W. A. Galston, *Liberal Pluralism: The Implications of Value Pluralism for Political Theory and Practice* (Cambridge University Press, 2002).
38. P. Norris, *Driving Democracy: Do Power-Sharing Institutions Work* (Cambridge University Press, 2008).
39. S. M. Lipset, The value patterns of democracy: A case study in comparative analysis. *Am. Sociol. Rev.* **28**, 515–531 (1963).
40. S. M. Lipset, S. Rokkan, *Party Systems and Voter Alignments: Cross-National Perspectives* (Free Press, 1967).
41. S. A. Stouffer, *Communism, Conformity, and Civil Liberties: A Cross-Section of the Nation Speaks Its Mind* (Transaction Publishers, 1955).
42. E. A. Nordlinger, *Conflict Regulation in Divided Societies* (Center for International Affairs, Harvard University, 1972).
43. S. Levitsky, D. Ziblatt, *How Democracies Die* (Broadway Books, 2018).
44. S. Roccas, M. B. Brewer, Social identity complexity. *Pers. Soc. Psychol. Rev.* **6**, 88–106 (2002).
45. J. R. Gubler, J. S. Selway, Horizontal inequality, crosscutting cleavages, and civil war. *J. Conflict Resolut.* **56**, 206–232 (2012).
46. D. Siroky, M. Hechter, Ethnicity, class, and civil war: The role of hierarchy, segmentation, and cross-cutting cleavages. *Civ. Wars* **18**, 91–107 (2016).
47. J. S. Selway, Cross-cuttingness, cleavage structures and civil war onset. *Br. J. Polit. Sci.* **41**, 111–138 (2011).
48. L. Mason, *Uncivil Agreement: How Politics Became Our Identity* (University of Chicago Press, 2018).
49. A. Abramowitz, "Partisan polarization and the rise of the Tea Party movement" in *APSA 2011 Annual Meeting Paper*, A. Abramowitz, Ed. (APSA, 2011).
50. A. I. Abramowitz, K. L. Saunders, Ideological realignment in the US electorate. *J. Polit.* **60**, 634–652 (1998).
51. N. McCarty, K. T. Poole, H. Rosenthal, Polarized America: The *D*ance of *I*deology and *U*nequal *R*iches (MIT Press, 2016).
52. T. Carothers, A. O'Donohue, *Democracies Divided: The Global Challenge of Political Polarization* (Brookings Institution Press, 2019).
53. D. Baldassarri, A. Gelman, Partisans without constraint: Political polarization and trends in American public opinion. *AJS* **114**, 408–446 (2008).
54. S. Praet, A. M. Guess, J. A. Tucker, R. Bonneau, J. Nagler, What's not to like? Facebook page likes reveal limited polarization in lifestyle preferences. *Polit. Commun.* **39**, 311–338 (2022).
55. A. Giddens, *Modernity and Self-Identity: Self and Society in the Late Modern Age* (Stanford University Press, 1991).

56. T. Bennett *et al.*, *Culture, Class, Distinction* (Routledge, 2009).
57. E. Currid-Halkett, The sum of small things: A theory of the aspirational class (Princeton University Press, Princeton, NJ, 2017).
58. D. DellaPosta, Y. Shi, M. Macy, Why do liberals drink lattes? *AJS* **120**, 1473–1511 (2015).
59. W. G. Jacoby, Is there a culture war? Conflicting value structures in American public opinion. *Am. Polit. Sci. Rev.* **108**, 754–771 (2014).
60. J. McCoy, T. Rahman, M. Somer, Polarization and the global crisis of democracy: Common patterns, dynamics, and pernicious consequences for democratic polities. *Am. Behav. Sci.* **62**, 16–42 (2018).
61. R. K. Garrett, Echo chambers online?: Politically motivated selective exposure among Internet news users. *J. Comput. Mediat. Commun.* **14**, 265–285 (2009).
62. P. Barberá, J. T. Jost, J. Nagler, J. A. Tucker, R. Bonneau, Tweeting from left to right: Is online political communication more than an echo chamber? *Psychol. Sci.* **26**, 1531–1542 (2015).
63. S. Goel, W. Mason, D. J. Watts, Real and perceived attitude agreement in social networks. *J. Pers. Soc. Psychol.* **99**, 611–621 (2010).
64. E. Bakshy, S. Messing, L. A. Adamic, Exposure to ideologically diverse news and opinion on Facebook. *Science* **348**, 1130–1132 (2015).
65. E. Colleoni, A. Rozza, A. Arvidsson, Echo chamber or public sphere? Predicting political orientation and measuring political homophily in Twitter using big data. *J. Commun.* **64**, 317–332 (2014).
