Political salience and regime resilience[*]

Sebastian Schweighofer-Kodritsch, HU Berlin[†]
Steffen Huck, WZB Berlin Social Science Center
Macartan Humphreys, WZB, HU Berlin, TCD

August 7, 2025

Abstract

We introduce political salience into a canonical model of attacks against political regimes, as scaling agents' expressive payoffs from taking sides. Equilibrium balances heterogeneous expressive motives with incentives to avoid sanctions by joining the winning side. We examine comparative statics in political salience, which we characterize in terms of equilibrium stability as well as attack size. A main insight is that when regime sanctions are weak, increases from low to middling salience can pose the greatest threat to seemingly safe regimes: ever smaller shocks become sufficient to drastically escalate into full-blown attacks, i.e., the regime becomes less resilient. Stronger regime safeguards not only directly reduce incentives to attack but can overturn these effects, such that increases in salience boost regime resilience. Our results speak to the charged debates about democracy, by identifying how safeguards determine when a rise in citizen interest in political action can lead to a threat to democracy.

JEL Classification: C72, D74, D91

Keywords: political conflict, salience, democracy, sanctions

[*] Our thanks to the editor and two anonymous reviewers for exceptionally helpful and constructive comments, in addition to Daniel Markovits, Ethan Bueno de Mesquita, Stephan Lauermann, Tom Palfrey, Carlo Prato, and Haoyu Zhai for generous advice on this project, and to seminar participants at HU Berlin for helpful remarks. Schweighofer-Kodritsch gratefully acknowledges financial support by the *Deutsche Forschungsgemeinschaft* through CRC TRR 190 (project number 280092119). Declarations of interest: none.

[†] Corresponding author. Email address: sebastian.kodritsch@gmail.com. Postal address: Prof. Sebastian Schweighofer-Kodritsch, School of Business and Economics, Humboldt-Universität zu Berlin, Spandauer Straße 1, 10099 Berlin, Germany.

1 Introduction

Alongside rising concerns regarding the resilience of American democracy, there has been a new focus on better understanding why and how democracies become vulnerable. We contribute to this work by returning to a canonical model of collective action and introducing a focus on *political salience*—encapsulating how much citizens care about expressing their individual regime preference—to assess how such salience, in and of itself, affects the vulnerability of political regimes.

In some accounts, it is disinterest in politics (low salience) that threatens democracy. The "slow slump in interest in politics and current events," according to Putnam [2000], can be one source of vulnerability. Other work highlights increasing salience in terms of the rising stakes of political decision-making. Levitsky and Ziblatt [2018], for instance, describe the erosion of democratic norms as politics becomes polarized and conflicts become more total. There are thus straightforward, if conflicting, logics through which changes in political salience can threaten political regimes.[1]

In this paper, we point to a critical interplay between political salience on the one hand and regime safeguards on the other.

Following Kuran [1989], Medina [2007] and others, we examine a model in which a large population of citizens individually decide whether to take a stance against ("attack") or in support of ("defend") an incumbent regime under threat by an opponent. The probability that the attack is successful is increasing in the share of citizens that join it.

The citizens are heterogeneous in their intrinsic political values. Some favor the regime (its supporters), others favor the alternative (its opponents), and with varying intensities. Formally, in our model, citizens' political values are continuously distributed on a line segment around zero as the point of indifference. While being negligible in determining the collective outcome, citizens have expressive motives to act in line with their values: they experience a payoff that depends on their own action only, which reflects their intrinsic political values, including intensities. Crucially, we model political salience via a parameter scaling these values in citizens' utilities and thus determining the general importance of expressive motives. Our interpretation of political salience is that it captures the general public's "bottom-up" attention to political values in the conflict, in line with the seminal approach to modeling psychological salience in economics [for a recent authoritative review see Bordalo et al., 2022]. As such, it is likely affected by the behavior of political elites and the media, in complex ways.[2] We

[1] We leave aside the empirical question of whether political salience is rising or falling. In some accounts it is falling, as in Putnam [2000]. In many journalistic accounts it is rising: Prior and Bougher [2018] cites many examples, while also showing that political interest has historically been quite constant on average. Of course, this can mask variation in politically active subgroups.

[2] For related evidence that mass polarization follows rather than drives the polarization of

analyze the effects of exogenous changes in political salience while remaining agnostic about their source.

Our second key parameter of interest captures the regime's safeguards, modeled as sanctions imposed on failed insurgents, where we also allow for sanctions on failed regime defenders upon a successful attack. Concerns about sanctioning give rise to (homogeneous) coordination incentives to join the winning side and thus escape its sanctions. All else equal, when more are expected to attack, more become willing to attack. Depending on their expectations of the attack, citizens may therefore face a trade-off between expressing their true values and avoiding sanctions, in choosing their action.

In analogy to the so-called "bandwagon effect" in voting, we will refer to acting contrary to one's political values in order to align with the "winning side" as *bandwagoning*.[3] Thus, a regime supporter is bandwagoning if they join an attack on the regime to avoid sanctions should the regime be defeated, and a regime opponent is bandwagoning if they defend the regime for fear of sanctions should the insurgency be defeated.[4] By incentivizing bandwagoning, sanctions act as a moderator of the effect of changes in political salience.

We solve the ensuing simultaneous-move game for its Nash equilibria. We focus especially on stable equilibria but examine also the properties of unstable equilibria as "threat points" to stable ones (determining their resilience, see below). In cases with low political salience, sanctioning concerns dominate expressive motives across the entire population; in the stable equilibria, citizens successfully coordinate and strategically conform to either all attacking or none attacking, thus avoiding sanctions altogether. In cases with high political salience, citizens act (almost) purely expressively, resulting in a unique equilibrium that is pinned down by the distribution of individual values. The most interesting cases lie in-between, where there can be a rich array of equilibria and variation across citizens in whether sanctioning concerns or expressive motives dominate. Due to the one-dimensional heterogeneity in political values and implied expressive motives, however, all equilibrium bandwagoning occurs on the same side of the spectrum: bandwagoning takes place either entirely among the supporters or entirely among the opponents of the regime.

We show that, generically, interior equilibrium comparative statics in salience are fully characterized by the side of the conflict on which bandwagoning arises,

political elites see Cinar and Nalepa [2022].

[3] For instance, Bernard C. Hennessy writes: "The bandwagon effect is supposed to induce the voter, regardless of party or other factors, to support a candidate simply because he appears to be a winner" [taken from a related experimental study by Fleitas, 1971, footnote 2]. See also Gallup and Rae [1940] on pre-election polls, highlighting a distinction between incentives and psychological aspects as explanations, or Simon [1954] on public election predictions that account for their own effects on the outcome; Schmitt-Beck [2015] offers a compact encyclopedia entry.

[4] Note that given asymmetries in sanctions, bandwagoning in our setting does not strictly require a side to be *winning* with greater probability but only to have a sufficiently large probability of winning for it to impose greater expected sanctioning.

together with whether the equilibrium is stable. An increase in salience strengthens expressive motives over sanctioning concerns across the board. The *direct "best-response" effect*—holding the assumed attack size constant—is always to reduce bandwagoning, because those that already acted in line with their values have now even more reason to do so. Thus, the numbers willing to attack increase when higher salience activates latent opposition and decrease when higher salience instead activates latent support.

The actual effect of a small increase in salience, given this direct effect, depends on whether the equilibrium is stable, however: stability determines the *indirect "equilibration" effect* that moves the attack size to rebalance the lower willingness to bandwagon with sanctioning concerns. With stable equilibria, the starting point must get attracted by the equilibrium after the change, so (myopic) best-response reasoning dynamics converge to it. Equilibration then reinforces the direct effect of reducing bandwagoning. By contrast, equilibration for unstable equilibria must counter the reduced willingness to bandwagon with an increase in actual bandwagoning and therefore greater risks of punishment for expressing true values, for the side of the conflict where there already was bandwagoning. Take, for instance, an equilibrium in which some intrinsic regime opponents are bandwagoning and defending the regime: higher political salience then means the attack grows if the equilibrium is stable, whereas the attack shrinks—producing greater punishment risk for attackers—if the equilibrium is unstable.

We then zoom in on "regime-optimal (stable) equilibria"—the stable equilibria with the smallest attack size—and study how changes in political salience alter not only the equilibrium size of the attack itself but also the size of collective deviations ("shocks") required to transition to a more threatening stable equilibrium. We take the latter as a measure of *regime resilience*, and it is determined by the unstable equilibrium that, in terms of attack size, lies in-between the regime-optimal stable equilibrium and the stable equilibrium with the next lowest attack size. The idea here is that any shock that pushes beliefs about how many (others) attack just above this size would lead citizens to reason their way to the still more threatening stable equilibrium. In this sense, the "adjacent" unstable equilibrium serves as a threat point to the regime-optimal stable equilibrium.

Our key findings relate to how changes in political salience and, hence, expressive motives, affect regime resilience. At low levels of salience, coordination incentives dominate, and the regime-optimal equilibrium has full coordination on "none attack." The direction of the effect of a rise in salience on resilience then depends on how weak or strong the regime's safeguards (sanctions on failed insurgents) are. If safeguards are weak, increases in salience from lower levels render this regime-optimal equilibrium less resilient, by producing more accessible threat points, because not only the equilibrium itself but also its threat point has bandwagoning by regime opponents, which—due to the threat point's unstable nature—expands upon a rise in salience. At middling levels of salience, the smallest shock can tip society and even give rise to a unique equilibrium with full coordination on the opposite extreme, where it is the regime supporters who collectively bandwagon against the regime for fear of sanctions by successful insurgents. Conversely, when regime safeguards are strong, increases in salience from low to middling ranges gradually shift out threat points, rendering the regime-optimal equilibrium more resilient and leading, possibly, to "none attack" as the unique equilibrium. In this case, there is bandwagoning by the entire opposition, all kept at bay by the threat of large regime sanctions. At yet higher levels of salience, expressive motives ultimately dominate any coordination incentives from sanctioning concerns. In this case, bandwagoning vanishes and there is a unique equilibrium involving social conflict between regime supporters and opponents, with an uncertain outcome.

There is thus a very simple message that arises from our analysis. Regime threats depend on the interplay between political salience and regime safeguards. Threats are greatest when safeguards are weak and salience increases from low to middling ranges. In these settings, small shocks that shift the composition of defenders versus attackers (e.g., through the rise of a small local anti-regime movement) suffice to activate otherwise latent regime opposition, which then gains further strength in its attack from bandwagoning by intrinsic supporters attempting to escape sanctions by a successful insurgency. If safeguards are strong, however, the same changes in salience can have opposite effects, further protecting regimes.

In our conclusion, we return to our main motivation, the role of political salience for the fate of democracies, relating our insights more specifically to it.

Related literature

We examine a model in the spirit of the classic accounts of Granovetter [1978] and Kuran [1989], in which a collection of players trade off the direct rewards and punishments of taking a stance against the intrinsic gains of acting in line with personal policy preferences over democratic and autocratic outcomes.

Medina [2007] gives perhaps the most comprehensive formal account of games of this form. We build on his work by providing analytic results on equilibria as a function of political salience for a heterogeneous population.

We can relate our contribution to work in the field with respect to the players, their preferences, and the information structure.

For players, our model keeps a focus on citizen action rather than elite behavior. Elite behavior has been a central motivation in the study of democratic backsliding. Much recent work focuses, for instance, on information or preference manipulation by regimes (Edmond [2013], and Grillo and Prato [2023]), or on effects of signals about the regime's vulnerability (Angeletos et al. [2006]). We do not doubt the importance of elite politics but focus on popular position-taking as a background condition for the success of elite strategies. Our results thus connect to contributions by Svolik [2019] and Miller [2021] on citizen attitudes and backsliding, and Carey et al. [2022] and Gidengil et al. [2022] on citizen support for backsliding elites.

We examine a setting in which payoffs depend on participation (in attack or defense of a regime) and on success (to avoid sanctions), but without assuming a prospect for players to be pivotal. This contrasts sharply with approaches such as those in Gieczewski and Kocak [2024] and Mutluer [2024] in which there are finite populations and actors focus on marginal effects of participation on success rates. In our setting, we instead assume a large population and calculations that are based on possibly conflicting payoffs from the act of participation itself. These include direct payoffs from participation similar to the psychological benefits studied, e.g., by de Mesquita and Shadmehr [2023] and Wood [2003], but they also include payoffs that depend on success. Unlike de Mesquita and Shadmehr [2023], we do not assume that these stem from the division of a fixed pie, however.

