British Journal of Political Science (2023), **53**, 980–996
doi:10.1017/S0007123422000540

British Journal of Political Science

ARTICLE

Dynamics of Polarization: Affective Partisanship and Policy Divergence

Daniel Diermeier[1] and Christopher Li[2]*

[1]Department of Political Science and Owen School of Management, Vanderbilt University, Nashville, USA and [2]Department of Economics,Vanderbilt University, Nashville, USA
*Corresponding author. Email: christopher.m.li@vanderbilt.edu

(Received 21 March 2021; revised 2 June 2022; accepted 21 September 2022; first published online 2 December 2022)

Abstract

We explore the dynamics of affective partisanship and policy divergence in a behavioral voting model. Voters are adaptive and influenced by partisan affect, while political parties are rational and office motivated. We show that the affective partisanship of the electorate and the divergence of party platforms can be mutually reinforcing, thus providing an explanation for the observed co-movement of affective and elite polarization in recent decades. Whether the induced behavioral path exhibits low polarization or high polarization depends on the salience of group identity and the number of moderate voters. Thus, shocks to those factors, perhaps due to such events as economic crises or war, can lead to the polarization or depolarization of the electorate and of the elite.

Keywords: affective polarization; policy extremism; behavioral voting

The political elite in the United States have become more partisan over the last forty years (Barber and McCarty 2015; Layman, Carsey, and Horowitz 2006; McCarty, Poole, and Rosenthal 2016). This development has spurred extensive research and debate. One controversy is whether elite polarization is driven by the polarization of the electorate in terms of values and ideology. While there is some evidence in favor of this hypothesis (Abramowitz 2010; Abramowitz and Saunders 2008), others have argued that the electorate's issue preferences have remained fairly stable over the years (see Fiorina and Abrams 2008; Fiorina, Abrams, and Pope 2011; Hetherington 2009; Hill and Tausanovitch 2015; Levendusky 2009).

Traditionally, research on mass polarization focuses on the public's preferences on issues and policies. Yet, polarization can manifest itself directly in terms of the feelings and attitudes of one group toward another. Research in social psychology suggests that individuals have inherently positive feelings and attitudes toward members of their own social group, the "in-group," and negative feelings and attitudes toward members of the "out-group" (Tajfel 1970; Tajfel and Turner 1979). This is true even when group membership is defined via trivial characteristics or assigned randomly (Chen and Li 2009; Landa and Duell 2015).

In a political context, social groups are naturally defined along party lines, that is, the party that a person affiliates with defines the person's "in-group," while the other party defines the "out-group." In-group members include voters, politicians, public officials, or party staff—anyone who supports one particular party and not the other. According to this perspective, joining or supporting a party is less a rational decision and more a reflection of a person's group identity, which triggers emotions typical of in-group–out-group dynamics.

The role of affect and emotions in voting behavior was recognized as early as the 1950s. Classic studies by the Columbia and Michigan Schools found that voting behavior is often influenced by habit, social influence, and emotions rather than careful comparisons of issue positions (Berelson, Lazarsfeld, and McPhee 1954; Campbell et al. 1960).[1] These insights led to the idea of party identification, which has dominated the empirical study of voting ever since (see Bartels 2000; Campbell et al. 1960; Erikson, Mackuen, and Stimson 2002; Lewis-Beck et al. 2008; Miller, Shanks, and Shapiro 1996).[2]

Recent research provides ample evidence of rising affect-based partisanship, or *affective polarization*, over the past several decades in the United States (Iyengar and Krupenkin 2018; Iyengar, Sood, and Lelkes 2012; Iyengar et al. 2019).[3] Generally speaking, there has been an increase in animosity, distrust, and stereotyping between Republicans and Democrats. In 1960, for instance, less than 5 per cent of Republican and Democratic voters said they would be "displeased" if their children married across party lines. By 2010, the numbers had risen to 49 per cent for Republicans and 33 per cent for Democrats (Iyengar, Sood, and Lelkes 2012). A recent study by Abramowitz and Webster (2016) shows that partisan affect is a good predictor of partisan voting in US presidential and congressional elections.

Some notable studies have alluded to a potential connection between partisan affect and elite behavior. Using National Election Studies (NES) surveys and House budget votes, Coleman (1996) documents a strong correlation between elite polarization and mass partisanship. Bartels (2000) shows that US voters have demonstrated increasing party loyalty since the 1970s, coinciding with elite polarization.

Importantly, existing evidence indicates that affective polarization is distinct from the polarization of policy attitudes. For example, the correlation between policy preferences and measures of partisan affect is weak. Also, strong partisans with committed policy positions, that is, liberal Democrats and conservative Republicans, do not exhibit similar increases in interparty animus (Iyengar, Sood, and Lelkes 2012).

Given the empirical findings, it is natural to posit affective polarization as a potential driver of elite polarization. There is also the intriguing possibility of a causal effect in the opposite direction. In discussing the evidence of rising mass partisanship, Bartels (2000) proposes elite polarization as a potential cause. For example, candidates for political office may want to make the differences between the parties more salient to increase turnout. More recent work by Rogowski and Sutherland (2016) and Banda and Cluverius (2018) offers direct evidence that elite behavior can intensify voters' partisan affect.

Informal discussions notwithstanding, there is little systematic and rigorous exploration of the various threads suggested by the empirical evidence. One obstacle to such explorations is incompatible research traditions. While much of the empirical research is grounded in social psychology, the formal study of elections, beginning with Downs (1957), conceptualizes voters as rational decision makers who vote based on the evaluation of the "expected party differential." Notably, one of the robust findings of the Downsian model is the convergence of office-motivated

[1] Recent work suggests that these features do not simply rest on the lack of information or attention, shortcomings that may be alleviated by the use of heuristics like cue taking (Popkin 1994) or retrospective voting (Healy and Malhotra 2013), but are grounded in the psychological processes of opinion formation and candidate evaluation (Achen and Bartels 2017; Lodge and Taber 2013; Zaller et al. 1992).

[2] Various scholars have equated party identification with affect-based partisanship (see, for example, Burden and Klofstad 2005; Green, Palmquist, and Schickler 2004; Greene 2004). Indeed, this view can be traced back to the conceptualization of party identification in *The American Voter* (Campbell et al. 1960). Accounts of party identification based on group identity and emotional affiliations are not the only approaches. Alternative accounts of party identification include such concepts as "running tallies" (Fiorina 1981) and "macropartisanship" (MacKuen, Erikson, and Stimson 1989).

[3] Iyengar, Sood, and Lelkes (2012) use thermometer ratings of parties from the US National Election Studies as a measure of partisan affect and show that Democrats' and Republicans' ratings for their own parties have remained fairly stable but the ratings for the opposing parties have seen a 15-point drop (on a 100-point scale) since 1988. Abramowitz and Webster (2016) find similar patterns.

https://doi.org/10.1017/S0007123422000540 Published online by Cambridge University Press

candidates to the center.[4] However, formal theories based on the rational-choice paradigm are of limited use in exploring group identity and affective polarization.

In a recent approach, Diermeier and Li (2019) incorporated affective partisanship into a spatial-voting model and showed how affective polarization can induce policy divergence. However, their model is static and therefore cannot address questions about the dynamic interaction between mass partisanship and elite behavior.

In this article, we explore these questions by extending Diermeier and Li's (2019) framework to a dynamic context. Our model takes some basic building blocks from standard spatial-voting models. Voters have ideal points on a policy space, and parties are rational and office motivated: they choose policies to maximize their vote share. However, unlike in the traditional approach, voters act according to a behavioral heuristic that maps their partisan attachment and their experiences under implemented policies into voting behavior. Unlike in Downsian-type models, voters act in response to past experiences; they need not be aware of their own ideological positions, the parties' policy locations, or any other aspects of the model. It is worth noting that affective polarization is often presented in the literature as capturing how voters feel toward each other. Given the social identity conception of partisanship, it is reasonable to assume that the feelings toward co- and out-partisans apply equally to party elites, as they are prominent members of the in-/out-group. Several studies have adopted this broader interpretation of affective partisanship.[5]

The heuristic that guides voter behavior is derived from two principles established in the literature on group identity. The first principle—*in-group favoritism*—is the positive bias in the evaluation of behavior and characteristics of in-group members (see Mullen, Brown, and Smith 1992). The second principle—*in-group responsiveness*—is the observation that individuals tend to *be more sensitive* to the behavior and traits of in-group members than to those of out-group members (Frimer and Skitka 2020; Marques and Paez 1994; Marques, Yzerbyt, and Leyens 1988). One explanation for in-group responsiveness is that individuals apply stricter norms to in-group members and, therefore, judge more severely in-group members who violate such norms (Mackie and Cooper 1984; Pinto et al. 2010).