66. E. Lawrence, J. Sides, H. Farrell, Self-segregation or deliberation? Blog readership, participation, and polarization in American politics. *Perspect. Polit.* **8**, 141–157 (2010).
67. C. Vaccari *et al.*, Of echo chambers and contrarian clubs: Exposure to political disagreement among German and Italian users of Twitter. *Soc. Media Soc.* **2**, 1–24 (2016).
68. S. Yardi, D. Boyd, Dynamic debates: An analysis of group polarization over time on twitter. *Bull. Sci. Technol. Soc.* **30**, 316–327 (2010).
69. G. De Francisci Morales, C. Monti, M. Starnini, No echo in the chambers of political interactions on Reddit. *Sci. Rep.* **11**, 2818 (2021).
70. M. Gentzkow, J. M. Shapiro, Ideological segregation online and offline. *Q. J. Econ.* **126**, 1799–1839 (2011).
71. D. C. Mutz, P. S. Martin, Facilitating communication across lines of political difference: The role of mass media. *Am. Polit. Sci. Rev.* **95**, 97–114 (2001).
72. D. C. Mutz, J. J. Mondak, The workplace as a context for cross-cutting political discourse. *J. Polit.* **68**, 140–155 (2006).
73. A. Keuchenius, P. Törnberg, J. Uitermark, Why it is important to consider negative ties when studying polarized debates: A signed network analysis of a Dutch cultural controversy on Twitter. *PLoS One* **16**, e0256696 (2021).
74. R. K. Garrett *et al.*, Implications of pro-and counterattitudinal information exposure for affective polarization. *Hum. Commun. Res.* **40**, 309–332 (2014).
75. C. A. Bail *et al.*, Exposure to opposing views on social media can increase political polarization. *Proc. Natl. Acad. Sci. U.S.A.* **115**, 9216–9221 (2018).
76. C. S. Taber, M. Lodge, Motivated skepticism in the evaluation of political beliefs. *Am. J. Pol. Sci.* **50**, 755–769 (2006).
77. A. Corner, L. Whitmarsh, D. Xenias, Uncertainty, scepticism and attitudes towards climate change: Biased assimilation and attitude polarisation. *Clim. Change* **114**, 463–478 (2012).
78. C. S. Taber, D. Cann, S. Kucsova, The motivated processing of political arguments. *Polit. Behav.* **31**, 137–155 (2009).
79. D. M. Kahan, Misconceptions, misinformation, and the logic of identity-protective cognition (2017). https://doi.org/10.2139/ssrn.2973067. Accessed 17 September 2022.
80. J. N. Druckman, M. C. McGrath, The evidence for motivated reasoning in climate change preference formation. *Nat. Clim. Chang.* **9**, 111–119 (2019).
81. S. J. Hill, Learning together slowly: Bayesian learning about political facts. *J. Polit.* **79**, 1403–1418 (2017).
82. A. Coppock, *Persuasion in Parallel: How Information Changes Minds about Politics* (Chicago Studies in American Politics, University of Chicago Press, Chicago, 2021).
83. B. Nyhan, J. Reifler, When corrections fail: The persistence of political misperceptions. *Polit. Behav.* **32**, 303–330 (2010).
84. M. B. Brewer, The social self: On being the same and different at the same time. *Pers. Soc. Psychol. Bull.* **17**, 475–482 (1991).
85. M. W. Macy, J. A. Kitts, A. Flache, S. Benard, "Polarization in dynamic networks: A Hopfield model of emergent structure" in *Dynamic Social Network Modeling and Analysis*, R. Breiger, K. Carley, P. Pattison, Eds. (National Academies Press, Washington, DC, 2003), pp.162–173.
86. D. Baldassarri, P. Bearman, Dynamics of political polarization. *Am. Sociol. Rev.* **72**, 784–811 (2007).
87. D. M. Kahan *et al.*, The polarizing impact of science literacy and numeracy on perceived climate change risks. *Nat. Clim. Chang.* **2**, 732–735 (2012).
88. D. M. Kahan, "The politically motivated reasoning paradigm, part 1: What politically motivated reasoning is and how to measure it" in *Emerging Trends in the Social and Behavioral Sciences: An Interdisciplinary, Searchable, and Linkable Resource*, R. A. Scott, S. M. Kosslyn, Eds. (John Wiley & Sons, Hoboken, NJ, 2015), pp. 1–16.
89. A. Guess, A. Coppock, Does counter-attitudinal information cause backlash? Results from three large survey experiments. *Br. J. Polit. Sci.* **50**, 1497–1515 (2020).
90. T. Wood, E. Porter, The elusive backfire effect: Mass attitudes' steadfast factual adherence. *Polit. Behav.* **41**, 135–163 (2019).
