Schweighofer-Kodritsch, Sebastian; Huck, Steffen; Humphreys, Macartan

Working Paper

Political Salience and Regime Resilience

CESifo Working Paper, No. 12116

Provided in Cooperation with:
Ifo Institute – Leibniz Institute for Economic Research at the University of Munich

Suggested Citation: Schweighofer-Kodritsch, Sebastian; Huck, Steffen; Humphreys, Macartan (2025) : Political Salience and Regime Resilience, CESifo Working Paper, No. 12116, Munich Society for the Promotion of Economic Research - CESifo GmbH, Munich

This Version is available at:
https://hdl.handle.net/10419/331582

CES ifo Working Papers

12116
2025

September 2025

Political Salience and Regime Resilience

Sebastian Schweighofer-Kodritsch, Steffen Huck, Macartan Humphreys

CES ifo

Imprint:

CESifo Working Papers

ISSN 2364-1428 (digital)

Publisher and distributor: Munich Society for the Promotion
of Economic Research - CESifo GmbH

Poschingerstr. 5, 81679 Munich, Germany
Telephone +49 (0)89 2180-2740

Email office@cesifo.de
https://www.cesifo.org

Editor: Clemens Fuest

An electronic version of the paper may be downloaded free of charge
· from the CESifo website: www.ifo.de/en/cesifo/publications/cesifo-working-papers
· from the SSRN website: www.ssrn.com/index.cfm/en/cesifo/
· from the RePEc website: https://ideas.repec.org/s/ces/ceswps.html

Political salience and regime resilience*

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

June 7, 2024

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 concerns with material bandwagoning incentives. We examine comparative statics in salience that fully characterize the stability of equilibria. A main insight is that when regime sanctions are weak, increases from low to middling salience can pose the greatest threat to regimes – even very small shocks can suffice to drastically escalate attacks. Our results speak to the charged debates about democracy, by identifying conditions under which heightened interest in political decision-making can pose a threat to democracy in and of itself.

JEL Classification: C72, D74, D91

Keywords: political conflict, salience, democracy, sanctions

*Our thanks to Daniel Markovits, Ethan Bueno de Mesquita, 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* assessing how salience, in and of itself, can contribute to the vulnerability of political regimes.

Salience encapsulates how much voters care about the type of regime that prevails which we will model by introducing an expressive component into voters' utility function that is scaled by a common parameter capturing different political climates where voters may care more or less about taking a stance. Dynamic changes in public debates, how heated they are and their prominence in mainstream or social media, can change salience rendering the expressive component more or less impactful.

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 the rising stakes of political decision-making (that is, increasing salience). Levitsky and Ziblatt [2018], for instance, describe the erosion of democratic norms as politics become polarized and conflicts more total. There are thus straightforward, if conflicting, logics through which changes in the salience of politics can threaten political regimes.[1]

In this short 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 citizens individually decide whether to take a stance against ("attack") or in support of ("defend") a regime. Departing from existing models, citizens' preferences depend on political salience scaling heterogeneous expressive values from action. Our interpretation of political salience is that it captures the general public's "bottom-up" attention to political values, 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, though in likely rather complex ways.[2] We analyze the effects of exogenous changes in political salience while remaining agnostic about their source. Our second key variable captures the regime's safeguards modelled as sanctions imposed on (failed) insurgents. Sanctions will act as a moderator of the effect of changes in political salience.

[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 political elites see Cinar and Nalepa [2022].

We solve the ensuing simultaneous-move game for Nash equilibria which we require to be stable. In cases with low political salience, "bandwagoning" concerns dominate as strategic citizens conform to avoid sanctioning. In cases with high political salience, citizens act purely expressively. The most interesting cases lie in-between, where there can be a rich array of equilibria and variation across citizens in whether bandwagoning or expressive incentives dominate.

We then zoom in on "regime-optimal equilibria" and study how changes in political salience alter both the equilibrium size of protest groups and the size of collective deviations required to transition to a more threatening equilibrium.