The relative weight placed on direct returns to participation in our model—which we term "salience"— derives from the psychology of bottom-up attention that is discussed and modeled in the survey of Bordalo et al. [2022]. Thus, we can consider salience as reflecting the public attention that the political conflict receives. It is also similar to a weight placed on civil duty as in Riker and Ordeshook [1968], though with an important difference: Our model features heterogeneity regarding whether such duty inspires attack or defense of the incumbent regime.[5]

Finally, ours is a setting in which information is symmetric. Although our model has connections with the recent literature on global games (Carlsson and van Damme [1993], Shadmehr and Bernhardt [2011]), these focus more specifically on information asymmetries and inference, which we bracket here. Information structures in global games typically serve the elimination of multiplicity; by contrast, there is inherent equilibrium multiplicity in our setting, and we are interested in characterizing how it is affected by changes in political salience as a structural feature.

2 Model and results

There is a unit mass of citizens ("players"), each deciding whether to defend or attack an incumbent regime. Each citizen i is identified with a location ϵ_i on the interval $[-1, \alpha]$, which denotes their idiosyncratic payoff from attacking (relative to defending) the regime, with $0 < \alpha < \infty$ and hence positive for some while negative for others. We refer to citizens with $\epsilon_i < 0$ as regime supporters, on account of their preferences and regardless of their actions, and refer to those with $\epsilon_i > 0$ as opponents. The distribution of citizens over such payoffs

[5] Though we do not explore this here, there are plausibly connections to the λ parameter in Medina [2007], at least to the extent that both of these capture weights placed on strategic considerations only or own actions only, with others' actions treated as fixed.

is given by an increasing and differentiable cdf F, with associated density f. When a mass $m \in [0,1]$ of citizens attack the regime, the probability that it is overthrown is $p(m)$, where $p(0) = 0$, $p'(m) > 0$ and $p(1) = 1$.[6] Let $\rho_A > 0$ (resp., $\rho_D > 0$) denote sanctions, which are the punishments imposed by winning attackers (resp., winning defenders) on citizens who have acted against them. Accordingly, the expected sanctioning from joining (resp., defending against) an attack of size m equals $(1 - p(m))\rho_D$ (resp., $p(m)\rho_A$).

We formalize *political salience* in this setting as the importance of idiosyncratic action payoffs relative to (expected) sanctions, via weighting parameter $\sigma \in [0,1]$. For an assumed attack size m, citizen i will attack rather than defend the regime if and only if:

$$\sigma\epsilon_i - (1-\sigma)(1-p(m))\rho_D \geq \sigma \cdot 0 - (1-\sigma)p(m)\rho_A.$$

Expressing expected sanctioning for attacking in net terms, this is equivalent to:

$$\sigma\epsilon_i + (1-\sigma)(\rho_D + \rho_A)\left(p(m) - \frac{\rho_D}{\rho_D + \rho_A}\right) \geq 0. \tag{1}$$

As any individual citizen's decision has a negligible effect in determining whether the regime survives, their action payoff captures heterogeneous expressive concerns, $\epsilon_i \in [-1, \alpha]$, whether to attack or to defend the regime. Salience measures how strongly motivated citizens are to act in line with their expressive concerns. We can see the moderating effect of σ on ϵ_i in the first term of (1). From the second term of (1) we can see that if $\sigma < 1$ the net expected gain from attacking is increasing in m. This captures coordination incentives arising from homogeneous sanctioning concerns: the more citizens take one side, the more likely it succeeds and hence the greater the expected sanctioning for taking the other side.

For a given level of salience σ, citizens face a potential trade-off between these coordination incentives (to minimize expected sanctioning) and their individual expressive motives (to act in line with their values). The former are the same for all citizens, pinned down by the punishment levels together with the size of the attack, while the latter vary in both sign and strength across the population according to the distribution F of political values $\epsilon_i \in [-1, \alpha]$. Changes in salience will affect this trade-off as they change the relative importance of expressive motives and sanctioning concerns.

[6] By mass m, we mean the Lebesgue measure of the subset of agents attacking, whenever it is a measurable one. The threshold structure of incentives in our model will allow us to deal with the issue of (strategy profiles implying) non-measurable subsets by assigning any non-measurable such subset an arbitrary success "probability" $q \in [0,1]$. However this is done, citizens' best-responding behavior always results in a measurable partition of citizens, corresponding to a threshold $t_q \in [-1, \alpha]$ such that citizens with $\epsilon_i > t_q$, measuring mass $1 - F(t_q)$, will attack. In this sense, the measurability issue pertains only to *defining* equilibrium without restricting to measurable strategy profiles, but not to equilibrium itself.

We will say that an individual is *bandwagoning* if coordination incentives dominate their individual expressive motives in this trade-off. That is, an individual is bandwagoning if they take an action that is counter to their political values—to "defend" rather than "attack" for citizens with $\epsilon_i > 0$, and vice versa for citizens with $\epsilon_i < 0$—because they expect the actions of others to render the risks of sanctioning too great should they act in line with their own political preferences.

A profile of actions is a Nash equilibrium if, given the actions of other players, no player has an incentive to change their own action. Let $\mu(m)$ denote the "attack response function," giving the share of players that weakly prefer to attack assuming a share m of players were to attack. (This will indeed be a function rather than a correspondence, except for a boundary case.) A Nash equilibrium then corresponds to a fixed point of μ, where $\mu(m) = m$. Graphically, the set of Nash equilibria is the set of intersections of $\mu : [0, 1] \to [0, 1]$ with the 45-degree line. Given the threshold structure of any equilibrium, which we formally establish in (3) below, we abuse terminology and identify any equilibrium simply by the corresponding attack size. We call an equilibrium m^* "stable" if there exists some $\delta > 0$ such that $|\mu(m) - m^*| < |m - m^*|$ for all $m \in [0, 1]$ with $0 < |m - m^*| < \delta$ (i.e., around m^*, μ is a local contraction). Otherwise we call it unstable. Stability requires local robustness of equilibrium, whereby at least in some neighborhood best-response dynamics—or corresponding equilibrium reasoning—should always converge (back) to it.

Many concrete applications may fit this general reduced-form model. Our main application, following our introductory motivation, assumes that the incumbent regime is democratic and describes citizens as choosing to attack or defend democracy vis-à-vis an autocratic agitator. We next sketch its concrete microfoundation from a standard median voter setting. Throughout what follows, our exposition—in particular, our terminology and interpretation—will emphasize it. The scope of our results extends beyond this specific application, however.

2.1 Application: Attacks on democracy

We provide a concrete/full microfoundation from a standard median voter setting for this application in Appendix A. It yields an interpretation of ϵ_i as reflecting citizens' single-peaked preferences over policy outcomes under the autocratic vs. the democratic regime (the latter resulting in the median ideal point). For the case of quadratic disutility from policy distance, we explicitly derive ϵ_i as i's (normalized) net policy gain under a successful autocratic attack. This is, of course, negative for a majority of citizens $F(0) > 0.5$. (The setting we use for our main illustrations in Section 2.5 indeed assumes this.)

To see this, let $u_i(A)$ be citizen i's policy payoff in case the autocratic attack on democracy is successful, and let it be $u_i(D)$ otherwise, so $\mathbb{E}u_i(m) = p(m)u_i(A) + (1 - p(m))u_i(D)$ is the expected policy payoff given mass m attack. This citizen will then attack if

$$\sigma u_i(A)+(1-\sigma)(\mathbb{E}u_i(m)-(1-p(m))\rho_D) > \sigma u_i(D)+(1-\sigma)(\mathbb{E}u_i(m)-p(m)\rho_A),$$

which is equivalent to

$$\sigma \cdot (u_i(A) - u_i(D)) + (1-\sigma)(p(m)\rho_A - (1-p(m))\rho_D) > 0.$$

Hence, ϵ_i in (1) here corresponds to the difference in policy payoffs $(u_i(A)-u_i(D))$, as the basis for expressive motives in driving political action.[7]

In view of this microfoundation, we can think of σ as capturing a form of *affective* polarization (Iyengar et al. [2019]), without political polarization (in the sense that policy preferences of all citizens remain the same). It measures how intensely the expressive value of action reflects the differences in outcomes that would arise when different groups control government and thus the stakes of political control (see also Chiopris et al. [2024] on platform divergence and attitudes to backsliding).

2.2 Equilibrium, stability and salience

We are ultimately interested in how threats to a democracy-optimal stable equilibrium, which has the smallest attack size among all stable equilibria, depend on political salience. To prepare this analysis, we first derive general results on the equilibrium set and the comparative statics of stable as well as unstable equilibria in salience. We begin by characterizing the boundary cases, which involve the following two population shares:

$$m_0 := p^{-1}\left(\frac{\rho_D}{\rho_D + \rho_A}\right) \text{ and } m_1 := 1 - F(0). \tag{2}$$

Note that both of these shares are interior.

Lemma 1. *Boundary cases:*

(i) If $\sigma = 0$, there exist three equilibria m^*, given by attack sizes $m^* = 0$ ("none attack"), $m^* = 1$ ("all attack"), and $m^* = m_0$. The two extreme equilibria are stable, while the interior equilibrium is unstable.

(ii) If $\sigma = 1$, there exists a unique equilibrium, given by attack size $m^* = m_1$. This interior equilibrium is stable.

Proof. (i) If $\sigma = 0$, this is a symmetric game with pure coordination incentives. From (1) and $p(m_0) = \rho_D/(\rho_D + \rho_A)$, the net expected gain from attacking

[7] This further illustrates how ϵ_i is indeed an expressive, psychological payoff. The policy outcome is fully determined by the aggregate actions wherein any individual i's action is negligible, so any separable (expected) payoff concerning this outcome "cancels out."

then equals $(\rho_D + \rho_A)(p(m) - p(m_0))$ for every citizen. Hence, all are indifferent between attacking and not attacking for $m = m_0$ (i.e., μ there takes the set-value $[0, 1]$), all prefer attacking for $m > m_0$ (i.e., $\mu(m) = 1$ for all such m), and all prefer not attacking for $m < m_0$ (i.e., $\mu(m) = 0$ for all such m). Since $0 < m_0 < 1$, we thus obtain the three equilibria as in the claim. To establish stability of the extreme equilibria, take $\delta < \min\{m_0, 1 - m_0\}$ in each case and observe that $m < \delta$ implies $\mu(m) = 0$ while $m > 1 - \delta$ implies $\mu(m) = 1$. To establish that the interior equilibrium at $m^* = m_0$ is unstable, note that for any $m \neq m_0$, $\mu(m) \in \{0, 1\}$, so hits the boundary.

(ii) If $\sigma = 1$, the net expected gain from attacking for citizen i equals simply ϵ_i, independent of how many others m attack. The "game" is then one with pure expression incentives. Hence, $\mu(m) = 1 - F(0) = m_1$ for all m, and $m^* = m_1$ is the unique equilibrium. It is clearly stable, since $|\mu(m) - m_1| = 0$ for all $m \in [0, 1]$. □

Case (i) has only sanctioning concerns and hence pure coordination incentives, which uniformly apply to all citizens. Accordingly, both extremes where all choose the same action—either "none attack" or "all attack," fully avoiding any sanctions—are stable equilibria. The unstable interior equilibrium has an attack size m_0 such that its implied success probability exactly equalizes expected sanctions from attacking vs. defending democracy, so that all citizens are indifferent as to their action. Sanctioning concerns alone favor attack on democracy whenever $m > m_0$ and favor defense whenever $m < m_0$. We will refer to m_0 as "countervailing-coordination" share/equilibrium. Note that to equalize expected sanctions across actions, greater democratic (resp., autocratic) sanctions require a larger attack (resp., defense), so they are balanced by a lower probability of success; i.e., m_0 is increasing in ρ_D as well as decreasing in ρ_A.

By contrast, the opposite boundary case (ii) has only expressive motives, hence pure expression incentives, because any strategic considerations due to sanctioning concerns are completely ignored. Accordingly, there is a unique and stable equilibrium, where all citizens simply express their political values by taking the corresponding side, corresponding to outright social conflict with an uncertain outcome. Exactly the share m_1 of intrinsic opponents to democracy attack, which we will refer to as "full-expression" share/equilibrium.

Neither of m_0 and m_1 depend on salience (see (2)). However, they will feature prominently in our results concerning equilibria for various levels of salience below.Given m_0 is increasing in democratic sanctions while m_1 measures the share of intrinsic opponents to democracy, pinned down entirely by the distribution of values, we will refer to democracy as having (relatively) *weak safeguards* when $m_0 < m_1$ and *strong safeguards* when $m_0 > m_1$. Note that in both boundary cases the overall democracy-optimal equilibrium is stable: it is the "none attack" equilibrium in (i) and "full expression" as the unique equilibrium in (ii).