A key feature of the present model that sets it apart from Diermeier and Li's (2019) is that elite behavior can also influence voters' affective partisanship. This creates a bidirectional linkage between elite behavior and affective polarization, whereas Diermeier and Li (2019) explore the causal link in only one direction (that is, from affective polarization to elite behavior). The evolution of affective partisanship is modeled via a Markov process. Affective partisanship is the "state variable" and is updated each period based on the policy implemented by the incumbent. We assume that positive experience enhances a voter's affective association with the incumbent and that negative experience decreases it.[6] This postulate is congruent with the "law of effect"—"the most important principle in learning theory" (Hilgard and Bower 1966, 481).[7] The adaptive nature of partisanship allows for potential feedback between elite behavior and affective partisanship: politicians adjust policies in response to mass polarization and, at the same time, policies shape partisanship going forward.

We show how affective partisanship and elite behavior can reinforce each other. Importantly, qualitatively different behavioral paths may arise, depending on the electoral environment. If group identity is not salient or the number of moderate voters is large, then the behavioral path features low polarization, in which the parties choose centrist policies and affective partisanship moderates over time. On the other hand, if group identity is salient

[4] For a survey, see Duggan (2012).

[5] For an example, see Landa and Duell (2015).

[6] Rogowski and Sutherland (2016) show that voters' affective partisanship responds to elite policy positions. More generally, a long line of research in psychology suggests that an individual's affective state is a function of experience and stimuli (see, for example, Lazarus 1991; Lerner and Keltner 2000). This condition is consistent with the evidence that party identification responds to the performance and actions of incumbents (MacKuen, Erikson, and Stimson 1989).

[7] For experimental evidence, see Bendor (2010) and Woon (2012).

https://doi.org/10.1017/S0007123422000540 Published online by Cambridge University Press

and moderate voters are few, a high polarization path emerges, in which parties choose extreme policies and affective partisanship is high. Thus, our model can account for the observed correlation between elite polarization and mass partisanship. More important, an implication of these observations is that shocks to the electoral environment, possibly due to such events as economic crises and war, may lead to a shift from modest political polarization to heightened polarization, or vice versa.

Our article adds to the burgeoning body of formal work that incorporates behavioral agents. Our approach takes inspiration from the literature on adaptive learning (see Börgers and Sarin 1997; Hart 2005), which has seen increasing application in political science. Bendor, Diermeier, and Ting (2003) explore the "paradox of voting" in an adaptive-voting model. Diermeier and Li (2017) study electoral accountability with behavioral voters. Andonie and Diermeier (2019) examine multiparty elections. More closely related to this model, Bendor, Kumar, and Siegel (2010) examine how adaptive voters can develop partisanship toward candidates in the long run. However, the candidates in their model are not strategic; rather, their policy positions are fixed. Bendor et al. (2011) numerically analyze a model where both candidates and voters are adaptive.

While the adaptive-learning framework is useful in capturing a general lack of sophistication, other approaches focus on more specific behavioral biases. For example, several recent studies have explored the implications of biases in statistical reasoning, such as correlation neglect and motivated reasoning (Levy and Razin 2015; Little 2019; Minozzi 2013; Ortoleva and Snowberg 2015). Bisin, Lizzeri, and Yariv (2015) and Lizzeri and Yariv (2017) study government policy in the presence of time-inconsistent voters. Patty and Penn (2020) study the impact of identity and culture on individual behavior in organizations.

The Model

Two parties, A and B, compete in elections at the end of date $t = 1, 2, \ldots$. The winner of the election at date $t - 1$ becomes the incumbent at date t and chooses policy θ_t from the interval $[-1, 1]$. The objective of the incumbent is to maximize their vote share in the date t election. The parties are ex ante homogeneous; they do not differ in exogenous valence, nor are they policy motivated. It should be noted that this would rule out policy divergence in a typical Downsian setting (see Roemer 1994).

There is a unit continuum of voters, divided into three blocs according to their bliss point $b \in \{l, m, r\}$, where $l = -1$, $m = 0$, and $r = 1$. Let κ_b denote the measure of voters with bliss point b. We assume that $\kappa_l = \kappa_r$, that is, the distribution of bliss points is symmetric around the median. Voters with bliss point b will be referred to collectively as "b voters."

Voters are adaptive: their voting behavior responds to prior experiences with implemented policies. Such experiences, however, do not presuppose any knowledge or evaluation of policies. Rather, voters vote "how they feel." A voter's propensity to vote for a given candidate also depends on their affective partisanship, which is the emotional attachment, or *affinity*, to one of the two parties (Dias and Lelkes 2022). Formally, we denote a (generic) voter's affinity for the in-party at date t by p_t. We assume for tractability that affinity is binary: a voter either has high affinity for the in-party or for the out-party.[8] That is to say, p_t takes one of two values $\{h, l\}$, with h indicating high affinity for the in-party and l indicating high affinity for the out-party.[9] The probability that a voter votes for the incumbent at the date t election is a function of p_t and θ_t. For tractability, we focus on a linear functional form and explore a

[8] Adding additional gradients of affinity will complicate the analysis significantly without adding meaningful insights.

[9] This allows one to think of polarization in terms of negative partisanship (Abramowitz and Webster 2016), as high affinity for one party is equivalent to low affinity for the other party.

https://doi.org/10.1017/S0007123422000540 Published online by Cambridge University Press

more general setting in Appendix 1. Specifically, the probability of a voter with bliss point b voting for the incumbent is assumed to be:

$$\alpha(p_t, \theta_t) = -\tau(p_t)|\theta_t - b| + 2\tau(p_t), \tag{1}$$

where τ determines both the level of α and its sensitivity to the distance between the policy and the voter's bliss point. Letting $\tau_h \equiv \tau(h)$ and $\tau_l \equiv \tau(l)$, we make the following assumption (see Figure 1).

Assumption 1: $1/2 > \tau_h > \tau_l > 0$.

The fact that τ_h and τ_l are positive means that a voter is more likely to vote for the incumbent the more congruent the adopted policy is with their bliss point. This is reminiscent of standard Downsian preferences, but, importantly, we *do not* presume that voters conduct any rational evaluation of policy platforms. Rather, voters are more likely to have a positive experience when the policy is closer to their ideal point, and they respond positively to such experience. They may not be conscious of the underlying mechanism that links policy with experience, or even the proximity of policy to their own bliss points (Joesten and Stone 2014). It should be noted that a consequence of linearity is the symmetry in a voter's responses to deviations of policy toward and away from their bliss point.[10] This is not essential to our result. The generalization of the model in Appendix 1 allows for concavity or convexity of α with respect to policy. The restriction that τ_h and τ_l are less than 1/2, together with the functional form of α, ensures that α remains in the interior of [0, 1]. One may adopt a more general piecewise linear functional form for α that relaxes this restriction and still obtain qualitatively similar insights.

The assumption $\tau_h > \tau_l$ plays a key role in the model, and its implications are twofold. First, it implies that $\alpha(h, \theta) > \alpha(l, \theta)$, meaning that a voter is more likely to vote for the candidate they have high affinity for. This is an instance of *in-group favoritism*, a well-established finding in social psychology (Mullen, Brown, and Smith 1992). Jessee (2010) documents in-group favoritism in the 2008 US presidential election. Specifically, Democratic and Republican partisans "show a strong tendency to vote for their party's candidate even in situations in which they are ideologically closer to the other candidate" (Jessee 2010, 328) Secondly, the assumption implies that for $\theta \neq b$, $\tau_h = |(\partial\alpha(h, \theta)/\partial\theta)| > |(\partial\alpha(l, \theta)/\partial\theta)| = \tau_l$. In other words, a change in the incumbent's policy stance has a bigger impact on the propensity of its "co-partisans" (voters with high affinity) than on that of "out-partisans" (voters with low affinity). This is an example of *in-group responsiveness* as identified in various research in social psychology (Biernat, Vescio, and Billings 1999; Marques, Yzerbyt, and Leyens 1988; Mendoza, Lane, and Amodio 2014). In particular, studies have shown that people tend to punish in-group members more severely than out-group members for norm-violating behavior.[11] Here, deviations from the "party" line are similar to norm violations. We interpret the difference $\tau_h - \tau_l$, which determines how sensitive voter behavior is to affective partisanship, as measuring how salient group identity is.