Our key findings relate to how changes in political salience and, hence, expressive concerns, affect regime resilience. The direction of the effect of salience on resilience depends on whether sanctions for siding with unsuccessful anti-regime movements are low or high relative to sanctions for siding with failing regimes. If low, then increases in salience from lower levels render the regime-optimal equilibrium ("none attack") less resilient by producing more accessible threat points. This is due to the fact that the equilibrium relies on bandwagoning by its opponents, which gets weakened with greater salience. At middling levels of salience, this can give rise to a unique equilibrium involving full opposition to the regime ("all attack"), whereas at high levels expressive concerns dominate, resulting in outright social conflict with an uncertain outcome. Conversely, when sanctions for siding with unsuccessful anti-regime movements are relatively large, then increases in salience from low to middling ranges gradually remove threat points, rendering regime support robust. Further increases in salience result in anti-regime actions among regime opponents but without gains from bandwagoning by others.

There is thus a very simple message that arises from our analysis. Regime threats depend on the interplay between political salience and 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 movement) suffice to activate otherwise latent opposition, which then gains further strength from bandwagoning by others. If safeguards are strong however, the same changes in salience can have opposite effects, further protecting regimes.

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.

2 Model and results

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.

The model has connections with the recent literature on global games (Carlsson and van Damme [1993], Shadmehr and Bernhardt [2011]), though these focus more specifically on information asymmetries, which we bracket here.

Our model also keeps a focus on citizen action rather than elite behavior. Elite behavior has been a central motivation to 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], resp.), 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.

There is a unit mass of citizens ("players"), each deciding whether to take an action to defend or attack a 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; regarding the distribution of citizens over such payoffs, we assume a cdf F that is strictly increasing and differentiable. Let $p(m)$ denote the probability that the incumbent regime is overthrown by the attack when a mass $m \in [0, 1]$ of players take actions against it, and assume $p(0) = 0$, $p'(m) > 0$ and $p(1) = 1$.[3] Let $\rho_A > 0$ (resp., $\rho_D > 0$) denote punishments imposed by winning attackers (resp., winning defenders) on citizens who have taken actions against them. Let "salience" $\sigma \in [0, 1]$ denote the importance of action payoffs relative to punishment concerns.

The expected sanctioning from joining an attack of size m is then $(1 - p(m))\rho_D$ while the expected sanctioning from defending is $p(m)\rho_A$. The expected utility *gain* from attacking the regime rather than defending it, given m, is then:

$$\sigma\epsilon_i + (1 - \sigma)(p(m)\rho_A - (1 - p(m))\rho_D), \tag{1}$$

and i will attack (resp., defend) the regime if this expected utility gain is positive (resp., negative). Given the probability of a successful attack is monotonically

[3] 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 of attackers as follows: Assign any non-measurable subset of agents attacking an arbitrary number in $[0, 1]$ and thus a success "probability;" however this is done, citizens' best-responding behavior, given any $m \in [0, 1]$, always results in a measurable partition of citizens, corresponding to a threshold t_m in payoff space $[-1, \alpha]$, such that the attackers are those with payoffs in (t_m, α), of which there is mass $1 - F(t_m)$. In this sense, the measurability issue pertains only to *defining* equilibrium without restricting to measurable strategy profiles, but not to equilibrium itself.

increasing in m, so is (1); i.e., punishment concerns give rise to bandwagoning incentives to conform with the majority action and thus reduce expected sanctioning (where majority is qualified by punishment levels). These are to be weighed against individuals' heterogeneous action preferences $\epsilon_i \in [-1, \alpha]$, whose relative importance is governed by σ.

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:" the share of players that weakly prefer to attack given that a share m of players choose to attack. A Nash equilibrium is then a fixed point of μ.[4] We call an equilibrium m^* "stable" if there exists some $\delta > 0$ such that $|\mu(m) - m^*| < |m - m^*|$ for all m with $0 < |m - m^*| < \delta$ (i.e., around m^*, $\mu(\cdot)$ is a local contraction). Otherwise we call it unstable.

Many concrete applications may fit this rather general reduced-form model. Our main application, following our introductory motivation, assumes that the incumbent regime is democratic and describes citizens as attacking or defending 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 focus on it. The scope of our general results on how political salience and polarization affect regime resilience extends beyond this case, however.

2.1 Application: Attacks on democracy

We provide a concrete/full microfoundation from a standard median voter setting for this application in supplementary material A. It yields an interpretation of ϵ_i as citizens' expressive payoffs reflecting their single-peaked preferences over policy outcomes, as they apply to the autocratic policy versus the democratic policy (the latter being the median ideal point). For the case of quadratic policy disutility, 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 $F(0) > 0.5$. (All visualizations below assume distributions F that indeed satisfy this property.)