91. Z. Krizan, R. S. Baron, Group polarization and choice-dilemmas: How important is self-categorization? *Eur. J. Soc. Psychol.* **37**, 191–201 (2007).
92. M. Mäs, A. Flache, Differentiation without distancing. Explaining bi-polarization of opinions without negative influence. *PLoS One* **8**, e74516 (2013).
93. R. S. Nickerson, Confirmation bias: A ubiquitous phenomenon in many guises. *Rev. Gen. Psychol.* **2**, 175 (1998).
94. D. M. Kahan, Ideology, motivated reasoning, and cognitive reflection: An experimental study. *Judgm. Decis. Mak.* **8**, 407–424 (2012).
95. K. Takács, A. Flache, M. Mäs, Discrepancy and disliking do not induce negative opinion shifts. *PLoS One* **11**, 1–21 (2016).
96. A. M. Guess, P. Barberá, S. Munzert, J. Yang, The consequences of online partisan media. *Proc. Natl. Acad. Sci. U.S.A.* **118**, e2013464118 (2021).
97. M. Jurkowitz, A. Mitchell, E. Shearer, M. Walker, US media polarization and the 2020 election: A nation divided (Pew Research Center, 2020). https://www.pewresearch.org/journalism/2020/01/24/u-s-media-polarization-and-the-2020-election-a-nation-divided/. Accessed 17 September 2022.
98. G. Evolvi, Hate in a tweet: Exploring internet-based islamophobic discourses. *Religions (Basel)* **9**, 307 (2018).
99. R. Moernaut, J. Mast, M. Temmerman, M. Broersma, Hot weather, hot topic. Polarization and sceptical framing in the climate debate on Twitter. Inform. Commun. *Soc.* **25**, 1047–1066 (2022).
100. A. Gruzd, J. Roy, Investigating political polarization on Twitter: A Canadian perspective. *Policy Internet* **6**, 28–45 (2014).
101. R. Hegselmann, U. Krause, Opinion dynamics and bounded confidence models, analysis, and simulation. *J. Artif. Soc. Soc. Simul.* **5**, 1–33 (2002).
102. R. P. Abelson, "Mathematical models of the distribution of attitudes under controversy" in *Contributions to Mathematical Psychology*, N. Fredericksen, H. Gullicksen, Eds. (Holt, Rinehart, and Winston, New York, NY, 1964), pp.142–165.
103. P. Bonacich, P. Lu, *Introduction to Mathematical Sociology* (Princeton University Press, 2012).
104. R. Axelrod, The dissemination of culture: A model with local convergence and global polarization. *J. Conflict Resolut.* **41**, 203–226 (1997).
105. N. E. Friedkin, E. C. Johnsen, Social influence and opinions. *J. Math. Sociol.* **15**, 193–206 (1990).
106. A. Flache, R. Torenvlied, When will they ever make up their minds? The social structure of unstable decision making. *J. Math. Sociol.* **28**, 171–196 (2004).
107. A. Flache, M. W. Macy, What sustains cultural diversity and what undermines it? Axelrod and beyond. arXiv [Preprint] (2006). https://arxiv.org/abs/physics/0604201 (Accessed 17 September 2022).
108. A. Flache, M. W. Macy, Small worlds and cultural polarization. *J. Math. Sociol.* **35**, 146–176 (2011).
109. N. P. Mark, Culture and competition: Homophily and distancing explanations for cultural niches. *Am. Sociol. Rev.* **68**, 319–345 (2003).
110. S. Banisch, E. Olbrich, Opinion polarization by learning from social feedback. *J. Math. Sociol.* **43**, 76–103 (2019).
111. J. M. Epstein, *Generative Social Science: Studies in Agent-Based Computational Modeling* (Princeton University Press, 2006).
112. A. Steger, N. C. Wormald, Generating random regular graphs quickly. *Combin. Probab. Comput.* **8**, 377–396 (1999).
113. P. Holme, B. J. Kim, Growing scale-free networks with tunable clustering. *Phys. Rev. E Stat. Nonlin. Soft Matter Phys.* **65**, 026107 (2002).
114. J. Bednar, Polarization, diversity, and democratic robustness. *Proc. Natl. Acad. Sci. U.S.A.* **118**, e2113843118 (2021).
115. E. Klein, *Why We're Polarized* (Simon and Schuster, 2020).
116. D. J. Hopkins, *The Increasingly United States: How and Why American Political Behavior Nationalized* (University of Chicago Press, 2018).
117. R. Collins, C-escalation and d-escalation: A theory of the time-dynamics of conflict. *Am. Sociol. Rev.* **77**, 1–20 (2012).
118. P. Törnberg, Model of the dynamics of partisan sorting. GitHub. https://github.com/cssmodels/tornberg2022pnas. Deposited 16 September 2022.