Consider now intermediate cases with $\sigma \in (0, 1)$, in which citizens place weight both on the actions of others, through potential sanctions, and on their own action, through expressive motives related to their individual political values. To avoid clutter we will define $\tilde{\sigma} := \sigma/(1-\sigma) > 0$. Using the definition of m_0 in (2) to substitute $p(m_0)$ for $\rho_D/(\rho_D + \rho_A)$, citizen i is indifferent to taking part in the attack against democracy if and only if:

$$\epsilon_i = \frac{1}{\tilde{\sigma}}(\rho_D + \rho_A)(p(m_0) - p(m)) =: \tau(m). \tag{3}$$

For any assumed attack size $m \in [0,1]$, the function $\tau(m)$ yields the unique threshold such that all citizens with values above (resp., below) the threshold will attack (resp., defend) democracy. It is decreasing in the assumed attack size, meaning the more are expected to attack the more will actually attack, which reflects the coordination incentives from sanctioning concerns. In general, $\tau(m)$ may well fall outside of the support $[-1, \alpha]$ of the values distribution F, meaning all citizens would choose the same side when assuming that m attack. Threshold $\tau(m)$ is interior (to the support) if and only if:

$$\underline{m} := p^{-1}\left(\frac{\rho_D - \min\{\rho_D, \alpha\tilde{\sigma}\}}{\rho_D + \rho_A}\right) < m < p^{-1}\left(\frac{\rho_D + \min\{\rho_A, \tilde{\sigma}\}}{\rho_D + \rho_A}\right) =: \overline{m}. \tag{4}$$

The threshold structure implies that, for any commonly assumed attack size m, bandwagoners are all from the same side of the values spectrum. If $0 < \tau(m)$, expected sanctioning for attacking exceeds that for defending, so the bandwagoners are all (relatively weak) regime opponents, with values $\epsilon_i \in (0, \tau(m))$, who nevertheless defend the regime. If $\tau(m) < 0$, expected sanctioning for defending exceeds that for attacking, so the bandwagoners are (relatively weak) regime supporters, with values $\epsilon_i \in (\tau(m), 0)$, who nevertheless attack the regime. The effects of political salience on bandwagoning will be key to understanding its equilibrium effects.

Salience affects the threshold's responsiveness to the coordination incentives that arise from sanctioning concerns. From Equation 3 it is easy to see that at the countervailing-coordination share m_0 the "action indifferent" citizen coincides with the politically indifferent one—i.e., $\tau(m_0) = 0$, for any $\sigma > 0$. For other values of m, both left and right of m_0, increases in salience move $\tau(m)$ towards zero (that is, the value of the politically indifferent citizen), meaning more citizens will act in line with their political values. This captures the overall muting of coordination incentives by stronger expressive motives. Hence, the direct effect of rising salience is to reduce bandwagoning, regardless of the side of the political values spectrum on which it occurs.

The attack response function is then simply:

$$\mu(m) = 1 - F\left(\tau(m)\right). \tag{5}$$

Since τ is decreasing and F is non-decreasing, μ is non-decreasing. For assumed attack sizes m that imply an interior threshold $\tau(m) \in (-1, \alpha)$, $\mu(m)$ is interior, and μ is then actually increasing because F is. At the countervailing-coordination share m_0, where $\tau(m_0) = 0$, we have full expression of values as citizens' best response, so that $\mu(m_0) = m_1$, meaning no bandwagoning whatsoever. Increases in salience therefore pivot the attack response function clockwise around the point (m_0, m_1), as illustrated in Figure 1 (see Appendix D for details on this example).[8]

To understand this direct (or myopic) *best-response effect* of higher salience for any given m, recall that an increase in salience strengthens expressive motives relative to sanctioning concerns across the board. Those that were already willing to act in line with their political values have now all the more incentive to do so. But the trade-off for bandwagoners is affected: as the importance of coordination incentives relative to expressive motives is diminished, bandwagoning is reduced. Given a smaller attack size $m < m_0$, the interior threshold $\tau(m)$ is positive, meaning bandwagoning is by some weak opponents of democracy, some of whom will instead join the attack if salience is greater. This increases the attack size (i.e., $\mu(m)$ moves upward). Conversely, for a larger attack size $m > m_0$, bandwagoning is by weak supporters of democracy, and an increase in salience would similarly activate some of this latent support, in this case shrinking the attack (i.e., $\mu(m)$ moves downward). This summarizes the direct effects on willingness to attack $\mu(m)$ from (myopic) best responding behavior, fixing attack size m. Equilibrium comparative statics will depend on the indirect *equilibration effect* of how m changes to restore equilibrium given this direct change in incentives.

The attack response function μ is differentiable at any m where $\tau(m)$ is interior, so that also $\mu(m) \in (0, 1)$. In this case, we have:

$$\mu'(m) = -f(\tau(m))\tau'(m) = \frac{1}{\tilde{\sigma}}(\rho_D + \rho_A) f(\tau(m)) p'(m), \tag{6}$$

explicitly showing the strict increasingness of μ. Stability of an interior equilibrium $m^* = \mu(m^*) \in (0, 1)$ is then equivalent to $\mu'(m^*) < 1$, i.e., that μ crosses the 45-degree line from above at this equilibrium point.

For the analysis that follows we will rule out pathological (tangency) cases in which the slope of μ is exactly 1 at an interior equilibrium point.

Assumption 1 (Genericity)**.** *For any $m \in (0, 1)$, $\mu(m) = m$ implies $\mu'(m) \neq 1$.*

This assumption guarantees local uniqueness and well-defined comparative statics in salience of all interior equilibria, as differentiable functions $m^*(\cdot)$ of salience in a neighborhood of the initial σ. The assumption lets us simplify the exposition while otherwise remaining general with respect to the shapes of both the

[8] We are immensely grateful to an anonymous reviewer for suggesting a version of this figure and even providing us with code for it.

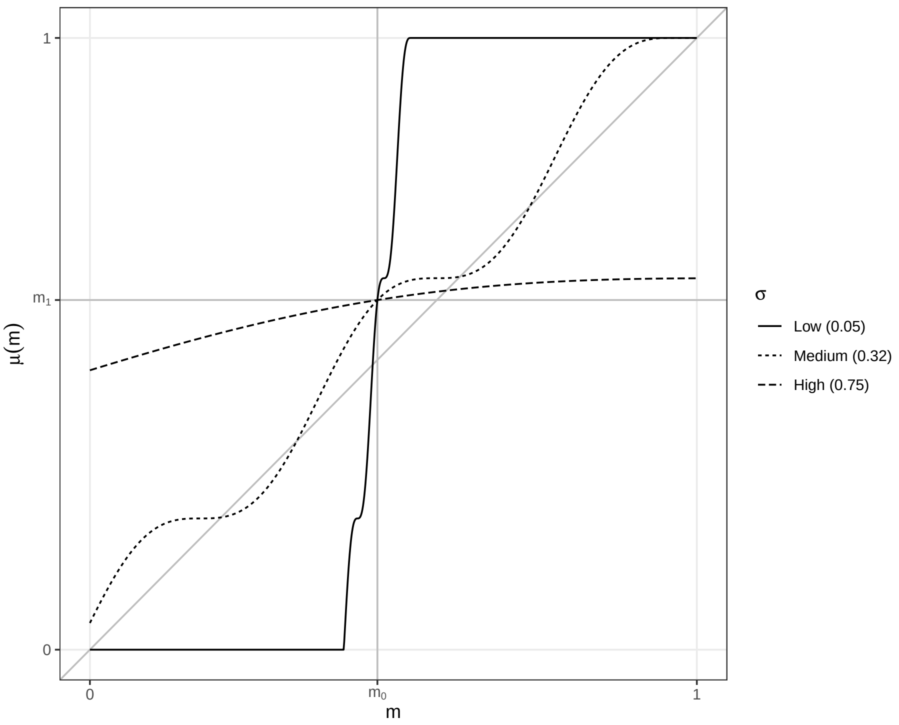

Figure 1: An increase in salience pivots the attack response function clockwise. The curved lines show three response functions for different salience levels, all passing through (m_0, m_1), which here lies above the 45-degree line. The two boundary cases' response functions would appear here as (i) a correspondence when $\sigma = 0$, going from (singleton) zero for $m < m_0$ to (singleton) one for $m > m_0$, while giving the whole interval $[0, 1]$ at $m = m_0$ (vertical line), and as (ii) the horizontal line at m_1 when $\sigma = 1$. Changes in salience always pivot the response function in the area in-between these two. Associated equilibria with the response functions are the intersections with the 45-degree line. The curve corresponding to the low σ value is close to the $\sigma = 0$ boundary case from Lemma 1 with two stable extreme equilibria and one unstable interior equilibrium slightly below m_0. The curve corresponding to the high σ value is close to the $\sigma = 1$ boundary case from Lemma 1 with one stable interior equilibrium slightly above m_1. The case of a medium σ value here exhibits multiple interior equilibria, together with an extreme equilibrium where all attack. See Appendix D for details underlying the figure as well as further illustration of equilibria as a function of salience.

distribution of values F and the success function p. It is not without loss of generality however. In Appendix B we explain what equilibrium configurations and corresponding "technicalities" we thus avoid, including also how these equilibria would be affected by changes in salience, and we relate the assumption to model primitives, in particular conceptualizing genericity in terms of salience levels. We highlight that our comparative statics results on σ and τ_D below relate to marginal effects at points where Assumption 1 is satisfied, and we do not seek to make claims about effects of discrete changes of parameters.

Our general results regarding comparative statics in salience of various equilibria, depending on their stability as well as attack size, are summarized in Proposition 1, which we prove in Appendix C.1. We provide illustrations in Section 2.5 and also in Appendix D. Given an equilibrium m^* for salience level σ, we write $\partial m^*/\partial\sigma$ for the marginal effect of changing salience on this equilibrium.

Proposition 1. *Given Assumption 1 and $\sigma \in (0,1)$:*

(i) A stable equilibrium exists. In particular:

(1) "None attack" is an equilibrium $m^* = 0$ if and only if $\tilde{\sigma} \leq \rho_D/\alpha$. It is stable if $\tilde{\sigma} < \rho_D/\alpha$, and in this case also satisfies $\frac{\partial m^*}{\partial \sigma} = 0$.

(2) "All attack" is an equilibrium $m^* = 1$ if and only if $\tilde{\sigma} \leq \rho_A$. It is stable if $\tilde{\sigma} < \rho_A$, and in this case also satisfies $\frac{\partial m^*}{\partial \sigma} = 0$.

(3) There exists a stable interior equilibrium $m^* \in (0,1)$ if $\tilde{\sigma} > \max\{\rho_D/\alpha, \rho_A\}$.

(ii) If $m_0 \neq m_1$, then there is no equilibrium in $[\min\{m_0, m_1\}, \max\{m_0, m_1\}]$. If $m_0 = m_1 = \hat{m}$, then $m^* = \hat{m}$ is an interior equilibrium; it is stable if and only if $\tilde{\sigma} > (\rho_D + \rho_A) f(\hat{m}) p'(\hat{m})$, and in any case satisfies $\frac{\partial m^*}{\partial \sigma} = 0$.

(iii) An interior equilibrium $m^* < m_1$ is stable if and only if $\frac{\partial m^*}{\partial \sigma} > 0$. An interior equilibrium $m^* > m_1$ is stable if and only if $\frac{\partial m^*}{\partial \sigma} < 0$.

Jointly with Lemma 1, Proposition 1 (i) establishes existence of a stable equilibrium, and in particular a *democracy-optimal* stable equilibrium with minimal attack size among all stable equilibria, for any salience level $\sigma \in [0,1]$. For low salience, in the sense of $\tilde{\sigma} < \rho_D/\alpha$, this equilibrium has "none attack." Given no one else attacks, doing so would surely fail and result in sanction ρ_D; at any salience $\tilde{\sigma} < \rho_D/\alpha$ even the most fervent opponent at $\epsilon_i = \alpha$ is then kept from expressing their opposition. A similar argument delivers also "all attack" as ("autocracy-optimal") stable equilibrium when salience is low, so that $\tilde{\sigma} < \rho_A$, while if salience is so high that neither of these extreme equilibria exists, there is always a stable interior equilibrium, and the democracy-optimal stable equilibrium entails risk that democracy gets overthrown.[9]

[9] Note that (1)–(3) of part (i) of Proposition 1 do not cover existence of a stable equilibrium in knife-edge cases where $\tilde{\sigma} = \max\{\rho_A, \rho_D/\alpha\}$, which are dealt with separately in the proof, see Appendix C.1.