To illustrate, 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

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

[4] We note a small abuse of terminology. A Nash equilibrium is a profile of strategies, not the number of people employing a particular strategy. Here, however, incentives have a threshold structure, so any equilibrium strategy profile, mapping values ϵ_i into the binary action, is fully characterized by the share of players attacking.

Hence, ϵ_i in (1) here corresponds to the difference in policy payoffs $(u_i(A)-u_i(D))$, as the basis for expressive concerns in driving political action.[5]

Substantively, we think of σ as reflecting the psychological importance placed on political action (relative to material costs associated with sanctioning). Our "salience" terminology derives from the psychology of bottom-up attention that is discussed and modeled in the survey of Bordalo et al. [2022]. Thus, we consider σ as being determined by the public attention that the political conflict receives. In view of our microfoundation, σ translates into 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). 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.[6]

2.2 Equilibrium, stability and salience

Our interest is in how equilibria, stable or unstable, depend on σ. We begin by characterizing the boundary cases.

Lemma 1. *Boundary cases:*

(i) If $\sigma = 0$, there exist three equilibria: share $m^* = 0$ ("none attack"), $m^* = 1$ ("all attack"), and $m^* = m_{\sigma=0} := p^{-1}\left(\frac{\rho_D}{\rho_A+\rho_D}\right) \in (0,1)$ attack. The two extreme equilibria are stable, the interior equilibrium is unstable.

(ii) If $\sigma = 1$, there exists a unique equilibrium: share $m^* = m_{\sigma=1} := 1 - F(0)$ attack. This equilibrium is stable.

Proof. (i) If $\sigma = 0$, this is a symmetric game of pure coordination with net utility from attack of $p(m)\rho_A - (1-p(m))\rho_D$. The claim follows from the fact that players are indifferent between attacking and not for $m = m_{\sigma=0}$, strictly prefer attacking for $m > m_{\sigma=0}$ and strictly prefer not attacking for $m < m_{\sigma=0}$. Our assumptions on p imply that $0 < m_{\sigma=0} < 1$. To establish stability of the extreme equilibria, take $\delta < \min\{m_{\sigma=0}, 1-m_{\sigma=0}\}$. To establish that the interior equilibrium at $m^* = m_{\sigma=0}$ is unstable, note that for any $m \neq m_{\sigma=0}$, either all or none will attack.

[5] This illustration also explicitly shows that ϵ_i is indeed an expressive, psychological payoff. The policy outcome is fully determined by the aggregate wherein any individual i's action is negligible, so any material policy payoff "cancels out."

[6] 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.

(ii) With $\sigma = 1$, utility from attacking equals ϵ_i, independent of how many others m attack. The same share of citizens i with $\epsilon_i \geq 0$ will attack, regardless of m, so $m^* = 1 - F(0)$ is the unique equilibrium, and it is stable. □

We will refer to the interior equilibria $m_{\sigma=0}$ and $m_{\sigma=1}$ from (i) and (ii) as the "pure coordination" and "pure expression" equilibria. Note that in both cases the democracy-optimal equilibrium is stable: it is the stable "none attack" equilibrium under pure coordination; and under pure expression it is the unique equilibrium, which is stable, of course.

Consider now cases with $\sigma \in (0, 1)$ in which players place weight on both the actions of others, through potential sanctions, and their own policy preferences. A citizen i is indifferent to taking part in attack against democracy if:

$$\epsilon_i = (-p(m)\rho_A + (1 - p(m))\rho_D) \cdot (1 - \sigma)/\sigma.$$

To avoid clutter we will define $\tilde{\sigma} := \sigma/(1 - \sigma) > 0$.

The attack response function is then:

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

Observe that, for a fixed/assumed attack size m, whether an increase in salience σ and hence in expressive concerns leads to an increase in how many citizens $\mu(m)$ will attack depends on whether bandwagoning incentives favor attack or defense (i.e., whether $(1 - p(m))\rho_D$ is smaller or greater than $p(m)\rho_A$). If these incentives favored attack (resp., defense), then they are in line with expressive concerns for all opponents (resp., supporters) to democracy and stronger such concerns have no effect on their optimal action; by contrast, stronger expressive concerns cause a change in the optimal action among at least some supporters (resp., opponents) of democracy, who had previously bandwagoned but now will act in line with their intensified intrinsic preference.