Consistent with in-group responsiveness, affectively partisan voters tend to be more zealous, that is, they exhibit stronger emotional responses to their own party's actions, such as its policy positions. The Tea Party movement is a good example. While Tea Party supporters on the whole hold a similar ideology as average Republicans (Newport 2010), the former tend to be more impassioned (Kimball, Anthony, and Chance 2018). Accordingly, Tea Party supporters are keener to hold their elected representatives accountable for holding conservative policy positions.

[10] Specifically, fix affinity, the change in the probability of voting for the incumbent is the same whether the incumbent moves the policy ϵ closer to or away from the voter's bliss point.

[11] One explanation is that deviant behavior from in-group members is perceived as a betrayal and a greater threat to the positive group image and, more generally, the sense of identity (Akerlof and Kranton 2000; Castano et al. 2002).

https://doi.org/10.1017/S0007123422000540 Published online by Cambridge University Press

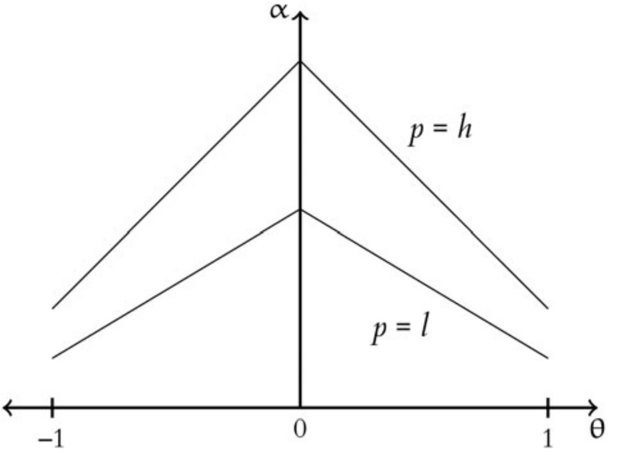

Fig. 1. An example of $\alpha(p, \theta)$ that satisfies Assumption 1.

Republican congresspeople reportedly felt immense pressure from the Tea Party to remain uncompromising politically, as "any deviation from the conservative line is met with a flood of phone calls and a credible threat of a primary challenge" (Lee 2013).[12]

Such in-group responsiveness is important whenever deviations from a group's core identity are at issue, that is, when elected politicians appear to stray from the pure party line. On other types of behavior, in-group responsiveness is not important. For example, some empirical studies suggest that voters are more lenient toward co-partisan politicians for general misbehavior or lack of performance, such as corruption or poor economic performance (see, for example, Eggers et al. 2014; Kayser and Wlezien 2011). Such behavior is deplorable, and bad outcomes are unwelcome, but they do not threaten group identity. In such contexts, in-group favoritism applies but not in-group responsiveness, that is, members of one's own group are given the benefit of the doubt, while members of the out-group are evaluated harshly. In other words, on general misbehavior not tied to features that distinguish the two parties, members of one's own party are treated with leniency (in-group favoritism). On the other hand, deviations from the party line, for example, taking a different stance on highly polarized topics like immigration or abortion, are viewed as betrayals and punished severely (in-group responsiveness).

To complete the model, we now describe how affective partisanship evolves over time. Here, we follow the tradition of adaptive-voting models (see, for example, Bendor et al. 2011) and assume party affinity evolves according to a Markov process. Specifically, let I be the date t incumbent, then the probability of a generic voter having high affinity for I at date $t+1$ is $\alpha(p_t, \theta_t)$. In other words, $\alpha(p_t, \theta_t)$ captures both the probability that a voter votes for party I and the probability that the voter will have high affinity for I at date $t+1$ (see Figure 2). As an illustration of the sequence of events, suppose the Democratic Party is in power at time t and implements its platform. Voters may have positive or negative experiences during the election cycle, which is partially influenced by government policies but also depends on individual circumstances, such as their partisan leanings and other random events. When it comes time for an election at the end of date t, a voter considers their experience and votes to reelect the Democratic Party if their experience was positive and votes for the Republican party otherwise.

[12] In our model, Tea Party members are just as disaffected with Republican politicians for adopting more extreme positions compared to them as they are when Republican politicians adopt more moderate positions. This symmetry assumption can be relaxed without affecting our results.

https://doi.org/10.1017/S0007123422000540 Published online by Cambridge University Press

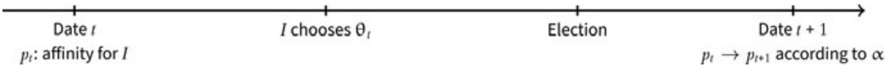

Fig. 2. The transition process of affective partisanship.

To put it more succinctly, voters consider the famous question posed by Ronald Reagan—“Are you better off than you were four years ago?”—and vote accordingly.

The intertemporal linkage between elite behavior and affective partisanship is established through the Markov process. Affective partisanship, as reflected in the distribution of party affinity, influences the incumbent’s policy choice. Specifically, the incumbent seeks to maximize vote share $\int\alpha(p_t, \theta_t)dF_t$, where F_t is the distribution of p_t.

At the same time, the incumbent’s policy choice shapes affective partisanship in the future. The fact that α is decreasing in δ_t means that a voter is more likely to develop high affinity for the incumbent the more the voter likes the incumbent’s policy. This represents the “law of effect.”

Analysis

To simplify notation, the time subscript is omitted whenever possible. Also, define $\delta_t = |\theta_t - b|$ as the proximity of policy to a voter’s bliss point; we will sometimes use δ_t as an argument in α instead of θ_t, that is, $\alpha(p_t, \delta_t) = -\tau(p_t)\delta_t + 2\tau(p_t)$. Let g_b denote the proportion of b voters that have high affinity for party A.[13] In a given period, the triple (g_l, g_m, g_r) fully describes the affective partisanship of the electorate. Given this, affective polarization is defined as follows:

Definition 1: The electorate is affectively polarized if $g_l > 1/2 > g_r$ or $g_l < 1/2 < g_r$.

In other words, the electorate is affectively polarized if, on average, l voters and r voters favor different parties. It should be noted that the notion of affective partisanship is independent of voters’ ideologies. The former may evolve over time in response to policies, while the ideological composition of the electorate is assumed to be fixed. By treating the two as distinct concepts, we can isolate the implications of affective partisanship and help explain the observation that partisanship has increased while attitudes on policy have largely been constant (Fiorina and Abrams 2008; Fiorina, Abrams, and Pope 2011; Hetherington 2009; Levendusky 2009). Without loss of generality, we assume that affective polarization takes the form of $g_l > 1/2 > g_r$, that is, l voters favor party A, while r voters favor party B. For the following discussion, we interpret the difference $g_l - g_r$ as the intensity of affective polarization at the societal level.

First, we establish a sufficient condition for the existence of a behavioral path in which affective partisanship is moderate and parties take on centrist positions. This “low polarization” path depends on the initial affective partisanship, the salience of group identity, and the size of moderate voters:

Proposition 1 (low polarization): If g_l and g_r in the initial period satisfy:

$$0 \le g_l - g_r \le \left(\frac{\tau_l}{\tau_h - \tau_l}\right)\frac{2\kappa_m}{1-\kappa_m}, \tag{2}$$

then either party, when in power, adopts the median policy. Moreover, affective polarization disappears in the long run, that is, $g_{l,t} - g_{r,t}$ converges monotonically to 0 as $t \to \infty$.

[13] The distribution of affinity for party B is implicitly defined given that a voter’s affinity for party A is high iff their affinity for B is low.

https://doi.org/10.1017/S0007123422000540 Published online by Cambridge University Press

When affective polarization is low, that is, $g_l - g_r$ is small, there is little electoral advantage for the incumbent to appeal to extreme voters. For illustration, consider the case where $g_l = g_r$. If the incumbent were to deviate from the median, say to the left, then the gains they make with l voters will be canceled out by the loss of r voters. On top of it, they lose support from m voters. Thus, there is a net loss from the deviation, and there is no incentive for the incumbent to deviate. The moderate stance of the parties will, in turn, limit affective partisanship, especially among voters at the extremes of the spectrum, thereby creating a "virtuous cycle" of centrist policies and moderate affective partisanship.

An immediate consequence of Proposition 1 is the following corollary:

Corollary 1: A low polarization path arises if group identity is not salient (that is, $\tau_h - \tau_l$ is small) or the size of moderate voters is large (that is, κ_m is large).

For intuition, consider the extreme case where group identity is not operative, that is, $\tau_h = \tau_l$. Thus, regardless of ideological position, voters are equally responsive to the incumbent's policy. The incumbent therefore has nothing to gain from biasing their policy toward either l or r voters, as any gains in support from one group will be canceled out by the loss of support from the other group. Besides the salience of group identity, the distribution of voters on the ideological spectrum is also an important factor in the incumbent's calculus. Specifically, the greater the proportion of moderate voters, the more costly it is for the incumbent to deviate from the center.