Proposition 1 (ii) establishes a non-existence region for any $\sigma \in (0, 1)$, between the countervailing-coordination share m_0 and the full-expression share m_1. By Lemma 1, these attack shares correspond to the interior equilibria in the two boundary cases of pure coordination incentives ($\sigma = 0$) and pure expression incentives ($\sigma = 1$). To understand this, recall that an attack size equal to the countervailing-coordination share m_0 balances relative sanctioning exactly so that the action-indifferent/marginal citizen is also politically indifferent, i.e., $\tau(m_0) = 0$. The number of citizens willing to attack is hence pinned down entirely by expressive motives: $\mu(m_0) = 1 - F(0) = m_1$. Unless the setting is in the knife-edge case of $m_0 = m_1$, m_0 is not itself an equilibrium for any $\sigma \in (0, 1)$, and since willingness to attack increases in assumed attack from others around m_0, there is then no equilibrium in the entire range between m_0 and m_1 (these bounds included). This implies that any equilibrium features bandwagoning, on exactly one side of the political values spectrum. Moreover, in any interior equilibrium m^*, that side of the conflict also has some citizens express their (sufficiently strong) values through their action rather than bandwagoning, with the indifferent agent's value $\tau(m^*)$ interior and dividing this side. Given the distribution of values is continuous, even small changes in salience will shift the indifferent agent's value and thus how many are willing to attack, which drives the comparative statics in (iii), see below. The knife-edge case where $m_0 = m_1 = \hat{m}$ has $\mu(\hat{m}) = \hat{m}$, meaning full expression leads to the countervailing-coordination share, neutralizing sanction concerns and thus supporting itself. Graphically, the pivot point (m_0, m_1) of μ in salience then sits right on the 45-degree line, as $(m_0, m_1) = (\hat{m}, \hat{m})$, and hence constitutes an equilibrium. This is then true for any salience level (including boundary cases); its slope $\mu'(\hat{m})$ and hence stability depends on salience, however, according to (6), and in line with the stability properties of the two boundary cases' respective interior equilibria in Lemma 1.

Most importantly, Proposition 1 essentially characterizes equilibrium comparative statics in salience.[10] While the stable "none attack" and "all attack" equilibria when $\tilde{\sigma} < \rho_D/\alpha$ and $\tilde{\sigma} < \rho_A$, respectively, do not respond to marginal changes in salience, of course, a lesson of (i) is also that increases in σ from intermediate levels can destroy both of these (with the former disappearing first if $\rho_A < \rho_D/\alpha$). Intuitively, for sufficiently high weight placed on their expressive payoffs, the

[10] The qualification by "essentially" concerns not only Assumption 1 but also the "none attack" and "all attack" equilibria in their knife-edge cases of $\tilde{\sigma} = \rho_D/\alpha$ and $\tilde{\sigma} = \rho_A$, respectively. Their stability then depends on the local curvature of μ (see the proof of existence of a stable equilibrium when $\tilde{\sigma} = \max\{\rho_A, \rho_D/\alpha\}$ in Appendix C.1). However, in either knife-edge case, the extreme equilibrium is not a differentiable function of salience. The effect of increased salience on the "none attack" equilibrium when $\tilde{\sigma} = \rho_D/\alpha$ depends on its stability, increasing the attack size away from zero if stable—similar to interior stable equilibria $m^* < m_1$—but destroying the equilibrium if unstable, because unlike in interior unstable equilibria, there is no way to increase punishment risk for attackers further. By contrast, "none attack" remains in place with any lower salience, regardless of its stability, and becomes "stabilized" if initially unstable (see (4) for how the lower bound $\underline{m}$ on m for τ and hence μ to be interior depends on salience, starting from $\tilde{\sigma} = \rho_D/\alpha$). Effects on "all attack" when $\tilde{\sigma} = \rho_A$ are analogous (increased salience decreasing the attack size away from one if stable and destroying the equilibrium if unstable, and lower salience keeping it in place while "stabilizing" it; see upper bound $\overline{m}$ in (4)).

most extreme citizens on both sides would then refuse to bandwagon even if they were sure that they thus join the losing side and incur the sanctions. As a result, there then exist only interior equilibria (including a stable one), so there is certain conflict.

For such interior equilibria, the comparative statics in salience are tightly linked to their stability, via bandwagoning, as in (iii). Take any interior equilibrium $m^* < m_1$, stable or unstable. This equilibrium has fewer citizens attack democracy than there are intrinsic opponents to it, i.e., it features bandwagoning on the opposition side (and only there). Indeed, by (ii), also $m^* < m_0$, so the action-indifferent agent has an interior value on the opposition side, $\tau(m^*) \in (0, \alpha)$, and the bandwagoning opponents are those with "weaker" values $\epsilon_i \in (0, \tau(m^*))$. A (small) rise in salience focuses all citizens more on their expressive motives and hence intrinsic values. Holding the attack size constant, the supporters of democracy, all of whom were already defending it, would see their existing motivations to defend only bolstered, whereas this would activate (some of) the latent opposition that was bandwagoning to now attack instead. Graphically, τ flattens around $(m_0, 0)$, so the action-indifferent agent for attack size m^* becomes closer to political indifference, whereby μ pivots clockwise around (m_0, m_1) and therefore gets pushed upwards at $m^* < \min\{m_0, m_1\}$. The direct/myopic (best-response) effect of higher salience, given m^*, is thus to reduce opponents' bandwagoning and increase the attack. Similarly, any interior equilibrium $m^* > m_1$—where also $m^* > m_0$ by (ii)—has bandwagoning by (some) supporters of democracy, and the direct effect of a rise in salience is again to reduce bandwagoning, though here this implies a decrease in the attack.

Whether the equilibrium is stable matters, however, for how the attack size adjusts to maintain (or restore) equilibrium, given this direct effect of an increase in salience to reduce bandwagoning; we refer to this as the indirect equilibration effect. Intuitively, with stable interior equilibria, we can simply follow the (myopic) best-response dynamics to obtain equilibration: given a small increase in salience, the changed equilibrium contains the starting point in its basin of attraction, so those dynamics converge there. Graphically, a stable interior equilibrium m^* has $\mu'(m^*) < 1$, so the direct effect of a reduction in bandwagoning is reinforced by the dynamics, but less than one-for-one, implying convergence. Thus, a stable interior equilibrium with a small attack $m^* < \min\{m_0, m_1\}$ and bandwagoning by opponents of democracy has the attack grow as salience rises, whereas a stable interior equilibrium with a large attack $m^* > \max\{m_0, m_1\}$ and bandwagoning by supporters of democracy has the attack shrink as salience rises.

By contrast, unstable interior equilibria have the counter-intuitive property that a rise in salience promotes equilibrium bandwagoning, which is the opposite of the direct effect: to maintain equilibrium despite its instability, whereby best-response dynamics would diverge, citizens' increased interest in expressing political values must be countered by greater punishment risk on the very side that has bandwagoning, requiring yet more such bandwagoning. Graphically, an unstable interior equilibrium m^* has $\mu'(m^*) > 1$, so more bandwagoning is indeed required to offset the reduced willingness to doing so. For instance, take an unstable interior equilibrium $m^* < \min\{m_0, m_1\}$ with bandwagoning by opponents of democracy: being unstable means that the slightest increase in the assumed attack size above this equilibrium m^* would entail a discontinuous "explosion" in the attack size under best-response reasoning/dynamics, because $\mu'(m^*) > 1$; however, the strong responsiveness of best responses to the assumed attack size also means that equilibration in view of greater willingness to attack is obtained at a smaller attack with additional bandwagoning from opponents, which enhances expected sanctioning for attacking through a greater risk of failure. In such an unstable equilibrium, a rise in salience therefore leads to fewer and yet more fervent opponents who attack while facing greater punishment risk. Analogously, for unstable interior equilibria $m^* > \max\{m_0, m_1\}$ with bandwagoning by supporters of democracy, where a rise in salience leads to a larger attack, leaving democracy with fewer but more fervent supporters who still defend it under greater punishment risk.

For graphical intuition, see the attack response function for the medium value of salience in Figure 1, which oscillates around the 45-degree line and shows four interior equilibria: two with smaller attack sizes $m < m_1$, hence bandwagoning by opponents, and two with larger attack sizes $m > m_1$, hence bandwagoning by supporters. Moreover, within each pair, the equilibrium with the smaller attack is stable while that with the larger attack is unstable. Recalling that an increase in salience would pivot the attack response function around the point (m_0, m_1), where no one will bandwagon, it is easy to see that this would move the stable equilibrium below m_1 up the 45-degree line, and that above m_1 down the 45-degree line, while the opposite is true for the corresponding unstable equilibrium in each case.

A joint lesson of (ii) and (iii) then concerns the equilibrium effect of an increase in salience from low levels on the unique interior and unstable equilibrium that corresponds to (or emerges from) the countervailing-coordination equilibrium m_0 when $\sigma = 0$. At low levels of salience, this is indeed the unique unstable equilibrium, and it then coexists with the stable "none attack" and "all attack" equilibria (for $\tilde{\sigma} < \min\{\rho_D/\alpha, \rho_A\}$). Whether rising salience entails a growing or shrinking attack size in this interior equilibrium depends on the relative sizes of the countervailing-coordination and full-expression shares m_0 and m_1. Given there is no equilibrium in-between them, by (ii), there is only one direction for the unstable countervailing-coordination equilibrium m_0 to move as salience increases from $\sigma = 0$, and as long as this unstable interior equilibrium exists, the effects of salience on it are in line with the counter-intuitive effects on bandwagoning in such equilibria, as discussed above with case (iii). Specifically, if democracy has strong safeguards, in the sense of $m_0 > m_1$, then this equilibrium has $m^* > m_1$ and therefore bandwagoning by supporters of democracy; the attack grows as salience rises, due to a further increase in bandwagoning from supporters, more of whom attack democracy. This is illustrated in Figure 2 below (e.g., see its top left panel). For the same reason, if democracy's safeguards are weak, $m_0 < m_1$, then this equilibrium has $m^* < m_1$, hence bandwagoning by opponents of democracy, and sees the attack size shrink as salience rises, because more opponents then engage in bandwagoning and defend democracy. Figure 2 below also illustrates this case (e.g., see its top right panel).

2.3 Regime safeguards

As seen, essentially every equilibrium involves bandwagoning, on exactly one of the two sides. The bandwagoning stems from the conflict between sanctioning concerns and expressive motives tied to heterogeneous political values continuously distributed around political indifference. It occurs among citizens on the same side of the conflict because expected sanctions are the same for all citizens, incentivizing coordination on the action with lower expected sanctions, whereas expressive payoffs reflect sides, further confirming that action on one side while causing a trade-off on the other. The citizens that distinguish themselves as bandwagoners are then those whose values are sufficiently weak so that expressing them is not worthwhile in view of expected sanctions for doing so.

The following result shows the effect on an *interior* equilibrium of changing democratic safeguards in the form of its sanctions ρ_D against failed insurgents, as the policy tool in our model (for given ρ_A). Analogous to salience, $\partial m^*/\partial \rho_D$ denotes the marginal (comparative-statics) effect of changing democratic sanctions on a given interior equilibrium m^*. Assumption 1 guarantees that this is well-defined.

Proposition 2. *Given Assumption 1 and $\sigma \in (0, 1)$, an interior equilibrium $m^* \in (0, 1)$ satisfies $\frac{\partial m^*}{\partial \rho_D} < 0$ if it is stable and $\frac{\partial m^*}{\partial \rho_D} > 0$ if it is unstable.*

We relegate the proof to Appendix C.2. The intuition is straightforward, however. An increase in ρ_D directly renders attacking less attractive, for any assumed interior attack size $m \in (0, 1)$. Graphically, the attack response function shifts downward, as the threshold function τ shifts upward (though how far may vary with m).[11] Starting from any interior equilibrium m^*, a small increase in ρ_D thus implies that the share willing to attack, when assuming m^* others will, falls below m^*. Stability of m^* means $\mu'(m^*) < 1$, so this direct best-response effect is reinforced by the indirect equilibration effect, and the equilibrium attack therefore shrinks indeed. Conversely, instability of m^* means $\mu'(m^*) > 1$, so the expectation of a smaller attack would lead the attack to discontinuously "implode," which the indirect equilibration effect prevents by compensating reduced willingness to attack for given attack m^* with an increase in the actual attack above m^*.

The proposition focuses on interior equilibria as well as interior salience, because these will be most relevant to our considerations of democratic resilience below. For completeness, however, note from part (i) of Proposition 1 that: on the one

[11] To see this shift in the threshold function from (3), it is important to keep in mind that an increase in democratic sanctions ρ_D increases m_0. Increasing only total sanctions while keeping relative sanctions and hence m_0 constant would pivot τ similar to when salience *decreases*.

hand, ρ_D is irrelevant in any "all attack" equilibrium, including whether it exists (the probability of being punished by the democratic regime is then zero, as it is sure to lose, only autocratic sanctioning ρ_A matters and must be large enough to get even the strongest supporters of democracy to bandwagon); and, on the other hand, marginal changes to ρ_D could also generally not upset any stable "none attack" equilibrium, which prevails as long as $\tilde{\sigma} < \rho_D/\alpha$. The remaining case is then the knife-edge case of a "none attack" equilibrium when $\tilde{\sigma} = \rho_D/\alpha$. Here, a change in ρ_D has the same effect as the opposite change in salience; in particular, any increase in democratic safeguards ρ_D would clearly bolster such an equilibrium, in fact "stabilizing" it if previously unstable.[12] Concerning the boundary cases of salience, from Lemma 1, with $\sigma = 1$, any sanctioning is irrelevant, because citizens face pure expression incentives; with $\sigma = 0$, citizens face pure coordination incentives, so existence and stability of "none attack" and "all attack" do not depend on the sizes of any sanctions, while the unique interior equilibrium of countervailing coordination behaves in ρ_D just like any interior unstable equilibrium covered by Proposition 2, when $\sigma \in (0, 1)$.