Note also that μ is differentiable, with the derivative μ' positive at any interior point $m \in (0, 1)$. Stability of an interior equilibrium m^* is then equivalent to $\mu'(m^*) < 1$, meaning that at m^* the attack response function μ crosses the 45-degree line from above.

For the analysis that follows we will rule out pathological (tangency) cases in which the slope of μ is exactly 1 at an equilibrium point—so that instability of an interior equilibrium is equivalent to $\mu'(m^*) > 1$—as well as the case that the pure coordination and the pure expression equilibria exactly coincide.

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

In addition, for equilibrium m^* given $\tilde{\sigma}$, we will abuse notation and write $m^*(\tilde{\sigma})$ to describe how equilibria vary in the neighborhood of m^* as a function of $\tilde{\sigma}$.

Our main results regarding stability and comparative statics in salience of various equilibria are summarized in Proposition 1, with illustrations of the results provided in Section 2.4's Figure 1 and in Appendix B.

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

(i) A stable equilibrium exists. In particular:

1. "None attack" is an equilibrium 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 $\frac{\partial m^*}{\partial \tilde{\sigma}} = 0$.

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

(ii) There is no equilibrium m^* with $\min\{m_{\sigma=0}, m_{\sigma=1}\} \leq m^* \leq \max\{m_{\sigma=0}, m_{\sigma=1}\}$.

(iii) An interior equilibrium $m^* < m_{\sigma=1}$ is stable if and only if $\frac{\partial m^*}{\partial \tilde{\sigma}}$ is positive; an interior equilibrium $m^* > m_{\sigma=1}$ is stable if and only if $\frac{\partial m^*}{\partial \tilde{\sigma}}$ is negative.

Proof. (i) As μ is a continuous mapping from the compact interval $[0,1]$ to itself, it satisfies the conditions of Brouwer's fixed-point theorem. The stronger result that a stable equilibrium exists follows from:

1. Equilibrium at $m = 0$ for $\tilde{\sigma} \leq \rho_D/\alpha$, and stability for $\tilde{\sigma} < \rho_D/\alpha$: Note that $\mu(0) = 1 - F(\rho_D/\tilde{\sigma})$, so $\mu(0) = 0$ if and only if $\rho_D/\tilde{\sigma} \geq \alpha$, 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 $\tilde{\sigma} < \rho_D/\alpha$, there is a $\delta > 0$ such that no one will attack also for any $m \in (0, \delta)$, by continuity of expected utility in m. For the same reason, marginal changes in salience then do not affect equilibrium, i.e., $\frac{\partial m^*}{\partial \tilde{\sigma}} = 0$.

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

3. Stable interior equilibrium if $\tilde{\sigma} > \max\{\rho_A, \rho_D/\alpha\}$: From the argument in 1., $\tilde{\sigma} > \rho_D/\alpha$ implies $\mu(0) > 0$ and, analogously, $\tilde{\sigma} > \rho_A$ implies $\mu(1) < 1$. Given this, by its continuity together with the genericity assumption, μ must *cross the 45-degree line from above* at some interior point $m \in (0,1)$, which is then an equilibrium; any such equilibrium m^* has $\mu'(m^*) < 1$, hence is stable.

It remains to establish existence of a stable equilibrium if $\tilde{\sigma} = \max\{\rho_A, \rho_D/\alpha\}$. Suppose that $\tilde{\sigma} = \rho_D/\alpha \geq \rho_A$, which implies $\mu(0) = 0$ and $\mu(1) \leq 1$. For the case that the "none attack" equilibrium is unstable, the genericity assumption implies that $\mu'(0) > 1$, whereby there exists $\hat{m} \in (0, \frac{1}{2})$ such that $\mu(\hat{m}) > \hat{m}$. If $\mu(1) < 1$, there exists a stable interior equilibrium by the argument given in 3.; if $\mu(1) = 1$ and the "all attack" equilibrium is unstable, then the genericity assumption implies that $\mu'(1) > 1$, whereby there exists $\tilde{m} \in (\frac{1}{2}, 1)$ such that

$\mu(\tilde{m}) < \tilde{m}$, so there exists a stable interior equilibrium $m^* \in (\hat{m}, \tilde{m})$, again by the argument given in 3. Existence of a stable equilibrium when $\tilde{\sigma} = \rho_A > \rho_D/\alpha$ follows analogously to when $\tilde{\sigma} = \rho_D/\alpha > \rho_A$.