Next, we provide a sufficient condition for a high-polarization path in which the parties choose extreme policies (that is, party A chooses -1 when in power and B chooses 1) and a high level of affective partisanship persists. To state the result, we define $\tilde{g}(\delta)$ to be the solution of the following equation:

$$\frac{\alpha(l, \delta)}{1 - \alpha(h, \delta)} = \frac{\tilde{g}(\delta)}{1 - \tilde{g}(\delta)}.$$

Intuitively, $\tilde{g}(\delta)$ is the stationary distribution of affinity among a bloc of voters, assuming one party is always in power and chooses a policy that is δ away from the voters' bliss point. We then have the following result:

Proposition 2 (high polarization): If g_l and g_r in the initial period satisfy the following conditions:

$$g_l - g_r > \left(\frac{\tau_l}{\tau_h - \tau_l} + 1\right)\frac{2\kappa_m}{1 - \kappa_m} \tag{3}$$

and

$$\max\{1 - \tilde{g}(0), \tilde{g}(2)\} \leq g_r < g_l \leq \min\{\tilde{g}(0), 1 - \tilde{g}(2)\}, \tag{4}$$

then party A chooses $\theta_t = -1$ and B chooses $\theta_t = 1$ whenever in power and:

$$\limsup_{t\to\infty} g_{r,t} \leq \max\{1 - \tilde{g}(0), \tilde{g}(2)\} < \frac{1}{2} < \min\{\tilde{g}(0), 1 - \tilde{g}(2)\} \leq \liminf_{t\to\infty} g_{l,t}. \tag{5}$$

If affective partisanship is sufficiently extreme, then it is in the incumbent's interest to appeal to its partisans by deviating from the center. Doing so will engage its base sufficiently to overcome any loss of votes from the other blocs. The extreme policy stance, in turn, reinforces affective

https://doi.org/10.1017/S0007123422000540 Published online by Cambridge University Press

partisanship, thereby creating a “vicious cycle” of policy divergence and high levels of affective partisanship.

Condition 3 mirrors Condition 2, for it is a lower bound on the initial level of affective polarization. The bound can be interpreted as a “tipping point” for polarization: if the initial level of polarization is above this threshold, then it will remain above this threshold forever because of the vicious cycle logic. How might one identify the tipping point empirically, say, in the case of the United States? Given appropriate time-series data, the level of polarization at time t may be considered as the tipping point if: (1) polarization in subsequent periods does not show moderation, minor volatility notwithstanding; and (2) there is a sharp rise in the ideological division between the Democratic and Republican parties around time t that persists afterward. Condition 4 provides bounds that allow us to establish monotonicity, in the sense that starting within this bound, g_l and g_r will eventually move outside of it.[14]

It is straightforward to see that Condition 3 is more easily satisfied if $\tau_h - \tau_l$ is large and κ_m is small, giving rise to the following corollary:

Corollary 2: A high polarization path is more likely to arise if group identity is salient (that is, $\tau_h - \tau_l$ is large) and the size of moderate voters is small (that is, κ_m is small).

In general, a stronger sense of group identity and fewer moderate voters increase the incumbent’s incentive to appeal to its partisan base. Suppose *Party A* is the incumbent, then strong salience of group identity means that l voters are much more responsive to *Party A*’s policy than r voters. This means that *Party A* would rather excite its base than “reach across the aisle” and sway r voters. Since moderate voters exert a centripetal force upon *Party A*, a strategy of “rallying the base” will indeed be optimal if moderate voters are few.

Discussion

One broad lesson from our results is that shocks to the electoral environment can have substantial long-term consequences with regard to political polarization. The idea is similar to a standard exercise in macroeconomic research, starting with the Nobel Prize-winning work of Kydland and Prescott (1982), which shows how exogenous shocks to such variables as monetary policy or technology can generate aggregate fluctuations in output and employment. Importantly, even temporary shocks can have a persistent impact on the long-run equilibrium path through changing players’ beliefs—a phenomenon known as “hysteresis” (see, for example, Cooper 1994; Morris and Yildiz 2019). Thus, the economy can dive into a persistent recession if people become (briefly) pessimistic about the future due to unforeseen events that have little direct effect on the economy’s fundamentals.

In the context of our model, shocks to the salience of group identity or the number of moderate voters can switch the equilibrium path from one of low polarization to one of high polarization, or vice versa. Suppose, for example, the system is initially on the low polarization path, where parties are choosing centrist policies and affective partisanship is low (that is, Condition 2 holds). Suppose at some date t, there is a shock to group salience such that $\tau_h - \tau_l$ increases to $\tilde{\tau}_h - \tilde{\tau}_l$. If it is the case that $g_{l,t} - g_{r,t} > ((\tilde{\tau}_l/\tilde{\tau}_h - \tilde{\tau}_l) + 1)(2\kappa_m/1 - \kappa_m)$, then the system will switch to the high polarization path, where the incumbent chooses extreme policies and affective partisanship is high. It should be noted that the shocks need not be permanent to induce a permanent change. If shocks to the salience of group identity described earlier persist through several periods, so that affective polarization increases significantly, then even if the salience of group

[14] Condition 4 is not necessary if one is only interested in showing that affective polarization remains above some threshold.

identity eventually reverts back to the original level, the high polarization path will persist. The aforementioned shocks can take the form of specific political events, such as an economic crisis or political scandals. Some observers have pointed to the end of the Cold War as an accelerator of political polarization in the United States (Blankenhorn 2018). The idea is that the liberal/conservative identity is less salient when there is a "common enemy," as in the Soviet Union. The disappearance of a common enemy then highlights group differences. Other candidates include the financial crisis and the subsequent outrage at the bailing out of banks and other financial institutions, which gave rise to the Tea Party. A particularly interesting extension of our model would include rhetorical strategies by political entrepreneurs to highlight group identity for their political benefit. The details of such a model are, however, beyond the scope of this article.

Another interesting observation from the model has to do with how affective polarization relates to ideological polarization. In standard Downsian models, the ideological composition of the electorate is of little consequence to elite behavior. Specifically, a main prediction of Downsian theory is policy convergence with office-motivated candidates, in the form of median- and mean-voter theorems, *for any distribution of voter ideologies*.[15] In our model, an increase in ideological division, that is, a decrease in κ_m, can have a material effect on elite behavior. This does, however, depend crucially on affective partisanship. If group identity is not salient, that is, $\tau_h - \tau_l$ is small, then Condition 2 for the low polarization path will be satisfied regardless of the extent of ideological division. In that case, one recovers the classic convergence result. On the other hand, if group identity is salient, then sufficient ideological division as reflected by a small κ_m can lead to high polarization. In other words, issue polarization is neither necessary nor sufficient for elite polarization. It will not matter unless affective polarization is sufficiently high. If, however, affective polarization is indeed sufficiently high, issue polarization will serve as an amplifier of polarization.

The model has various implications that could help guide empirical studies. One obvious pattern is the following:

Implication 1: The trend in affective polarization is correlated with policy divergence: affective partisanship decreases over time when differentiation between parties' platforms is small and increases when the differentiation is large.

It should be noted that we are careful not to assert a causal relationship in any particular direction. In our model, affective partisanship and elite behavior influence each other (that is, both are endogenous variables). From the data, it may appear that changes in the elite's behavior precede changes in the attitudes of the electorate (see Implication 5), and one may therefore be tempted to infer a particular causal relationship. Our model suggests this line of thinking may be misguided and that a more sophisticated empirical approach, for example, simultaneous equations or structural approaches, would be required to understand the connection between affective and elite polarization.

The next two empirical implications from the model identify factors associated with high or low polarization regimes:

Implication 2: The more salient is group identity, the more likely the high polarization regime will emerge; the less salient, the more likely the low polarization regime.

[15] Policy divergence in the Downsian setting requires policy-motivated candidates and uncertainty about the election outcome (see, for example, Calvert 1985; Wittman 1983), or, alternatively, one candidate having a valence advantage (see, for example, Ansolabehere and Snyder 2000; Groseclose 2001). One exception is Kamada and Kojima (2014), where policy divergence occurs with office-motivated candidates. However, their result relies crucially on the assumption that voters' preferences are *convex*, that is, risk loving.

https://doi.org/10.1017/S0007123422000540 Published online by Cambridge University Press

Implication 3: The smaller the segment of moderate voters, the more likely the high polarization regime will emerge; the smaller the segment of moderate voters, the more likely the low polarization regime.