2.4 Dynamic considerations and regime resilience

Although our model is static, much of the literature (e.g., Kuran [1997]) has been concerned with shifts between equilibria, which implies a dynamic conceptualization of the problem.

Our model speaks to these concerns to the extent that we think of agents adjusting attack behavior in a given period in response to aggregate attacks in the previous period. In this setting, at an equilibrium point, agents do not have incentives to adjust their behavior. Following a single-period *shock* to behavior, say from equilibrium m^* to attack $m = (m^* + \delta)$, the effects on next period's behavior, and movement toward or away from an equilibrium, can be read from the sign of $(\mu(m) - m)$.

Stability of an equilibrium is a local notion concerned with small shocks. It means that behavioral adjustments following a small shock lead society back to that equilibrium. Here, we will additionally consider a complementary notion of "resilience" of stable equilibria to larger shocks. Applied to democracy-optimal (stable) equilibria, it shall capture the latent danger of shifting to a higher attack equilibrium in the event of a non-negligible shock. An increase in salience may pose a threat to democracy—in particular, a stable "none attack" equilibrium—not only by directly moving equilibrium itself but also by making it less resilient.

The general idea is as follows: Take any two "adjacent" *stable* equilibria m' and $m'' > m'$, where by adjacent we mean that there is no further stable equilibrium in-between m' and m''. Under Assumption 1, there is, however, a unique unstable equilibrium $\tilde{m} \in (m', m'')$. We then refer to $\tilde{m}$ as the *threat point* of stable

[12] The effect of decreased ρ_D depends on stability, see the effect of increased σ in this knife-edge case in Footnote 10.

equilibrium m', and we take the distance $\tilde{m} - m' > 0$ to measure the *resilience* of m'. Any smaller shock moving the attack to $m < \tilde{m}$ would not seriously upset equilibrium m', since it is stable, whereas any larger shock resulting in an attack $m > \tilde{m}$ would lead society away from m' to a much increased attack size of (at least) *stable* equilibrium m'' in the longer run. We note that this need not be an actual dynamic adjustment process following a shock to how many actually attack but could also be a common "belief shock" regarding the imminent attack, which—depending on the shock size—immediately leads all the way to either stable m' or stable m''.

Applying this notion to a (clearly democracy-optimal) stable "none attack" equilibrium when $\tilde{\sigma} < \rho_D/\alpha$, resilience of this equilibrium captures *regime resilience*, because this equilibrium involves zero risk that the regime would be overthrown and is not directly affected by changes in salience within the whole range where $\tilde{\sigma} < \rho_D/\alpha$.[13] Proposition 1 implies the following result, concerning the effect of salience on democracy's resilience.[14]

Corollary 1. *Given Assumption 1, existence of a stable equilibrium $m'' > 0$, and salience σ such that $\tilde{\sigma} < \rho_D/\alpha$: "None attack" is a stable equilibrium $m' = 0$, there exists an unstable interior equilibrium $\tilde{m} \in (m', m'')$ as its threat point, and a marginal increase in salience renders "none attack" less resilient if $\tilde{m} < m_1$ and more resilient if $\tilde{m} > m_1$.*

Proof. "None attack" is a stable equilibrium for any salience σ such that $\tilde{\sigma} < \rho_D/\alpha$, by Proposition 1 (i) as well as Lemma 1 (i), as the claim covers the boundary case of $\sigma = 0$. Given "none attack" is a stable equilibrium $m' = 0$ and existence of another stable equilibrium $m'' > 0$, there exists $\delta > 0$ such that $0 < m < \delta$ implies $\mu(m) < m$ and $(m'' - \delta) < m < m''$ implies $\mu(m) > m$, by the proof of Proposition 1 (i) in Appendix C.1. Hence, μ then crosses the 45-degree line from below at some point between δ and $(m'' - \delta)$. By Assumption 1, there then exists an unstable interior equilibrium $\tilde{m} \in (m', m'')$ such that there are no equilibria in $(0, \tilde{m})$, whereby $\tilde{m}$ is the threat point of $m' = 0$. By Proposition 1, a marginal increase in salience does not affect "none attack" (i), whereas $\frac{\partial \tilde{m}}{\partial \sigma} < 0$ if $\tilde{m} < m_1$, and $\frac{\partial \tilde{m}}{\partial \sigma} > 0$ if $\tilde{m} > m_1$ (iii). □

A special case of this result applies in the boundary case when $\tilde{\sigma} = \sigma = 0$, for which Lemma 1 (i) characterizes equilibrium and which is covered by Corollary 1. The stable "none attack" equilibrium then has threat point $\tilde{m} = m_0$, and any shock coordinating beliefs onto an attack size above this level would lead

[13] Note that our general notion of resilience concerns resilience of a stable equilibrium, not the probability that a regime will be overthrown at that equilibrium. In our setting, if the stable equilibrium of interest were interior, then its (positive) attack and associated risk of regime defeat would directly vary with salience. A rise in salience might then simultaneously increase both the risk of regime defeat in that equilibrium and its resilience.

[14] Corollary 1 is silent on the knife-edge case when $m_0 = m_1 = \hat{m}$ and the interior equilibrium $\hat{m}$ is the threat point (hence unstable) of stable "none attack." It is immediate from Proposition 1 (i) and (ii) that a marginal increase in salience would then have no effect on either of the two equilibria, so resilience is unchanged.

to the stable "all attack" equilibrium. In view of Proposition 1 (i), a small increase in salience $\tilde{\sigma}$ does not affect the "none attack" equilibrium directly, since $\rho_D/\alpha > 0 = \tilde{\sigma}$. Yet, as discussed as a joint lesson of the proposition's parts (ii) and (iii), when safeguards are weak in the sense of $m_0 < m_1$, it brings this threat point $\tilde{m}$ closer, thus rendering "none attack" less resilient; when safeguards are strong, however, in the sense of $m_0 > m_1$, it moves $\tilde{m}$ further away and increases resilience. As indicated there already, this generalizes to any conflict settings with low salience, where the only difference in the equilibrium set from the boundary case of $\sigma = 0$ is that the interior unstable equilibrium is closer to an extreme, as the top right and top left panels of Figure 2 below illustrate.

More generally, even when safeguards are strong, there may be interior equilibria with attacks of size less than m_1, coexisting with stable "none attack" equilibria (and then necessarily including an unstable equilibrium as its threat point). Corollary 1 shows that, in such cases, rises in salience can entail ever diminishing resilience of democracy, up to the point where society tips from where none attack to a significant attack size.[15]

Note again that these comparative statics results are valid for small changes under Assumption 1. A discrete increase in σ could render "none attack" more resilient even if $\tilde{m} > m_1$. Indeed, if there are violations of Assumption 1 within the range of the increase in σ, the threat point could belong to a different equilibrium set or "none attack" may no longer exist as an equilibrium, as is the case for a change from $\sigma = 0$ to $\sigma = 1$.

Our simple model thus points out an essential risk to democracy from (small) increases in political salience especially when sanctions against insurgents are weak. While no attack whatsoever becomes apparent, ever smaller belief shocks would destroy it and may even move society to an "all attack" equilibrium, with certain autocracy (in the case where there is just one interior and hence unstable equilibrium). Proposition 2 then implies that strengthening democratic sanctions is effective in increasing democracy's resilience, by keeping threat points—which are unstable interior equilibria, by definition—distant.

2.5 Illustration

We illustrate using a case for which full analytic solutions are available. In Appendix D we provide additional illustrations for more complex examples.

We imagine linear $p(m) = m$ and uniform $\epsilon_i \sim U[-1, 0.5]$, so that $F(x) = \frac{2}{3}(x+1)$ and $f(x) = \frac{2}{3}$ for x in the support $[-1, 0.5]$. This pins down the full-expression

[15] If there are multiple interior equilibria with attacks smaller than m_1, coexisting with stable "none attack," then, in addition to the latter's threat point, these would include also its adjacent *stable* equilibrium. By Proposition 1, salience moves the threat point and this adjacent stable equilibrium in opposite directions. A rise in salience then not only means that smaller shocks suffice to break society away from none attacking, but also a greater risk that democracy is overthrown in the actual event of a sufficient shock, as the adjacent stable equilibrium has a larger attack than before. Figure 4 in Appendix D illustrates such a case.

share to $m_1 = \frac{1}{3}$, while the countervailing-coordination share $m_0 = \frac{\rho_D}{\rho_A+\rho_D}$ depends on sanctions, which we will vary. Democratic safeguards are weak in the sense of $m_0 < m_1$ if and only if $\rho_A > 2\rho_D$, and they are strong in the sense of $m_0 > m_1$ if and only if $\rho_A < 2\rho_D$.

Considering $\sigma \in (0,1)$, the threshold function τ is linear, because p is linear. For any m such that the threshold is interior, the attack response function μ is then linear as well, because also F is linear. Specifically, we have:

$$\begin{aligned}\mu(m) &= 1 - \frac{2}{3}(\tau(m)+1) = 1 - \frac{2}{3}\left(\frac{1}{\tilde{\sigma}}(\rho_D+\rho_A)(m_0-m)+1\right) \\ &= \frac{1}{3} - \frac{2}{3}\frac{1}{\tilde{\sigma}}(\rho_D+\rho_A)(m_0-m)\end{aligned}$$

if $m \in (\underline{m}, \overline{m})$, while $\mu(m) = 0$ if $0 \leq m \leq \underline{m} = (\rho_D - \min\{\rho_D, 0.5\tilde{\sigma}\})/(\rho_D+\rho_A)$ and $\mu(m) = 1$ if $1 \geq m \geq \overline{m} = (\rho_D + \min\{\rho_A, \tilde{\sigma}\})/(\rho_D+\rho_A)$.

"None attack" is an equilibrium if and only if $\tilde{\sigma} \leq 2\rho_D$, while "all attack" is an equilibrium if and only if $\tilde{\sigma} \leq \rho_A$; either is guaranteed to be stable whenever the respective inequality is strict. Because μ is linear, Assumption 1 here simply rules out that it overlaps the 45-degree line, which corresponds to the non-generic case where $\tilde{\sigma} = 2\rho_D = \rho_A$ (see Appendix B for details). Then, however, there is at most one interior equilibrium, and whenever it exists, it is given by:

$$m^* = \frac{0.5\tilde{\sigma} - \rho_D}{(0.5\tilde{\sigma} - \rho_D) + (\tilde{\sigma} - \rho_A)}.$$

We omit an explicit characterization of existence, which will become qualitatively clear below. However, it is immediate from the above that existence requires both of $(0.5\tilde{\sigma} - \rho_D)$ and $(\tilde{\sigma} - \rho_A)$ to have the same (non-zero) sign. Loosely, this means that σ must be either low or high. Indeed, as σ approaches a boundary, m^* approaches the corresponding interior equilibrium, unstable m_0 as $\sigma \to 0$ and stable m_1 as $\sigma \to 1$. Furthermore, taken as a function of salience σ (or $\tilde{\sigma}$), m^* is anyways decreasing if $2\rho_D < \rho_A$ ($m_0 < m_1$, weak safeguards) and increasing if $2\rho_D > \rho_A$ ($m_0 > m_1$, strong safeguards).

Equilibria are illustrated in Figure 2.[16] The figure confirms:

[16] A noteworthy property from this figure is that, in all cases, there is a unique branch continuously connecting the unique (and stable) full-expression equilibrium m_1 when $\sigma = 1$ to one of the extreme (and stable) pure coordination equilibria in the other boundary case when $\sigma = 0$. Which extreme equilibrium this is, follows immediately from the non-existence region: it is "none attack" if $m_1 < m_0$ and "all attack" if $m_1 > m_0$. This suggests a potential equilibrium selection argument in favor of the unique equilibrium on this branch [relatedly, see McKelvey and Palfrey, 1995]. However, the example considered in Appendix D, with a multimode distribution, shows that this argument cannot, in general, be straightforward: there, the unique connecting branch also bends backwards, meaning that for some values of σ, it would select multiple equilibria, including multiple stable ones but also unstable ones (see Figure 4).