(ii) Suppose first that $m^* \geq m_{\sigma=0}$ for some interior equilibrium m^*. Then $p(m^*) \geq p(m_{\sigma=0}) = \frac{\rho_D}{\rho_A+\rho_D}$. This implies that the indifferent citizen i in this equilibrium has policy preference

$$\epsilon_i = (\rho_D - p(m^*)(\rho_A + \rho_D))/\tilde{\sigma} \leq \left(\rho_D - \frac{\rho_D}{\rho_A + \rho_D}(\rho_A + \rho_D)\right)\frac{1}{\tilde{\sigma}} = 0$$

That is, the indifferent citizen i must be weakly leaning towards democracy in such an equilibrium. Hence, $m^* \geq 1 - F(0) = m_{\sigma=1}$. Analogously, $m^* \leq m_{\sigma=0}$ implies $m^* \leq m_{\sigma=1}$.

Finally, note that $\mu(m_{\sigma=0}) = 1 - F(0) = m_{\sigma=1}$, so neither of $m_{\sigma=0}$ or $m_{\sigma=1}$ is an equilibrium, by Genericity.

(iii) Define $\phi(m) := -p(m)\rho_A + (1 - p(m))\rho_D$; then, at an equilibrium point $m^* = \mu(m^*)$:

$$m^* - 1 + F(\phi(m^*)/\tilde{\sigma}) = 0$$

Consider then any interior equilibrium $m^* \in (0, 1)$. Let $\phi^* := \phi(m^*)$ and note that $m^* < m_{\sigma=1}$ (resp., $m^* > m_{\sigma=1}$) if and only if $\phi^* > 0$ (resp., $\phi^* < 0$). From the Implicit Function Theorem:

$$\frac{\partial m^*}{\partial \tilde{\sigma}} = \frac{f(\phi^*/\tilde{\sigma})\phi^*/\tilde{\sigma}^2}{1 - f(\phi^*/\tilde{\sigma})p'(m^*)(\rho_A + \rho_D)/\tilde{\sigma}}$$

The denominator is equal to $1 - \mu'(m^*)$, so it is positive (resp., negative) if and only if the equilibrium m^* is stable (resp., unstable). (Given our genericity assumption it cannot be zero.) Hence, $\frac{\partial m^*}{\partial \tilde{\sigma}}$ is of the same (resp., opposite) sign as ϕ^* if and only if m^* is stable (resp., unstable).

Finally, consider a stable "none attack" equilibrium. A marginal change in salience keeps $\tilde{\sigma} < \rho_D/\alpha$ intact, hence equilibrium unchanged. A similar argument applies to a stable "all attack" equilibrium. □

Jointly with Lemma 1, Proposition 1 establishes existence of a stable equilibrium and in particular a democracy-optimal stable equilibrium. For low salience, in the sense of $\tilde{\sigma} < \rho_D/\alpha$ this equilibrium has "none attack."

Most importantly, Proposition 1 essentially characterizes equilibrium comparative statics in salience via stability.[7] While the stable "all attack" and "none attack"

[7] We have omitted an explicit characterization for the knife-edge cases of "all attack" and "none attack" equilibria when $\tilde{\sigma} = \rho_A$ and $\tilde{\sigma} = \rho_D/\alpha$, respectively, in the proposition. However,

equilibria do not respond to marginal changes in salience, of course, a lesson of (i) is that increases in σ from intermediate levels can remove both of these (with the former disappearing first if $\rho_A \leq \rho_D/\alpha$).