As discussed earlier, ideological differences on issue only matter if group identity is sufficiently strong:

Implication 4: Parties' platforms will become divergent as more voters hold extreme ideological positions but only when there is a heightened sense of group identity based on partisan affiliation.

The next implication follows from the observation that a transition between the low and the high polarization regimes entails a drastic change in the parties' platforms, for example, a switch from centrist to extreme policies. However, within a given regime, parties' policy positions will not change much. On the other hand, the level of affective partisanship among the masses changes gradually, both within and across regimes:

Implication 5: Changes to the parties' policy positions are less frequent but more drastic than changes in the levels of affective partisanship.

Finally, we point out a direct consequence of in-group favoritism versus in-group responsiveness:

Implication 6: Voters of a party will be more forgiving for general lack of performance or personal misconduct of public officials of their own party, provided such cases are unrelated to the core differences between the parties. On issues related to such core differences, public officials of the same party will be treated more harshly if they deviate from the party line.

Implications 2, 3, and 6 are straightforward and consistent with prior insights in the polarization literature. They demonstrate that our model can account for known empirical regularities. Implications 1, 4 and 5, on the other hand, are novel and somewhat subtle. They point to the value of developing a formal model of affective polarization.

Conclusion

While elite polarization has been well established, there are continuing debates about its cause. Recent research documents a rise in affective partisanship along the same period, suggesting a potential link between the two phenomena. In this article, we study a dynamic behavioral-voting model in which the elite's policy stance and the affective partisanship of the masses can mutually sustain each other, creating virtuous or vicious cycles. This can account for the observed temporal correlation between mass and elite partisanship noted by Coleman (1996) and Bartels (2000). We show that a high polarization path exists when group identity is salient and ideological division is high, while a low polarization path exists when group identity is not salient and ideological division is low. Thus, such events as economic crises that magnify group identity can lead to changes in the trajectory of political polarization. On the other hand, when group identity is less salient, for example, during periods of national security challenges attributed to a common enemy, polarization in general will be lower.

Notably, these implications hold even when voters' attitudes on policies are stable. Indeed, an electorate's polarization on issues is irrelevant for elite polarization unless voters exhibit a sufficient level of dislike for members of the other party. While changes in affective polarization may

https://doi.org/10.1017/S0007123422000540 Published online by Cambridge University Press

be the result of external events, they may also result from rhetorical strategies by candidates or changes in the media environment, for example, the emergence of more polarizing news sources.

The interactive, dynamic nature of our model suggests that empirical studies need to account for the mutual influence of mass and elite polarization. This would require estimation techniques from, for example, macroeconomics or industrial organization that can address the simultaneity issues. This is no accident. In both macroeconomics and our model, we find multiple regimes sustained by positive feedback loops and hysteresis, that is, the long-term impact of temporary shocks.

Although we framed the model and results based on US national politics, the model could be applicable in other contexts as well, for example, state and local politics in the United States. Indeed, the model is mostly agnostic about institutional features. The only limit on the scope is the assumption of two-party systems. Extending the model to multiparty systems is not straightforward given how we define and measure affective polarization. Such an extension will be especially challenging in multiparty systems that involve coalition formation, for example, parliamentary democracies under proportional representation. With those qualifications in mind, we believe that the model, though highly stylized, is a useful first step in studying the dynamics of affective partisanship and elite behavior. We hope for more research on this topic in the future.

Acknowledgments. We thank the editor and four anonymous reviewers for their helpful comments.

Financial Support. None.

Competing Interests. None.

References

Abramowitz A (2010) *The Disappearing Center: Engaged Citizens, Polarization, and American Democracy*. New Haven, CT: Yale University Press.

Abramowitz A and Saunders K (2008) Is polarization a myth? *The Journal of Politics* **70**(2), 542–555.

Abramowitz A and Webster S (2016) The rise of negative partisanship and the nationalization of US elections in the 21st century. *Electoral Studies* **41**, 12–22.

Achen CH and Bartels LM (2017) *Democracy for Realists: Why Elections Do Not Produce Responsive Government*. Princeton, NJ: Princeton University Press.

Akerlof GA and Kranton RE (2000) Economics and identity. *The Quarterly Journal of Economics* **115**(3), 715–753.

Andonie C and Diermeier D (2019) Impressionable voters. *American Economic Journal: Microeconomics* **11**(1), 79–104.

Ansolabehere S and Snyder JM (2000) Valence politics and equilibrium in spatial election models. *Public Choice* **103**(3–4), 327–336.

Banda KK and Cluverius J (2018) Elite polarization, party extremity, and affective polarization. *Electoral Studies* **56**, 90–101.

Barber M and McCarty N (2015) Causes and consequences of polarization. In Mansbridge J and Martin CJ (eds), *Political Negotiation: A Handbook*. Washington, DC: Brooking Institution Press, 39–43.

Bartels LM (2000) Partisanship and voting behavior, 1952–1996. *American Journal of Political Science* **44**, 35–50.

Bendor J (2010) *Bounded Rationality and Politics*. Berkeley, CA: University of California Press.

Bendor J, Diermeier D and Ting M (2003) A behavioral model of turnout. *American Political Science Review* **97**(2), 261–280.

Bendor J, Kumar S and Siegel D (2010) Adaptively rational retrospective voting. *Journal of Theoretical Politics* **22**(1), 26–63.

Bendor J et al. (2011) *A Behavioral Theory of Elections*. Princeton, NJ: Princeton University Press.

Berelson B, Lazarsfeld P and McPhee W (1954) *Voting: A Study of Opinion Formation in A Presidential Campaign*. Chicago, IL: University of Chicago Press.

Biernat M, Vescio TK and Billings LS (1999) Black sheep and expectancy violation: integrating two models of social judgment. *European Journal of Social Psychology* **29**(4), 523–542.

Bisin A, Lizzeri A and Yariv L (2015) Government policy with time inconsistent voters. *American Economic Review* **105**(6), 1711–1737.

Blankenhorn D (2018) The top 14 causes of political polarization. *The American Interest* **16**, 126.

Börgers T and Sarin R (1997) Learning through reinforcement and replicator dynamics. *Journal of Economic Theory* **77**(1), 1–14.

Burden B and Klofstad C (2005) Affect and cognition in party identification. *Political Psychology* **26**(6), 869–886.

Calvert R (1985) Robustness of the multidimensional voting model: candidate motivations, uncertainty, and convergence. *American Journal of Political Science* **29**(1), 69–95.

https://doi.org/10.1017/S0007123422000540 Published online by Cambridge University Press

Campbell A et al. (1960) *The American Voter*. New York: John Wiley & Sons.

Castano E et al. (2002) Protecting the ingroup stereotype: ingroup identification and the management of deviant ingroup members. *British Journal of Social Psychology* **41**(3), 365–385.

Chen Y and Li SX (2009) Group identity and social preferences. *American Economic Review* **99**(1), 431–457.

Coleman JJ (1996) *Party Decline in America: Policy, Politics, and the Fiscal State*. Princeton, NJ: Princeton University Press.

Cooper R (1994) Equilibrium selection in imperfectly competitive economies with multiple equilibria. *The Economic Journal* **104**(426), 1106–1122.

Dias N and Lelkes Y (2022) The nature of affective polarization: disentangling policy disagreement from partisan identity. *American Journal of Political Science* **66**(3), 775–790.

Diermeier D and Li C (2017) Electoral control with behavioral voters. *The Journal of Politics* **79**(3), 890–902.

Diermeier D and Li C (2019) Partisan affect and elite polarization. *American Political Science Review* **113**(1), 277–281.

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

Duggan J (2012) A Survey of Equilibrium Analysis in Spatial Models of Elections. Unpublished manuscript.

Eggers AC et al. (2014) Partisanship and electoral accountability: evidence from the UK expenses scandal. *Quarterly Journal of Political Science* **9**(4), 441–472.

Erikson R, Mackuen M and Stimson J (2002) *The Macro Polity*. Cambridge, UK: Cambridge University Press.

Fiorina M (1981) *Retrospective Voting in American National Elections*. New Haven, CT: Yale University Press.

Fiorina M and Abrams S (2008) Political polarization in the American public. *Annual Review of Political Science* **11**, 563–588.

Fiorina MP, Abrams SJ and Pope J (2011) *Culture War? The Myth of a Polarized America*. New York, NY: Longman.

Frimer JA and Skitka LJ (2020) Americans hold their political leaders to a higher discursive standard than rank-and-file co-partisans. *Journal of Experimental Social Psychology* **86**, 103907.