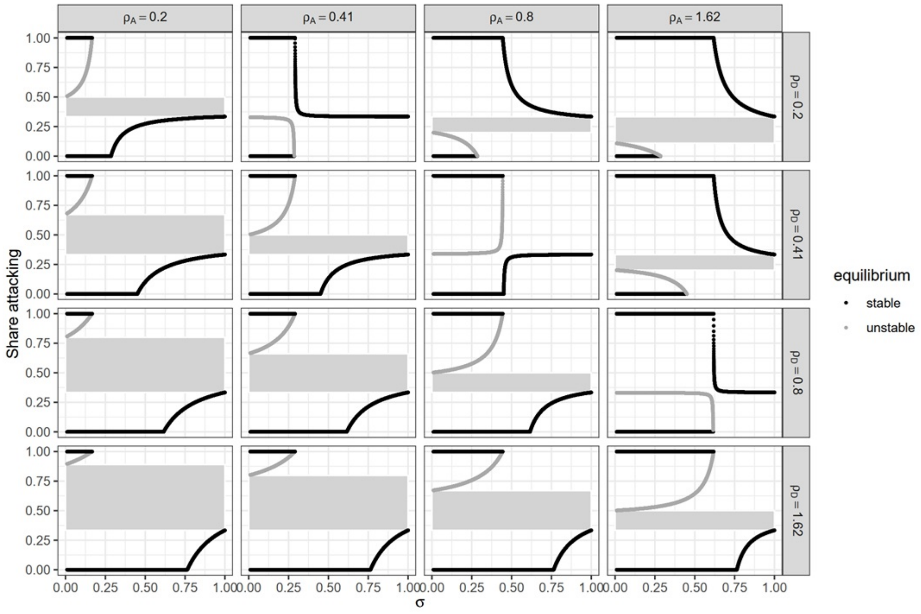

Figure 2: Equilibria in a linear-uniform model. Black points show stable equilibria. Dark grey points show unstable equilibria. There are no equilibria in the light-grey-shaded rectangular areas. Each panel varies salience $\sigma \in [0, 1]$ for given sanction values (ρ_A, ρ_D), and these vary between panels (ρ_A increases rightwards, ρ_D increases downwards). In the five upper right boundary panels, $2\rho_D < \rho_A$, meaning democratic safeguards are so weak that $m_0 < m_1$ and the unstable interior equilibrium is connected to the unstable countervailing-coordination equilibrium for $\sigma = 0$; it is then decreasing in salience σ until it hits the "none attack" equilibrium, where a further increase in salience tips society to the opposite extreme of "all attack" as a unique and stable equilibrium; at yet higher levels of salience, this equilibrium turns interior, connecting/limiting to the stable full-expression equilibrium for $\sigma = 1$. The other panels show mirror images of settings of strong safeguards, where $2\rho_D > \rho_A$.

1. Low salience always yields three equilibria, the two stable "none attack" and "all attack" equilibria as well as an unstable interior equilibrium connected to the unstable countervailing-coordination equilibrium under the pure coordination incentives with $\sigma = 0$; high salience always yields a unique equilibrium that is both stable and interior, connected to the stable full-expression equilibrium under the pure expression incentives with $\sigma = 1$.

2. Greater salience can increase or reduce risks of attack. In particular:

   - when democratic safeguards are weak, $m_0 < m_1$ (upper right region of Figure 2), an increase in salience from low to middling ranges renders "none attack" less and less resilient by pulling its threat point near and may eventually not only eliminate this equilibrium but yield a unique equilibrium where instead "all attack;"

   - when democractic safeguards are strong, $m_0 > m_1$ (lower left region of Figure 2), an increase in salience from low to middling ranges renders "none attack" more and more resilient by pushing its threat point away and may eventually turn "none attack" into a unique equilibrium.

While Figure 2 highlights effects of changing salience given sanctioning, Figure 3 illustrates the possible effects of changing democratic sanctions given salience, on interior equilibria, as established in Proposition 2. Fixing $\rho_A = 1$, it plots this equilibrium, which in our example is always unique (whenever it exists, see below), as a function of $\rho_D \in [0, 1]$, for different values of political salience σ. As the safeguards of democracy become stronger, the countervailing-coordination share m_0 increases, crossing that under pure expression, which is fixed at $m_1 = \frac{1}{3}$, at $\rho_D = \frac{1}{2}$.

The left panel illustrates three cases of low to middling salience, $\sigma < 0.5$. In these cases, whenever an interior equilibrium exists, it coexists with a stable "none attack" equilibrium and is the latter's threat point (unstable). The threat point gradually shifts upwards, thus becoming less accessible, with increases in ρ_D. This provides a non-trivial rationale for increasing democratic sanctions even in the democracy-optimal "none attack" equilibrium, to render it more resilient to shocks.

The right panel illustrates three cases of middling to high salience, $\sigma > 0.5$. In these cases, the (unique) equilibrium is stable and the attack size decreases in ρ_D. Here, increasing democratic sanctions directly affects the equilibrium and reduces the risk of a successful attack.

Critically, the effects of democratic sanctions depend on σ. For some ranges of σ, a small change in sanctioning costs can have dramatic strategic effects on bandwagoning, hence on resilience and the level of system defense, respectively. For instance, consider the two corresponding cases of middling salience in Figure 3, $\sigma = 0.495$ in the left panel and $\sigma = 0.505$ in the right panel. In other ranges,

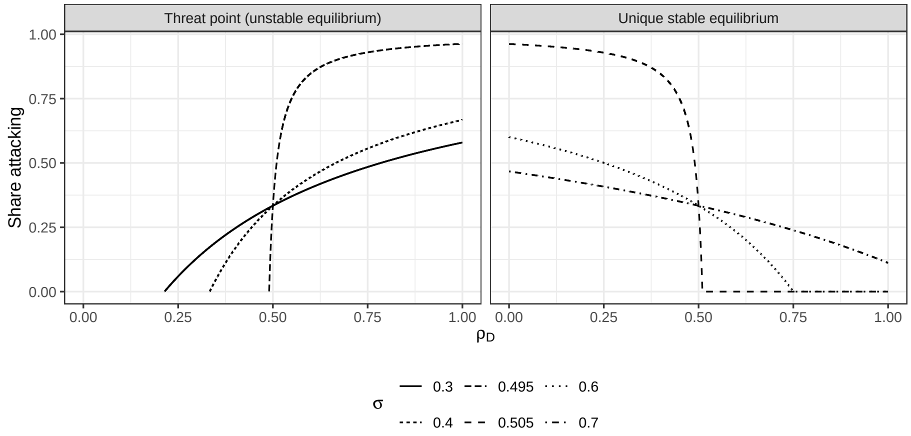

Figure 3: Illustration of (unique) interior equilibria as a function of $\rho_D \in [0, 1]$ when $\rho_A = 1$, for six different values of σ. The left panel has three values $\sigma < 0.5$, in which case this interior equilibrium is unstable, acts as a threat point to the coexisting stable “none attack” equilibrium, and is increasing in ρ_D. The right panel has three values $\sigma > 0.5$, in which case there is a unique equilibrium with the share attacking decreasing in ρ_D.

when σ is very low or very high, the effects of sanctions are likely very modest.

3 Conclusion

We study a simple model of attacks against regimes in a setting in which individuals differ in their desires to attack or defend institutions. Our key innovation is the consideration of an heterogeneous expressive utility component that is based on individuals’ intrinsic political preferences and scales with political salience. Our central results examine how changes in salience affect regime resilience. They hold for all regime-optimal stable equilibria, and, remarkably, for arbitrary distributions of policy preferences.

Applied to the potential challenges to democracy, our results suggest that when its safeguards are relatively weak, increases to middling levels of political salience can render democracies especially vulnerable. A heightened public focus on issues pertaining to democracy and more intense debates portrayed in mainstream and social media may, thus, become an Achilles heel for democracy. The intuition is that maintaining the democracy-optimal equilibrium relies on continued bandwagoning by latent opponents, who face sufficient punishment risk as long as sufficiently few other citizens attack. When sanctioning concerns are stronger for the anti-regime than the pro-regime equilibrium and increased salience renders sanctioning concerns generally less important, weakly safeguarded democracies may more easily tip. Thus, our model offers a lens through which the accounts of Putnam [2000] and Levitsky and Ziblatt [2018] of the dangers to democracy may be reconciled: disinterest in politics as low levels of political salience, in combination with institutional complacency in the form of little legal and executive safeguards, is exactly when increases in salience to middling levels and the resulting affective polarization put democracy under especially great risk of dramatically tipping in response to only small shocks.

In situations in which democracy's safeguards are strong, by contrast, increases in salience from low to middling levels can have the opposite effect of rendering democracies more resilient, preventing their intrinsic supporters from bandwagoning against the regime and instead continuing to actually express their support even when anticipating a serious attack. Yet, this can only be true up to a point, since increases in salience at high levels necessarily make it more difficult to keep opposition at bay. The intuition is that once sanctioning concerns are sufficiently muted by motives to express political values, some opponents are ready to express their opposition even in the democracy-optimal equilibrium. The indifferent agent is then an opponent who is indifferent only because of the threatened sanctions. An increase in political salience, further strengthening expressive motives over sanctioning concerns, thus shifts this agent to actively join the insurrection against the regime.

This finding has bearing on contemporaneous threats to democratic regimes. If citizens start to care more about political systems it may become important to bolster neglected safeguards for democracy and increase sanctions for insurgents. The long-run fate of democracies may, hence, be shaped by how governments react in the aftermath of events such as the attack on Capitol Hill. Our analysis suggests that leniency might generate heightened future threats.

References

George-Marios Angeletos, Christian Hellwig, and Alessandro Pavan. Signaling in a global game: Coordination and policy traps. *Journal of Political Economy*, 114(3):452–484, 2006.

Pedro Bordalo, Nicola Gennaioli, and Andrei Shleifer. Salience. *Annual Review of Economics*, 14:521—-544, 2022.

John Carey, Katherine Clayton, Gretchen Helmke, Brendan Nyhan, Mitchell Sanders, and Susan Stokes. Who will defend democracy? Evaluating tradeoffs in candidate support among partisan donors and voters. *Journal of Elections, Public Opinion and Parties*, 32(1):230–245, 2022.

Hans Carlsson and Eric van Damme. Global games and equilibrium selection. *Econometrica*, 61(5):989–1018, 1993.

Caterina Chiopris, Monika Nalepa, and Georg Vanberg. A wolf in sheep's clothing: Citizen uncertainty and democratic backsliding. *The Journal of Politics (just accepted)*, 2024. URL https://doi.org/10.1086/734253.

Ipek Cinar and Monika Nalepa. Mass or elite polarization as the driver of authoritarian backsliding? Evidence from 14 Polish surveys (2005–2021). *Journal of Political Institutions and Political Economy*, 3(3–4):433–448, 2022.

Ethan Bueno de Mesquita and Mehdi Shadmehr. Rebel motivations and repression. *American Political Science Review*, 117(2):734–750, 2023.

Chris Edmond. Information manipulation, coordination, and regime change. *Review of Economic Studies*, 80(4):1422–1458, 2013.

Daniel W Fleitas. Bandwagon and underdog effects in minimal-information elections. *American Political Science Review*, 65(2):434–438, 1971.

George Gallup and Saul Forbes Rae. Is there a bandwagon vote? *Public Opinion Quarterly*, 4(2):244–249, 1940.

Elisabeth Gidengil, Dietlind Stolle, and Olivier Bergeron-Boutin. The partisan nature of support for democratic backsliding: A comparative perspective. *European Journal of Political Research*, 61(4):901–929, 2022.

Germán Gieczewski and Korhan Kocak. Collective procrastination and protest cycles. *American Journal of Political Science (forthcoming)*, 2024.

Mark Granovetter. Threshold models of collective behavior. *American Journal of Sociology*, 83(6):1420–1443, 1978.

Edoardo Grillo and Carlo Prato. Reference points and democratic backsliding. *American Journal of Political Science*, 67(1):71–88, 2023.

Shanto Iyengar, Yphtach Lelkes, Matthew Levendusky, Neil Malhotra, and Sean J. Westwood. The origins and consequences of affective polarization in the united states. *Annual Review of Political Science*, 22:129–146, 2019.

Timur Kuran. Sparks and prairie fires: A theory of unanticipated political revolution. *Public choice*, 61(1):41–74, 1989.

Timur Kuran. *Private truths, public lies.* Harvard University Press, 1997.

Steven Levitsky and Daniel Ziblatt. *How Democracies Die.* Crown Publishing, New York, USA, 2018.

Richard D. McKelvey and Thomas R. Palfrey. Quantal response equilibria for normal form games. *Games and Economic Behavior*, 10(1):6–38, 1995.

Luis Fernando Medina. *A unified theory of collective action and social change.* University of Michigan Press, 2007.

Michael K. Miller. A republic, if you can keep it: Breakdown and erosion in modern democracies. *Journal of Politics*, 83(1):198–213, 2021.

Konuray Mutluer. Leading by example among equals. CERGE-EI Working Paper 791, October 2024.