For interior equilibria, stability and comparative statics in salience are tightly linked via bandwagoning. Specifically, a stable interior equilibrium $m^* < m_{\sigma=1}$ features bandwagoning by opponents to democracy, i.e., a positive mass of citizens with $\epsilon_i > 0$ for whom bandwagoning incentives dominate their opposing expressive concerns, so that they nonetheless defend democracy (recall that $m_{\sigma=1} = 1 - F(0)$ is the mass of citizens intrinsically supporting democracy). An increase in salience focuses them more on expressing their intrinsic values via their actions and thus activates (some of) this latent opposition. Analogously for the case of a stable interior equilibrium $m^* > m_{\sigma=1}$, which features bandwagoning by supporters of democracy. By contrast, unstable interior equilibria have the counter-intuitive property that increases in salience increase (rather than decrease) bandwagoning: To restore such equilibrium despite its instability, the increased interest in expressing political values must be countered by greater punishment risk; e.g., in an unstable interior equilibrium $m^* < m_{\sigma=1}$ with bandwagoning by opponents to democracy, increased salience leads to fewer and yet more fervent opponents who attack while facing accordingly greater punishment risk.

To understand the non-existence region (ii), observe that attack size $m_{\sigma=0}$ balances bandwagoning incentives exactly so that the indifferent/marginal citizen is also intrinsically action-indifferent, at $\epsilon_i = 0$. The number of citizens willing to attack is hence pinned down entirely by expressive motives: $\mu(m_{\sigma=0}) = 1 - F(0) = m_{\sigma=1}$. If $m_{\sigma=0} < m_{\sigma=1} = \mu(m_{\sigma=0})$, any $m > m_{\sigma=0}$ implies additional bandwagoning incentives to attack, so $\mu(m)$ rises even higher, above $m_{\sigma=1}$. Analogously, if $m_{\sigma=0} > m_{\sigma=1} = \mu(m_{\sigma=0})$, any $m < m_{\sigma=0}$ implies additional bandwagoning incentives to defend, so $\mu(m)$ falls even lower, under $m_{\sigma=1}$.

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 "almost" pure coordination, which then coexists with the stable "all attack" and "none attack" equilibria: Whether an increase in salience increases or decreases the attack size follows immediately from whether $m_{\sigma=0} > m_{\sigma=1}$ or $m_{\sigma=0} < m_{\sigma=1}$. Given there is no equilibrium in the range between the boundary cases of pure coordination and pure expression, there is only one direction for the unstable pure coordination equilibrium to move when salience increases, in line with the counter-intuitive effects on bandwagoning in unstable equilibria. If $m_{\sigma=0} > m_{\sigma=1}$, then such an interior equilibrium has $m^* > m_{\sigma=0}$, and its attack grows when salience increases, due to a further increase in bandwagoning by supporters of democracy. For the same reason, an interior equilibrium $m^* < m_{\sigma=0} < m_{\sigma=1}$ that features bandwagoning by opponents to democracy

these may be stable or unstable, and the characterization then is as in part (iii) for interior equilibria.

sees its attack shrink when salience increases.

Bandwagoning incentives themselves are directly affected by the sanctions inflicted upon those having joined the losing side. The following result concerns 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. We will again abuse notation and write $m^*(\rho_D)$ to describe these comparative statics.

Proposition 2. *Given Assumption 1, any stable interior equilibrium has* $\frac{\partial m^*}{\partial \rho_D} < 0$*, and any unstable interior equilibrium has* $\frac{\partial m^*}{\partial \rho_D} > 0$*.*

Proof. Similar to the proof of Proposition 1's part (iii), define $\phi(m) := -p(m)\rho_A + (1 - p(m))\rho_D$; then, at an equilibrium point $m^* = \mu(m^*)$:

$$m^* - 1 + F(\phi(m^*)/\tilde{\sigma}) = 0$$

Consider then any interior equilibrium $m^* \in (0, 1)$. Letting $\phi^* := \phi(m^*)$, from the Implicit Function Theorem:

$$\frac{\partial m^*}{\partial \rho_D} = \frac{-f(\phi^*/\tilde{\sigma})(1 - p(m^*))/\tilde{\sigma}}{1 - f(\phi^*/\tilde{\sigma})p'(m^*)(\rho_A + \rho_D)/\tilde{\sigma}}$$

Given an interior m^*, the numerator is negative. The denominator is equal to $1 - \mu'(m^*)$, so it is positive (resp., negative) if and only if the equilibrium m^* is stable (resp., unstable). (Given our genericity assumption it cannot be zero.) Hence, $\frac{\partial m^*}{\partial \rho_D} < 0$ if m^* is stable, and $\frac{\partial m^*}{\partial \rho_D} > 0$ if m^* is unstable. □

The intuition is straightforward: An increase in ρ_D directly renders attacking less attractive, for any assumed attack size (less than one). Starting from any interior equilibrium m^*, an increase in ρ_D thus implies that the share willing to attack assuming m^* falls below m^*. Stability (resp., instability) of m^* means $\mu'(m^*) < 1$ (resp., $\mu'(m^*) > 1$), which implies that maintaining the equilibrium yields a point where fewer attack (resp., where more attack).