Green D, Palmquist B and Schickler E (2004) *Partisan Hearts and Minds: Political Parties and the Social Identities of Voters*. New Haven, CT: Yale University Press.

Greene S (2004) Social identity theory and party identification. *Social Science Quarterly* **85**(1), 136–153.

Groseclose T (2001) A model of candidate location when one candidate has a valence advantage. *American Journal of Political Science* **45**(4), 862–886.

Hart S (2005) Adaptive heuristics. *Econometrica* **73**(5), 1401–1430.

Healy A and Malhotra N (2013) Retrospective voting reconsidered. *Annual Review of Political Science* **16**, 285–306.

Hetherington MJ (2009) Putting polarization in perspective. *British Journal of Political Science* **39**(2), 413–448.

Hilgard ER and Bower GH (1966) *Theories of Learning*. New York: Appleton-Century-Crofts.

Hill SJ and Tausanovitch C (2015) A disconnect in representation? Comparison of trends in congressional and public polarization. *The Journal of Politics* **77**(4), 1058–1075.

Iyengar S and Krupenkin M (2018) The strengthening of partisan affect. *Political Psychology* **39**, 201–218.

Iyengar S, Sood G and Lelkes Y (2012) Affect, not ideology: a social identity perspective on polarization. *Public Opinion Quarterly* **76**(3), 405–431.

Iyengar S et al. (2019) The origins and consequences of affective polarization in the United States. *Annual Review of Political Science* **22**, 129–146.

Jessee SA (2010) Partisan bias, political information and spatial voting in the 2008 presidential election. *The Journal of Politics* **72**(2), 327–340.

Joesten DA and Stone WJ (2014) Reassessing proximity voting: expertise, party, and choice in congressional elections. *The Journal of Politics* **76**(3), 740–753.

Kamada Y and Kojima F (2014) Voter preferences, polarization, and electoral policies. *American Economic Journal: Microeconomics* **6**(4), 203–236.

Kayser MA and Wlezien C (2011) Performance pressure: patterns of partisanship and the economic vote. *European Journal of Political Research* **50**(3), 365–394.

Kimball DC, Anthony J and Chance T (2018) Political identity and party polarization in the American electorate. In Green J, Coffeey D and Cohen D (eds), *The State of the Parties*. London: Rowman & Littlefield, 169–184.

Kydland FE and Prescott EC (1982) Time to build and aggregate fluctuations. *Econometrica* **50**(6), 1345–1370.

Landa D and Duell D (2015) Social identity and electoral accountability. *American Journal of Political Science* **59**(3), 671–689.

Layman G, Carsey T and Horowitz J (2006) Party polarization in American politics: characteristics, causes, and consequences. *Annual Review of Political Science* **9**, 83–110.

Lazarus RS (1991) Cognition and motivation in emotion. *American Psychologist* **46**(4), 352.

Lee T (2013) How the Tea Party broke the constitution. *Washington Post*, October 14.

Lerner JS and Keltner D (2000) Beyond valence: toward a model of emotion-specific influences on judgement and choice. *Cognition & Emotion* **14**(4), 473–493.

Levendusky M (2009) *The Partisan Sort: How Liberals Became Democrats and Conservatives Became Republicans*. Chicago, IL: University of Chicago Press.

Levy G and Razin R (2015) Correlation neglect, voting behavior, and information aggregation. *American Economic Review* **105**(4), 1634–1645.

https://doi.org/10.1017/S0007123422000540 Published online by Cambridge University Press

Lewis-Beck M et al. (2008) *The American Voter Revisited*. Ann Arbor, MI: University of Michigan Press.

Little AT (2019) The distortion of related beliefs. *American Journal of Political Science* **63**(3), 675–689.

Lizzeri A and Yariv L (2017) Collective self-control. *American Economic Journal: Microeconomics* **9**(3), 213–244.

Lodge M and Taber CS (2013) *The Rationalizing Voter*. New York, NY: Cambridge University Press.

Mackie D and Cooper J (1984) Attitude polarization: effects of group membership. *Journal of Personality and Social Psychology* **46**(3), 575.

MacKuen MB, Erikson RS and Stimson JA (1989) Macropartisanship. *American Political Science Review* **83**(4), 1125–1142.

Marques J and Paez D (1994) The black sheep effect: social categorization, rejection of ingroup deviates, and perception of group variability. *European Review of Social Psychology* **5**(1), 37–68.

Marques J, Yzerbyt V and Leyens J-P (1988) The "black sheep effect": extremity of judgments towards ingroup members as a function of group identification. *European Journal of Social Psychology* **18**(1), 1–16.

McCarty N, Poole KT and Rosenthal H (2016) *Polarized America: The Dance of Ideology and Unequal Riches*. Cambridge, MA: MIT Press.

Mendoza SA, Lane SP and Amodio DM (2014) For members only: ingroup punishment of fairness norm violations in the ultimatum game. *Social Psychological and Personality Science* **5**(6), 662–670.

Miller W, Shanks M and Shapiro R (1996) *The New American Voter*. Cambridge, MA: Harvard University Press.

Minozzi W (2013) Endogenous beliefs in models of politics. *American Journal of Political Science* **57**(3), 566–581.

Morris S and Yildiz M (2019) Crises: equilibrium shifts and large shocks. *American Economic Review* **109**(8), 2823–2854.

Mullen B, Brown R and Smith C (1992) Ingroup bias as a function of salience, relevance, and status: an integration. *European Journal of Social Psychology* **22**(2), 103–122.

Newport F (2010) Tea Party supporters overlap Republican base. *Gallup News*, July 2.

Ortoleva P and Snowberg E (2015) Overconfidence in political behavior. *American Economic Review* **105**(2), 504–535.

Patty J and Penn M (2020) Identity and Information in Organizations. Working paper.

Pinto IR et al. (2010) Membership status and subjective group dynamics: who triggers the black sheep effect? *Journal of Personality and Social Psychology* **99**(1), 107.

Popkin S (1994) *The Reasoning Voter: Communication and Persuasion in Presidential Campaigns*. Chicago, IL: University of Chicago Press.

Roemer JE (1994) A theory of policy differentiation in single issue electoral politics. *Social Choice and Welfare* **11**(4), 355–380.

Rogowski J and Sutherland J (2016) How ideology fuels affective polarization. *Political Behavior* **38**(2), 485–508.

Tajfel H (1970) Experiments in intergroup discrimination. *Scientific American* **223**(5), 96–103.

Tajfel H and Turner J (1979) An integrative theory of intergroup conflict. *The Social Psychology of Intergroup Relations* **33**(47), 74.

Wittman D (1983) Candidate motivation: a synthesis of alternative theories. *American Political Science Review* **77**(1), 142–157.

Woon J (2012) Democratic accountability and retrospective voting: a laboratory experiment. *American Journal of Political Science* **56**(4), 913–930.

Zaller JR et al. (1992) *The Nature and Origins of Mass Opinion*. Cambridge, UK: Cambridge University Press.

Appendix 1

General α

Denote the probability of a voter voting for the incumbent as $\alpha(p, \delta)$, with $\delta = |\theta - b|$. We will sometimes use the alternative expression $\alpha_b(p, \theta) \equiv \alpha(p, \delta)$ when it is more convenient for exposition. Let $\alpha(p, \delta)$ be differentiable with $\alpha(p, 0) = \lim_{\delta \to 0} \alpha(p, \delta)$ and $\partial\alpha(p, 0)/\partial\delta \equiv \lim_{\delta \to 0} \partial\alpha(p, \delta)/\partial\delta$.[16] We impose two further assumptions on $\alpha(p, \delta)$. The first is in-group favoritism:

Assumption 2 (in-group favoritism): $\alpha(h, \delta) > \alpha(l, \delta)$ for all δ and $(\partial\alpha(p, \delta)/\partial\delta) \leq 0$, with strict inequality for δ in the interior.

The second is in-group responsiveness:

Assumption 3 (in-group responsiveness): $|(\partial\alpha(h, \delta)/\partial\delta)| > |(\partial\alpha(l, \delta)/\partial\delta)|$.

First, we show that given initial affective polarization, parties develop distinct ideologies and voters develop stable party identification:

[16] Since the derivative is defined only in the interior of the domain, $\partial\alpha(p, 0)/\partial\delta$ is technically undefined. However, given α is continuous at $\delta = 0$, it is convenient for exposition purposes (in the case of finding the optimum) to define $\partial\alpha(p, 0)/\partial\delta$ as the right-hand limit.

https://doi.org/10.1017/S0007123422000540 Published online by Cambridge University Press

Proposition 3: Suppose the electorate is polarized initially, then for all future periods, *Party A* chooses leftist policies and *Party B* chooses rightist policies whenever in office (that is, $\theta_t^A \leq 0 \leq \theta_t^B$), and l and r voters are partisan for *Party A* and *Party A*, respectively (that is, $g_{l,t} > (1/2) > g_{r,t}$).