Markus Prior and Lori D. Bougher. "Like they've never, ever seen in this country"? Political interest and voter engagement in 2016. *Public Opinion Quarterly*, 82(S1):822–842, 2018.

Robert D. Putnam. *Bowling alone: The collapse and revival of American community.* Simon and Schuster, 2000.

William H. Riker and Peter C. Ordeshook. A theory of the calculus of voting. *American Political Science Review*, 62(1):25–42, 1968.

Rüdiger Schmitt-Beck. Bandwagon effect. In *The International Encyclopedia of Political Communication*, pages 1–5. John Wiley & Sons, Ltd, 2015. URL https://doi.org/10.1002/9781118541555.wbiepc015.

Mehdi Shadmehr and Dan Bernhardt. Collective action with uncertain payoffs: Coordination, public signals, and punishment dilemmas. *American Political Science Review*, 105(4):829–851, 2011.

Herbert A. Simon. Bandwagon and underdog effects and the possibility of election predictions. *Public Opinion Quarterly*, 18(3):245–253, 1954.

Milan W. Svolik. Polarization versus democracy. *Journal of Democracy*, 30(3): 20–32, 2019.

Elisabeth Jean Wood. *Insurgent collective action and civil war in El Salvador.* Cambridge University Press, 2003.

Appendix

A Median voter setting

We show here how our reduced form model can be microfounded by a median voter setting. Let there be a one-dimensional (non-empty and bounded) policy space $[\underline{v}, \overline{v}]$ over which citizens have preferences that are characterized by their ideal points in this space, such that a citizen i with ideal point v_i evaluates policy $\widehat{v}$ with utility function

$$u\left(\widehat{v}|v_i\right) = \overline{u} - t \cdot \left|v_i - \widehat{v}\right|^2, \tag{7}$$

for some preference parameters $\overline{u} > 0$ (the political "bliss" value when $\widehat{v} = v_i$) and $t > 0$ (the sensitivity to deviations of $\widehat{v}$ from v_i). Let citizens' ideal points v_i be distributed over the policy space according to a distribution function (cdf) G that is increasing and differentiable. Under democracy, the policy outcome shall be the median voter's ideal point $v^D = G^{-1}(0.5)$; without loss, let the policy outcome under the alternative regime be some $v^A > v^D$, and define $w := \left(v^A + v^D\right)/2$. (The main text's brief illustration's payoffs $u_i(X)$ correspond to $u(v^X|v_i)$ here, for $X \in \{A, D\}$.)

It is straightforward to derive that, for any ideal point v_i,

$$u\left(v^A|v_i\right) - u\left(v^D|v_i\right) = 2t \cdot \left(v^A - v^D\right) \cdot \left(v_i - w\right).$$

This relative policy gain under a successful attack on democracy by the alternative regime is linearly increasing in a citizen's ideal point v_i, from a minimum of $2t \cdot (v^A - v^D) \cdot (\underline{v} - w) < 0$ to a maximum of $2t \cdot (v^A - v^D) \cdot (\overline{v} - w) > 0$. Mapping any ideal point v_i into ϵ_i as

$$\epsilon_i := \frac{1}{2t \cdot (v^A - v^D) \cdot (w - \underline{v})} \cdot \left(u(v^A|v_i) - u(v^D|v_i)\right) = \frac{v_i - w}{w - \underline{v}},$$

we have that the range of ϵ_i equals $[-1, \alpha]$ for $\alpha = (\overline{v} - w)/(w - \underline{v}) > 0$, and its distribution F on this support is easily derived as $F(x) = G(w + (w - \underline{v})x)$. Thus it inherits the strict increasingness and differentiability from G, and it has $F(0) > 0.5$, since $\epsilon_i = 0$ if and only if $v_i = w > v^D$.

It should be clear that a similar though significantly more tedious derivation of our reduced form can be obtained for any policy preferences such that the square in (7) gets replaced by some other exponent greater than one. The linear case is special in that it results in a distribution F with two atoms, one at each end of the support. This is because all citizens with ideal points $v_i \leq v^D$ then have the same (negative) relative policy gain of $-(v^A - v^D)$, and this is similarly true for all citizens with ideal points $v_i \geq v^A$, who all gain $(v^A - v^D)$. There may then arise equilibria in which an atom of citizens are indifferent and break their indifference in a particular way; clearly, however, no such equilibrium is stable, whereby the main insights from our analysis carry over.

B Genericity

We first explain what kind of equilibrium configurations Assumption 1 rules out to guarantee well-defined comparative statics in salience, clarifying what happens should these configurations arise. We then discuss the assumption's restrictions in terms of model primitives and use this to provide intuitions for when we might expect violations of Assumption 1.

Throughout, we consider $\sigma \in (0,1)$, of course. Recall that (i) interior equilibria $m^* \in (0,1)$ satisfy $m^* \in (\underline{m}, \overline{m}) \subseteq (0,1)$ for the bounds given in Equation (4), as explained following Equation (3), defining the threshold function τ, and that (ii) increases (resp., decreases) in σ pivot μ clockwise (resp., counter-clockwise) around the point (m_0, m_1), as explained following Equation 5, defining the attack response function μ, and illustrated in Figure 1. In particular, any increase in σ pushes μ upward or downward at m^* depending on whether $m^* < m_0$ or $m^* > m_0$, and the respective opposite is true for any decrease in σ.

B.1 Types of violations of Assumption 1

Consider now any "tangency segments" of μ with the 45-degree line, in which $\mu(m) = m$ and $\mu'(m) = 1$ for all $m \in [a,b]$, $0 < a \leq b < 1$. Let a and b such that the segment is maximal, meaning μ changes slope both as m approaches a from below and as m approaches b from above. This covers tangency "points" for $a = b$. We distinguish four broad cases for the effect of a marginal increase in σ:[17]

1. if tangency is without crossing and from above—i.e., if $\mu' < 1$ left of a and $\mu' > 1$ right of b—then equilibria (i) vanish if $m_0 > b$, (ii) split into two equilibria coming from a and b instead if $m_0 \leq a$ (the former now stable, with smaller attack unless $a = m_0$, and the latter still unstable, with larger attack), and resolve to m_0 if $m_0 \in (a,b]$ (now stable);

2. if tangency is without crossing and from below—i.e., if $\mu' > 1$ left of a and $\mu' < 1$ right of b—then equilibria (i) vanish if $m_0 < a$, (ii) split into two equilibria coming from a and b instead if $m_0 \geq b$ (the former still unstable, with smaller attack, and the latter now stable, with larger attack unless $b = m_0$), and (iii) resolve to m_0 if $m_0 \in [a,b)$ (now stable);

3. if tangency is with crossing from above to below—i.e., if $\mu' < 1$ both left of a and right of b—then equilibria resolve to a single equilibrium (i) coming from b if $m_0 \geq b$ (now stable, with larger attack unless $b = m_0$), (ii) coming from a if $m_0 \leq a$ (now stable, with smaller attack unless $a = m_0$), and (iii) equal to m_0 if $m_0 \in (a,b)$ (now stable);

4. if tangency is with crossing from below to above—i.e., if $\mu' > 1$ both left of a and right of b—then equilibria resolve to a single equilibrium (i) coming from a if $m_0 \geq b$ (still unstable, with smaller attack unless $a = b = m_0$),

[17] We omit the "mirror" effect of a marginal *decrease*.

(ii) coming from b if $m_0 \leq a$ (still unstable, with larger attack unless $a = b = m_0$), and (iii) equal to m_0 if $m_0 \in (a, b)$ (still unstable).

As main takeaway, Assumption 1 ensures local uniqueness of equilibria, ruling out intervals of equilibria, and that comparative statics in salience are well-defined, ruling out cases where marginal changes in salience would lead to "destruction" or "splitting" of an equilibrium. Though not a focus of our analysis in this paper, it is worthwhile mentioning that the definitions of (stable) democracy-optimal equilibrium and threat points do not depend on Assumption 1, and—as long as they exist—one may still compare democracy-optimal equilibria and their resilience for different levels of salience.

B.2 Primitive configurations that give rise to violations of Assumption 1

Turning now to model primitives, recall that Assumption 1 is concerned with (interior) equilibria $m \in (\underline{m}, \overline{m})$ such that:

$$\mu'(m) = -f(\tau(m))\tau'(m) = \frac{1}{\tilde{\sigma}}(\rho_D + \rho_A)f(\tau(m))p'(m) = 1.$$

How many such points there are that violate this condition depends on the curvature of μ, which is determined by the curvatures of F and p.

If μ is strictly concave (resp., strictly convex), then there is at most a single interior equilibrium point $m^* = \mu(m^*) \in (0, 1)$, and there is also at most a single point m such that $\mu'(m) = 1$. In either case, increases in $\tilde{\sigma}$ rotate μ around (m_0, m_1), so there could only be a single value of salience $\tilde{\sigma}$ at which those two points would coincide and the unique interior equilibrium is exactly tangent to the 45-degree line. One way to satisfy the genericity assumption would then be to simply rule out this single "non-generic" value of salience.

Supposing, for simplicity, that both F and p are twice differentiable, whereby τ is so as well, we have that, for $m \in (\underline{m}, \overline{m})$:

$$\begin{aligned}
\mu''(m) &= -(f'(\tau(m))(\tau'(m))^2 + f(\tau(m))\tau''(m)) \\
&= -f(\tau(m))\tau'(m)\left(\frac{f'(\tau(m))}{f(\tau(m))}\tau'(m) + \frac{\tau''(m)}{\tau'(m)}\right) \\
&= \mu'(m)\left(\frac{f'(\tau(m))}{f(\tau(m))}\tau'(m) + \frac{p''(m)}{p'(m)}\right).
\end{aligned}$$

Sufficient conditions for strict concavity of μ (i.e., $\mu'' < 0$) for any $\tilde{\sigma} > 0$ would be that F is convex and p is concave, with one being strictly so. Analogously, F concave and p convex, again with one being strictly so, would imply strict convexity of μ (i.e., $\mu'' > 0$) for any $\tilde{\sigma} > 0$.

If both F and p are linear, as in the linear-uniform illustration of Section 2.5, then μ is linear, with $\mu'(m) = \frac{1}{\tilde{\sigma}} \frac{\rho_D + \rho_A}{1+\alpha}$. The only way the genericity assumption could be violated would be if μ overlapped the entire 45-degree line, meaning all $m \in [0,1]$ have $\mu(m) = m$ and are equilibria. (Note that in this case no stable equilibrium exists.) Since $\mu(m_0) = m_1$, genericity is then violated if and only if both $m_0 = m_1$ and $\tilde{\sigma} = (\rho_D + \rho_A)/(1+\alpha)$, which boils down to $\rho_D/\alpha = \rho_A = \tilde{\sigma}$. Hence, in a linear-uniform model, the genericity assumption rules out only this knife-edge case.

More generally, how many points $m \in (\underline{m}, \overline{m})$ there are such that $\mu'(m) = 1$ is determined by the number of inflection points of μ, and the same is true about the number of interior equilibria (see Figure 1 and Appendix D for illustrations with multiple inflection points); thus, the relation of the curvatures of F and p determines the potential for violations of Assumption 1 at certain "non-generic" values of $\tilde{\sigma}$ to be ruled out.

C Omitted proofs

C.1 Proof of Proposition 1

Proof. The proof follows the structure of the proposition.

(i) Since μ is a continuous mapping from the compact interval $[0,1]$ to itself, existence of equilibrium follows from Brouwer's fixed-point theorem. The stronger claim that a stable equilibrium exists follows from 1.–3., except for knife-edge cases, which we consider as 4. below:

1. Equilibrium $m^* = 0$ for $\tilde{\sigma} \leq \rho_D/\alpha$, and its stability as well as $\frac{\partial m^*}{\partial \sigma} = 0$ for $\tilde{\sigma} < \rho_D/\alpha$: Note that $\mu(0) = 1 - F(\tau(0)) = 0$ if and only if $\tau(0) \geq \alpha$, which is equivalent to $\rho_D/\alpha \geq \tilde{\sigma}$. Intuitively, for the most democracy hating person ($\epsilon_i = \alpha$), the psychological reward from attacking $\tilde{\sigma}\alpha$ is less than the certain punishment ρ_D. In case of strict inequality $\tilde{\sigma} < \rho_D/\alpha$, we have $\tau(0) > \alpha$, so by the continuity of τ in m, there is a $\delta > 0$ such that $\mu(m) = 0$ also for any $m \in (0, \delta)$, then implying stability. For marginal changes in salience, the strict inequality prevails, and so does therefore the equilibrium $m^* = 0$, i.e., $\frac{\partial m^*}{\partial \sigma} = 0$.

2. Equilibrium $m^* = 1$ for $\tilde{\sigma} \leq \rho_A$, and its stability $\frac{\partial m^*}{\partial \sigma} = 0$ for $\tilde{\sigma} < \rho_A$: Analogous to 1. above.