The proposition focuses on interior equilibria, because these will be most relevant to our considerations of democratic resilience below. For completeness, however, note that ρ_D is irrelevant in any "all attack" equilibrium (the probability of being punished by the regime is then zero), and marginal changes to ρ_D could also generally not upset any "none attack" equilibrium when $\tilde{\sigma} < \rho_D/\alpha$. The remaining case is then the knife-edge cases of "none attack" equilibrium, where $\tilde{\sigma} = \rho_D/\alpha$: Clearly, any increase in ρ_D bolsters such equilibrium, whereas any decrease destroys it; see Proposition 1's part (i).

2.3 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 democracies' "resilience" of stable (democracy-optimal) equilibria, capturing the latent danger of shifting to a higher attack equilibrium in the event of a larger 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.

For any stable equilibrium m' that does not have "all attack" (i.e., $m' < 1$), consider the interior equilibrium m'' with the smallest attack size greater than m'. If such m'' exists, it is necessarily unstable: If $m' = 0$ ("none attack"), then stability implies $\mu'(0) < 1$, whereby an interior m'' would have to be one where μ crosses the 45-degree line from below (by genericity). Moreover, if such m'' exists, then there also exists a stable equilibrium $m''' > m''$ adjacent to it (i.e., there are no equilibria $m \in (m'', m''')$), by a similar argument. We then refer to the unstable interior equilibrium m'' as the *threat point* of stable equilibrium m', and we take the distance $m'' - m' > 0$ to measure the *resilience* of m': Any shock such that $m < m''$ attack would not seriously upset the stable equilibrium m' in the longer run, whereas any shock such that $m > m''$ would lead society away from stable m' with a much increased attack size of (at least) *stable* m''' in the longer run. We note that this need not be an actual dynamic adjustment process but could also be a collectively shared reasoning process in face of 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 stable "none attack" equilibrium, Proposition 1's (ii) and (iii) imply the following result, concerning the effect of salience on democracy's resilience:

Corollary 1. *Given any $\tilde{\sigma} \geq 0$ and existence of an interior equilibrium, a marginal increase in salience renders a stable "none attack" equilibrium with threat point m^* less (resp., more) resilient if m^* is smaller (resp., greater) than $m_{\sigma=1}$.*

Proof. Given the arguments preceding the corollary, threat point m^* is an unstable interior equilibrium, and by Proposition 1's (iii), a marginal increase in salience does not affect "none attack," whereas $\frac{\partial m^*}{\partial \tilde{\sigma}}$ is negative if $m^* < m_{\sigma=1}$, and $\frac{\partial m^*}{\partial \tilde{\sigma}}$ is positive if $m^* > m_{\sigma=1}$. □

A special case of this result applies when $\tilde{\sigma} = \sigma = 0$, for which Lemma 1's (i) characterizes equilibrium. The stable "none attack" equilibrium has threat point $m^* = m_{\sigma=0}$. In view of Proposition 1's (i), a marginal increase in $\tilde{\sigma}$ does not change the "none attack" equilibrium directly, since $\rho_D/\alpha > 0 = \tilde{\sigma}$. Yet, as discussed as a lesson of the proposition's (ii) and (iii), it moves the threat point closer if $m_{\sigma=0} < m_{\sigma=1}$ and further away if $m_{\sigma=0} > m_{\sigma=1}$.

More generally, observe that when ρ_D is relatively low so that $m_{\sigma=0} < m_{\sigma=1}$, there exists an unstable equilibrium $m^* \leq m_{\sigma=0}$, as a threat point of a stable "none attack" equilibrium. Corollary 1 then tells us that an increase in salience always reduces its resilience. Furthermore, if there are more equilibria with attack size less than $m_{\sigma=0}$, then the increase in salience also increases the attack size in the stable equilibrium that obtains after "none attack" gets upset by a sufficiently large shock; otherwise this long-run resting point has an attack size even greater than $m_{\sigma=1}$, because of Proposition 1's (ii).[8] By contrast, when ρ_D is relatively high, so that $m_{\sigma=0} > m_{\sigma=1}$ and there exists no unstable equilibrium $m^* \leq m_{\sigma=0}$ (which would in fact require $m^* \leq m_{\sigma=1}$), then an increase in salience always increases the resilience of a stable "none attack" equilibrium, by raising the threat point (even further).