Intuitively, affective polarization drives policy divergence because of in-group responsiveness, as in Diermeier and Li (2019). Moreover, the process by which partisan affinity evolves implies that by adopting a biased policy, the incumbent builds rapport with one group of voters at the expense of alienating another. This induces affective polarization. The next two results provide a partial generalization of the conditions for the low and high polarization paths in the baseline model in the main text. First, Proposition 4 identifies conditions under which there exists a non-trivial lower bound on polarization:

Proposition 4: If there exists $\bar{g} > (1/2)$, such that:

$$\begin{aligned} \alpha_l(h, \bar{\theta})\bar{g} + \alpha_l(l, \bar{\theta})(1 - \bar{g}) &\geq \bar{g} \\ \alpha_r(h, \bar{\theta})(1 - \bar{g}) + \alpha_r(l, \bar{\theta})\bar{g} &\leq 1 - \bar{g}, \end{aligned} \tag{6}$$

where $\bar{\theta}$ is *Party A*'s optimal policy given $g_l = \bar{g}$, $g_m = 1$, $g_r = 1 - \bar{g}$, then if the initial distribution of affinities is such that $g_l \geq \bar{g}$ and $g_r \leq 1 - \bar{g}$, then it is the case that $g_{l,t} \geq \bar{g}, g_{r,t} \leq 1 - \bar{g}$, and $\theta_t^A \leq \bar{\theta} < 0 < 1 - \bar{\theta} \leq \theta_t^B$ for all t.

The proof (as well as the one for Proposition 5) can be found in Appendix 2. The proposition states that if $\bar{g}$ satisfies Condition 6, then it is a lower bound on polarization, in the sense that the policies are bounded away from the median, and the degree of affective polarization $g_l - g_r$ is at least $2\bar{g}$.[17] Next, Proposition 5 derives a sufficient condition for an upper bound on polarization:

Proposition 5: If there exists $\underline{g} > (1/2)$, such that:

$$\begin{aligned} \alpha_l(h, \underline{\theta})\underline{g} + \alpha_l(l, \underline{\theta})(1 - \underline{g}) &\leq \underline{g} \\ \alpha_r(h, \underline{\theta})(1 - \underline{g}) + \alpha_r(l, \underline{\theta})\underline{g} &\geq 1 - \underline{g}, \end{aligned}$$

where $\underline{\theta}$ is *Party A*'s optimal policy given $g_l = \underline{g}$, $g_m = 0$, $g_r = 1 - \underline{g}$, then if the initial distribution of affinities is such that $g_l \leq \bar{g}$ and $g_r \geq 1 - \underline{g}$, then for any $s > t$, it is the case that $g_{l,t} \leq \underline{g}$, $g_{r,t} \geq 1 - \underline{g}$, and $\underline{\theta} \leq \theta_t^A \leq 0 \leq \theta_t^B \leq 1 - \underline{\theta}$ for all t.

Given Propositions 4 and 5, one can show that if $\bar{g} > \underline{g}$, then there are two qualitatively different equilibrium paths: one in which affective polarization and policy divergence are moderate; another in which affective polarization and policy divergence are severe.

Appendix 2

Proofs

Preliminaries

In this section, we characterize the optimal policy for the incumbent in a generic period when there is affective polarization within the electorate. Denote

$$V(\theta|g_l, g_m, g_r) = \begin{cases} \sum_b \kappa_b[\alpha_b(h, \theta)g_b + \alpha_b(l, \theta)(1 - g_b)] & \text{if the incumbent is } A \\ \sum_b \kappa_b[\alpha_b(h, \theta)(1 - g_b) + \alpha_b(l, \theta)g_b] & \text{if the incumbent is } B \end{cases}$$

as the incumbent's vote share (that is, its objective function). It should be noted that V is continuous in θ; therefore, by compactness, an optimal policy exists. For simplicity, we assume that the optimum is unique.[18] Lemma 1 shows that a polarized electorate can drive policy divergence. It follows directly from Propositions 1 and 3 in Diermeier and Li (2019), and we omit the proof:

[17] One can derive a more general result where the bounds for g_l and g_r are not symmetric. The conditions would be more tedious to state, but the logic is similar.

[18] The results hold qualitatively even when the optimum is not unique.

https://doi.org/10.1017/S0007123422000540 Published online by Cambridge University Press

Lemma 1: Suppose the electorate is polarized, then if A is the incumbent, its optimal policy θ^A is negative (that is, $\theta^A \leq 0$), and if B is the incumbent, its optimal policy θ^B is positive (that is, $\theta^B \geq 0$). Moreover, θ^A is decreasing in g_l and increasing in g_m and g_r. θ^B is increasing in g_r and decreasing in g_m and g_l.

It should be noted that the result does not rule out the possibility that the optimal policy is the median. The following corollary identifies a sufficient condition for strict policy divergence:

Corollary 3: If $(1 - \kappa_m/2\kappa_m)(g_l - g_r)$ is sufficiently large, then $\theta^A < 0$ and $\theta^B > 0$.

Proof

Without loss of generality, suppose A is the incumbent (the argument for $\theta^B > 0$ is similar). A's optimal policy cannot be the median if $(\partial V/\partial\theta)|_{\theta=0} < 0$, which is equivalent to the following inequality:

$$\frac{1-\kappa_m}{2}(g_l - g_r)\left(\frac{\partial\alpha(l,\delta)}{\partial\delta}|_{\delta=1} - \frac{\partial\alpha(h,\delta)}{\partial\delta}|_{\delta=1}\right) > -\kappa_m\left(g_m\frac{\partial\alpha(h,\delta)}{\partial\delta}|_{\delta=0} + (1-g_m)\frac{\partial\alpha(l,\delta)}{\partial\delta}|_{\delta=0}\right).$$

It should be noted that we define $(\partial\alpha(p,\delta)/\partial\delta)|_{\delta=0} \equiv \lim_{\delta\to 0}(\partial\alpha/\partial\delta)(p,\delta)$ and recalled that $(\partial\alpha(h,\delta)/\partial\delta) < (\partial\alpha(l,\delta)/\partial\delta) < 0$. It is straightforward to see that if $(1-\kappa_m/2\kappa_m)(g_l - g_r)$ is sufficiently large, the aforementioned inequality is satisfied and therefore $\theta^A < 0$.

The condition in the corollary is obtained when $g_l - g_r$ is large. Thus, sufficiently high affective partisanship induces elite polarization. It should also be noted that the condition is satisfied if κ_m is sufficiently small. Thus, strict policy divergence can also occur when the number of moderate voters is low.

Proof for Proposition 1

Consider the case when *Party A* is the incumbent. The fact that α is linear in δ means the marginal gain in votes by moving to the left of the median is $(1 - \kappa_m/2)(g_l\tau_h + (1 - g_l)\tau_l)$, while the marginal loss of votes is $(1 - \kappa_m/2)(g_r\tau_h + (1 - g_r)\tau_l) + \kappa_m(g_m\tau_h + (1 - g_m)\tau_l)$. Noting that

$$\frac{1-\kappa_m}{2}(g_l\tau_h + (1-g_l)\tau_l) - \frac{1-\kappa_m}{2}(g_r\tau_h + (1-g_r)\tau_l) = \frac{1-\kappa_m}{2}(g_l - g_r)(\tau_h - \tau_l),$$

Party A thus has no incentive to deviate from the median if:

$$g_l - g_r \leq \left(\frac{\tau_l}{\tau_h - \tau_l} + g_m\right)\frac{2\kappa_m}{1-\kappa_m}.$$

It should be noted that the preceding inequality is implied by $g_l - g_r \leq (\tau_l/\tau_h - \tau_l)\frac{2\kappa_m}{1-\kappa_m}$. A similar argument extends to the case when *Party B* is the incumbent. Moreover, given that the incumbents choose the median policy, the difference between the next period g_l' and g_r' satisfies:

$$g_l' - g_r' = (\alpha(h,1) - \alpha(l,1))(g_l - g_r).$$

Since $\tau_h - \tau_l < 1$, this quantity is less than $g_l - g_r$. It also follows that in the limit, as $t \to \infty$, we have that $g_r = g_l$.