3. Stable interior equilibrium if $\tilde{\sigma} > \max\{\rho_A, \rho_D/\alpha\}$: From $\tilde{\sigma} > \rho_D/\alpha$ we have $\tau(0) < \alpha$ and hence $\mu(0) > 0$; analogously, $\tilde{\sigma} > \rho_A$ implies $\tau(1) > -1$ and hence $\mu(1) < 1$. Given this, the continuous function μ must *cross the 45-degree line from above* at some interior point $m \in (0,1)$, which is then a stable equilibrium.

4. Existence of a stable equilibrium if $\tilde{\sigma} = \max\{\rho_A, \rho_D/\alpha\}$. We distinguish three possible cases:

(a) If $\tilde{\sigma} = \rho_D/\alpha > \rho_A$, then $\mu(0) = 0$ and $\mu(1) < 1$. The genericity from Assumption 1 implies that if the one-sided derivative of μ at zero equals one, $\mu'_+(0) = 1$, then μ' changes when m increases: otherwise, there would be $\delta > 0$ such that $\mu(m) = m$ and $\mu'(m) = 1$ for interior $m \in (0, \delta)$, which the assumption rules out. If $\mu'_+(0) < 1$ or $\mu'_+(0) = 1$ with μ' locally decreasing as m increases from 0, then the "none attack" equilibrium is stable, establishing the claim. If $\mu'_+(0) > 1$ or $\mu'_+(0) = 1$ with μ' locally increasing as m increases from 0, then the "none attack" equilibrium is unstable. However, the latter case implies some small $\delta > 0$ such that $\mu(\delta) > \delta$, and since $\mu(1) < 1$, there exists a stable interior equilibrium in $(\delta, 1)$, by the argument given in the previous paragraph.

(b) If $\tilde{\sigma} = \rho_A > \rho_D/\alpha$, then $\mu(0) > 0$ and $\mu(1) = 1$, so the proof is analogous. If $\mu'_-(1) = 1$, then μ' changes when m decreases, by Assumption 1; if $\mu'_-(1) < 1$ or $\mu'_-(1) = 1$ with μ' locally increasing in

m as m increases towards 1, then the "all attack" equilibrium is stable, and otherwise there is some small $\delta > 0$ such that $\mu(1-\delta) < 1-\delta$, which together with $\mu(0) > 0$ implies existence of a stable interior equilibrium in $(0, 1-\delta)$.

(c) If $\tilde{\sigma} = \rho_D/\alpha = \rho_A$, then $\mu(0) = 0$ and $\mu(1) = 1$, and if neither of these extreme equilibria is stable, we can combine the above two cases' corresponding arguments for existence of some small $\delta \in (0, 1/2)$ such that both $\mu(\delta) > \delta$ and $\mu(1-\delta) < 1-\delta$, whereby μ must cross the 45-degree line from above at some point in $(\delta, 1-\delta)$.

(ii) Recall first that $\tau(m_0) = 0 \in (-1, \alpha)$ and hence $\mu(m_0) = m_1 \in (0,1)$ holds true for any value of $\tilde{\sigma}$. Hence the interior m_0 is an equilibrium if and only if $m_0 = m_1 = \hat{m}$. An equilibrium $m^* = \hat{m}$ does not depend on $\tilde{\sigma}$, whereby $\frac{\partial m^*}{\partial \sigma} = 0$; it is stable if and only if $\mu'(\hat{m}) < 1$, which, from (6), is equivalent to the condition given.

Now suppose $m_0 < m_1$. Since $\mu'(m_0) > 0$ while μ is generally nondecreasing, we then have that $m \in (m_0, m_1]$ implies $\mu(m) > \mu(m_0) = m_1 \geq m$. Hence there is no equilibrium in $[m_0, m_1]$. The case of $m_0 > m_1$ is analogous; then, $m \in [m_1, m_0)$ implies $\mu(m) < \mu(m_0) = m_1 \leq m$, hence no equilibrium in $[m_1, m_0]$.

(iii) Take any interior equilibrium point $m^* \in (0,1)$ given (normalized) salience $\tilde{\sigma}$, i.e., such that:

$$m^* = \mu(m^*) = 1 - F\left(\frac{1}{\tilde{\sigma}}(\rho_D + \rho_A)(p(m_0) - p(m^*))\right).$$

By Assumption 1, $\mu'(m^*) \neq 1$, so the conditions of the Implicit Function Theorem are satisfied; i.e., for some small $\delta > 0$, there is a unique differentiable function $g : (\tilde{\sigma} - \delta, \tilde{\sigma} + \delta) \to (0,1)$ such that $g(\tilde{\sigma}) = m^*$ and, for all $s \in (\tilde{\sigma} - \delta, \tilde{\sigma} + \delta)$:

$$\begin{aligned} g(s) &= 1 - F\left(\frac{1}{s}(\rho_D + \rho_A)(p(m_0) - p(g(s)))\right) \\ &= 1 - F\left(\frac{\tilde{\sigma}}{s}\tau(g(s))\right). \end{aligned}$$

Since $\tilde{\sigma}$ is increasing in σ, we obtain the sign of $\partial m^*/\partial \sigma$ from that of $g'(\tilde{\sigma})$, differentiating the above and evaluating at $\tilde{\sigma}$, where $g(\tilde{\sigma}) = m^*$:

$$
\begin{aligned}
& g'(s) = f\left(\frac{\tilde{\sigma}}{s}\tau(g(s))\right)\left(\frac{\tilde{\sigma}}{s}\frac{\tau(g(s))}{s} - \frac{\tilde{\sigma}}{s}\tau'(g(s))g'(s)\right) \\
\implies & g'(\tilde{\sigma}) = f\left(\tau(m^*)\right)\left(\frac{\tau(m^*)}{\tilde{\sigma}} - \tau'(m^*)g'(\tilde{\sigma})\right) = f\left(\tau(m^*)\right)\frac{\tau(m^*)}{\tilde{\sigma}} + \mu(m^*)g'(\tilde{\sigma}) \\
\implies & g'(\tilde{\sigma}) = \frac{f\left(\tau(m^*)\right)\tau(m^*)/\tilde{\sigma}}{1-\mu(m^*)}.
\end{aligned}
$$

If $m^* < m_1$, then $\tau(m^*) > 0$ because $m^* < m_0$ by (ii); from the above, we then have $\partial m^*/\partial\sigma > 0$ if $\mu'(m^*) < 1$, i.e., m^* is stable, and we have $\partial m^*/\partial\sigma < 0$ if $\mu(m^*) > 1$, i.e., m^* is unstable. Since $\mu'(m^*) = 1$ is ruled out by Assumption 1, $\partial m^*/\partial\sigma > 0$ if and only if m^* is stable. If $m^* > m_1$, then $\tau(m^*) < 0$ because $m^* > m_0$ by (ii), whereby $\partial m^*/\partial\sigma$ has the opposite sign given stability/instability; i.e., $\partial m^*/\partial\sigma < 0$ if and only if m^* is stable.

This concludes the proof of Proposition 1. □

C.2 Proof of Proposition 2

Proof. Take any interior equilibrium m^* for given democratic sanctions ρ_D. Similar to the proof of part (ii) of Proposition 1, due to Assumption 1, we can apply the Implicit Function Theorem, defining a differentiable function h that satisfies $h(\rho_D) = m^*$ and uniquely describes how equilibrium changes with democratic sanctions in a neighborhood of ρ_D:

$$
\begin{aligned}
h(r) &= 1 - F\left(\frac{1}{\tilde{\sigma}}(r+\rho_A)\left(\frac{r}{r+\rho_A} - p(h(r))\right)\right) \\
&= 1 - F\left(\frac{1}{\tilde{\sigma}}\left(r - (r+\rho_A)p(h(r))\right)\right).
\end{aligned}
$$

Differentiating h and evaluating the derivative h' at $r = \rho_D$, where $h(\rho_D) = m^*$, and then rearranging, yields $\partial m^*/\partial\rho_D$, i.e., as:

$$
\begin{aligned}
& h'(\rho_D) = -f(\tau(m^*))\frac{1-p(m^*)}{\tilde{\sigma}} + \mu(m^*)h'(\rho_D) \\
\implies & h'(\rho_D) = \frac{-f(\tau(m^*))(1-p(m^*))/\tilde{\sigma}}{1-\mu(m^*)}.
\end{aligned}
$$

Given an interior m^*, the numerator is negative. The denominator is equal to $1-\mu'(m^*)$, so it is positive if m^* is stable, and negative if m^* is unstable. Hence, $\partial m^*/\partial\rho_D < 0$ if m^* is stable, and $\partial m^*/\partial\rho_D > 0$ if m^* is unstable. □

D Additional example and visualization

Figure 1 shows how the attack response function pivots clockwise around (m_0, m_1) as salience σ increases from a low (0.05) to a medium (0.32) to a high (0.75) level, and also the corresponding equilibria. Underlying it is a setting with a multimodal distribution of political values. Specifically, we assume a distribution over the interval $[-1, 1]$ with cdf $F(x) = \frac{1}{2}((x+1) - (\cos(8x) - \cos(8))/8)$ and hence density $f(x) = \frac{1}{2}(1 + \sin(8x))$, together with a linear success function $p(m) = m$, and sanction levels that are fixed at $\rho_D = 0.45$ and $\rho_A = 0.5$. This implies $1 - F(0) = m_1 > m_0 = \rho_D/(\rho_D + \rho_A)$, so weak safeguards.

Figure 4 additionally shows the full set of equilibria for this example as a function of salience over the range $\sigma \in [0, 0.6]$. This includes the low and medium levels of salience considered in Figure 1, thus detailing how the three equilibria in the boundary case of $\sigma = 0$ change in the lower range, until additional (stable as well as unstable) interior equilibria emerge and subsequently the extreme equilibria vanish in the middling range. We omit higher values of σ with a unique and stable interior equilibrium that quickly approaches full expression m_1 (e.g., see Figure 1 for the high value omitted here), as the corresponding equilibrium in the boundary case of $\sigma = 1$. Relative to Figure 2, Figure 4 additionally illustrates how resilience of the democracy-optimal equilibrium is affected by rising salience for a setting that allows for multiple interior equilibria below m_1, as discussed specifically in Footnote 15.

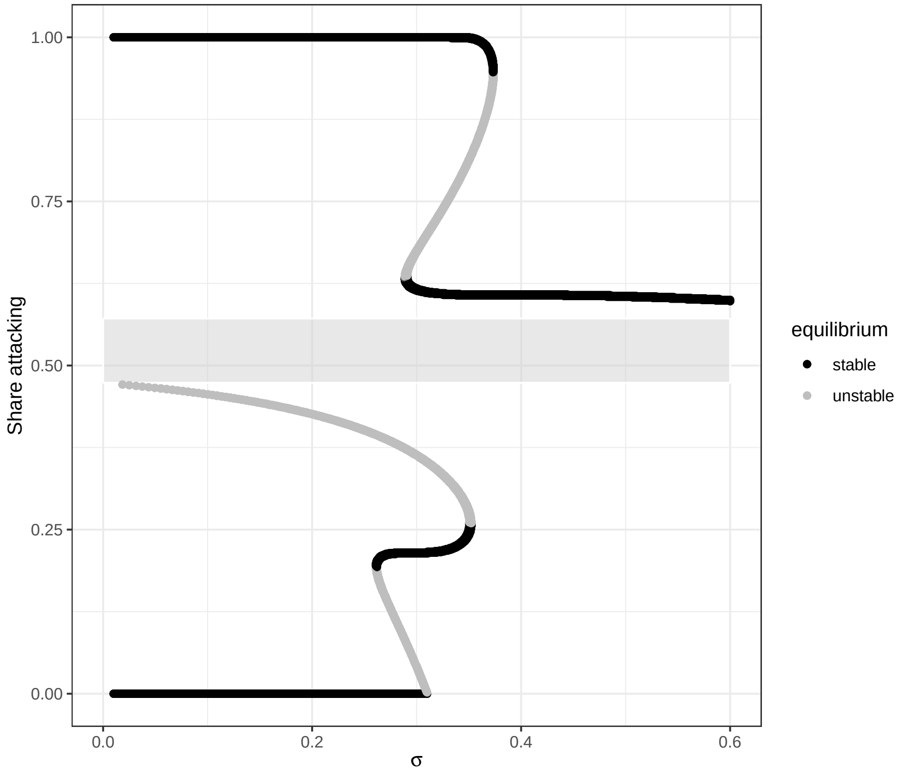

Figure 4: Equilibria as a function of σ over the range $[0, 0.6]$ for the same example as in Figure 1. Black points show stable equilibria. Dark grey points show unstable equilibria. There are no equilibria in the light-grey-shaded rectangular area.