Proof for Proposition 2

Following a similar argument as for Proposition 1, one can show that *Party A* (B) finds it optimal to choose the extreme policy if:

$$g_l - g_r > \left(\frac{\tau_l}{\tau_h - \tau_l} + g_m\right)\frac{2\kappa_m}{1-\kappa_m}.$$

Thus, if $g_l - g_r > ((\tau_l/\tau_h - \tau_l) + 1)(2\kappa_m/1 - \kappa_m)$, then the chosen policies diverge regardless of the distribution of party affinities g_b. Now, it should be noted that by the definition of $\tilde{g}(\cdot)$, if *Party A* is the incumbent at date t and implements

$\theta_t^A = -1$, and, moreover, if $g_{l,t} < \tilde{g}(0)$ ($g_{r,t} > \tilde{g}(2)$, respectively), then $g_{l,t} < g_{l,t+1} < \tilde{g}(0)$ ($g_{r,t} > g_{r,t+1} > \tilde{g}(2)$, respectively). Similarly, if *Party B* is the incumbent and implements $\theta_t^B = 1$, and if $g_{l,t} < 1 - \tilde{g}(2)$ ($g_{r,t} > 1 - \tilde{g}(0)$, respectively), then $g_{l,t} < g_{l,t+1} < 1 - \tilde{g}(2)$ ($g_{r,t} > g_{r,t+1} > 1 - \tilde{g}(0)$, respectively). Thus, given the conditions on the initial values of g_l and g_r, it must be that the polarization in the second period is at least as large as in the initial period. This argument can be iterated forward, and we will have that $g_{l,t} \geq g_l$ and $g_{r,t} \leq g_r$, and the incumbents always choose the extreme policy.

Now, define the function:

$$h(\delta, g) = \begin{cases} g_b\alpha(h, \delta) + (1 - g_b)\alpha(l, \delta) & \text{if the incumbent is } A \\ 1 - [(1 - g_b)\alpha(h, \delta) + g_b\alpha(l, \delta)] & \text{if the incumbent is } B \end{cases}. \tag{7}$$

This function computes the proportion of voters of a given bloc that would have high affinity for *Party A* in the next period when, today, the policy proximity is δ and the proportion of voters with high affinity for *Party A* is g. It should be noted that starting with any initial value of g, if one applies the function $h(\delta, \cdot)$ recursively, then in the limit, one obtains $\tilde{g}(\delta)$. Furthermore, the convergence is monotone. Now, suppose $\tilde{g}(0) < 1 - \tilde{g}(2)$. Since the incumbent chooses the extreme policy in every period, it must be that for any ϵ, there is a sufficiently large t where $g_{l,t} > \tilde{g}(0) - \epsilon$, that is, $liminfg_{l,t} \geq \tilde{g}(0)$. A similar argument applies for $1 - \tilde{g}(2) \leq \tilde{g}(0)$, and this implies that $\min\{\tilde{g}(0), 1 - \tilde{g}(2)\} \leq liminf_{t\to\infty}g_{l,t}$. As for $limsupg_{r,t}$, if $\tilde{g}(2) > 1 - \tilde{g}(0)$, then it must be that for any ϵ, there is a sufficiently large t where $g_{r,t} < \tilde{g}(2) + \epsilon$, that is, $limsup_{t\to\infty}g_{r,t} \leq \tilde{g}(2)$. If $\tilde{g}(2) < 1 - \tilde{g}(0)$, then a similar argument establishes that $limsup_{t\to\infty}g_{r,t} \leq 1 - \tilde{g}(2)$. Taken together, we have that $limsup_{t\to\infty}g_{r,t} \leq \max\{1 - \tilde{g}(0), \tilde{g}(2)\}$.

Proof for Proposition 3

It should be recalled that the proportion of b voters with high affinity for *Party A* given policy proximity δ and g_b is given by $g_b' \equiv h(\delta, g_b)$ (see Equation 7). Suppose the electorate is polarized and *Party A* is the incumbent at a particular date (the case where *Party B* is the incumbent is similar). Given Lemma 1, *Party A* chooses $\theta^A \leq 0$. We will argue that the electorate in the next period remains polarized. This follows simply from the fact that $h(\delta, g_b)$ is decreasing in δ and increasing in g_b. Specifically, $\theta^A \leq 0$ implies that $\delta_l \leq \delta_r$, and by assumption, we have $g_l > g_r$. Thus, $h(\delta_l, g_l) > h(\delta_r, g_r)$. Given that the electorate remains polarized in the subsequent period, Lemma 1 implies biased policies for either party. We iterate the preceding argument starting from date 1 and obtain the result.

Proof for Proposition 4

It should be recalled that $V = \sum_b \kappa_b[\alpha_b(h, \theta)g_b + \alpha_b(l, \theta)(1 - g_b)]$ is the objective function for *Party A*. We restrict our attention to θ in the interval $[-1, 0]$ given Lemma 1. First, it should be noted that $\bar{\theta} < 0$. Suppose $\bar{\theta} = 0$, then $\alpha_l(h, 0) = \alpha_r(h, 0) = \bar{\alpha}$ and $\alpha_l(l, 0) = \alpha_r(l, 0) = \underline{\alpha}$. The first inequality implies that $(\underline{\alpha}/1 - \bar{\alpha}) \geq (\bar{g}/1 - \bar{g})$, while the second inequality implies that $(\underline{\alpha}/1 - \bar{\alpha}) \leq (1 - \bar{g}/\bar{g})$; this is a contradiction because, by assumption, $\bar{g} > (1/2)$. Now, seeing V as a function of $-\theta$, it satisfies increasing differences with respect to $g_r, -g_m, -g_r$. This means that by monotone comparative statics, the optimal choice for *Party A*, $< ddollar > \theta^A$, is decreasing in g_r and increasing in g_m and g_r. Now, given $g_l = \bar{g}$ and $g_r = 1 - \bar{g}$, θ^A must be weakly smaller than $\bar{\theta}$ because $g_m \leq 1$. Furthermore, starting from the initial values $g_l \geq \bar{g}$ and $g_r \leq 1 - \bar{g}$, and supposing that *Party A* is the incumbent at date t, we have that $\theta_t^A \leq \bar{\theta}$. Now, if the inequalities given in Equation 6 are satisfied, then $\theta_t^A \leq \bar{\theta}$ implies that $g_{l,t+1} \geq \bar{g}$ and $g_{r,t+1} \leq 1 - \bar{g}$. This follows from the fact that $\alpha_l(h, \theta)g + \alpha_l(l, \theta)(1 - g)$ is decreasing in θ and increasing in g, and $\alpha_r(h, \theta)(1 - g) + \alpha_r(l, \theta)g$ is increasing in θ and decreasing in g. It should, in fact, be noted that $g_{l,t} \geq \bar{g}$ and $g_{r,t} \leq 1 - \bar{g}$ imply that $\theta_{t+1}^A \leq \bar{\theta}$. Now, given the symmetry of the setting, we see that whenever *Party B* is in power and $g_{l,t} \geq \bar{g}$ and $g_{r,t} \leq 1 - \bar{g}$, it must be that $\theta_{t+1}^B \geq 1 - \bar{\theta}$. Moreover, $\theta_t^B \geq 1 - \bar{\theta}$ implies that $g_{l,t+1} \geq \bar{g}$ and $g_{r,t+1} \leq 1 - \bar{g}$. The result then follows by induction.

Proof for Proposition 5

The proof is similar to that of Proposition 4, and we will therefore only provide an outline and omit the details. By monotone comparative statics, the optimal policy for *Party A* is decreasing in g_r and increasing in g_m and g_r. Assuming $g_l = \text{g}$ and $g_r = 1 - \text{g}$, θ^A is weakly greater than $\underline{\theta}$. Now, the inequalities in the proposition imply that whenever $\theta_t^A \geq \underline{\theta}$, it is the case that $g_{l,t+1} \leq \text{g}$ and $g_{r,t+1} \geq 1 - \text{g}$. This means that $\theta_{t+1}^A \geq \underline{\theta}$. Given the symmetry of the environment, *Party B*'s optimal policy would be the mirror opposite of *Party A*'s and will imply that $g_{l,t+1} \leq \text{g}$ and $g_{r,t+1} \geq 1 - \text{g}$. By induction we have that $g_{l,s} \leq \text{g}$, $g_{r,s} \geq 1 - \text{g}$, $\theta_s^A \geq \underline{\theta}$, and $\theta_s^B \leq 1 - \underline{\theta}$ for all $s \geq t$.

Cite this article: Diermeier D, Li C (2023). Dynamics of Polarization: Affective Partisanship and Policy Divergence. *British Journal of Political Science* **53**, 980–996. https://doi.org/10.1017/S0007123422000540