Our simple model thus points out an essential risk to democracy from increased political salience when its sanctions against insurgents are weak.[9] While no attack whatsoever becomes apparent, yet ever smaller shocks would destroy it and may even move society to an "all attack" equilibrium, with sure autocracy (this when there is but one interior and hence unstable equilibrium). Proposition 2 then implies that strengthening democratic sanctions is effective in increasing democracy's resilience, removing threat points (which are unstable interior equilibria, by definition).

2.4 Illustration

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

[8] While the latter case is easily illustrated (see Figure 1 below, in the upper right panels), the former case requires richer equilibrium multiplicity as illustrated only in supplementary material B. Specifically, see its Figure 4 and consider increases in salience within the range $\sigma \in (0.15, 0.2)$.

[9] The weakness of democratic sanctions concerns cases where the pure coordination equilibrium $m_{\sigma=0} = p^{-1}(\rho_D/(\rho_A + \rho_D))$ has fewer attacking than are intrinsically motivated to do so, i.e., than the pure expression equilibrium $m_{\sigma=1} = 1 - F(0)$. It is thus about democracy's punishments ρ_D being small relative to the combination of both the punishments ρ_A imposed by the autocratic attack and the latter's ideological support among citizens $1 - F(0)$.

We imagine $p(m) = m$ and $\epsilon_i \sim U[-1, 0.5]$, so that $F(x) = \frac{2}{3}(x+1)$ for $x \in [-1, 0.5]$. We have $m_{\sigma=0} = \frac{\rho_D}{\rho_A+\rho_D}$ and $m_{\sigma=1} = \frac{1}{3}$. Thus, $m_{\sigma=0}$ is larger (resp., smaller) than $m_{\sigma=1}$ if and only if $\rho_A < 2\rho_D$ (resp., $\rho_A > 2\rho_D$). For $\sigma \in (0, 1)$, the attack response function is linear in m:

$$\mu(m) = \frac{1}{3} - \frac{2}{3}\frac{1}{\tilde{\sigma}}(\rho_D - (\rho_A + \rho_D)m),$$

so there is at most one interior equilibrium, which then corresponds to fixed point:

$$m^* = \frac{\tilde{\sigma} - 2\rho_D}{3\tilde{\sigma} - 2(\rho_A + \rho_D)}.$$

Note that, written as a function, $m^*(\tilde{\sigma})$ approaches $m_{\sigma=0}$ as σ approaches 0 (and so $\tilde{\sigma}$ approaches 0) and $m^*(\tilde{\sigma})$ approaches $m_{\sigma=1}$ as σ approaches 1 (and so $\tilde{\sigma}$ approaches infinity). Moreover, $m^*(\tilde{\sigma})$ is increasing in σ if and only if $\rho_A < 2\rho_D$, or, equivalently, $m_{\sigma=0} > m_{\sigma=1}$; analogously, decreasingness in σ is equivalent to $m_{\sigma=0} < m_{\sigma=1}$. From Proposition 1's (iii), interior equilibrium m^* is therefore stable if and only if either democracy's sanctioning is relatively high ($\rho_A < 2\rho_D$) and there is *pro*-democracy bandwagoning ($m^* < \frac{1}{3}$) or autocracy's sanctioning is relatively high ($\rho_A > 2\rho_D$) and there is *anti*-democracy bandwagoning ($m^* > \frac{1}{3}$).

Equilibria are illustrated in Figure 1.[10] The figure confirms:

1. Low salience always yields three equilibria, the two stable "none attack" and "all attack" equilibria as well as the unstable (more or less pure) coordination equilibrium; high salience eliminates these extreme equilibria and results ultimately—when $\sigma = 1$—in a unique (stable) pure expression equilibrium with outright conflict and an uncertain outcome.

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

    - when $m_{\sigma=0} < m_{\sigma=1}$, an increase in salience from a low to a middling range renders "none attack" less and less resilient by pulling threat points near and may—even in the absence of any shock—not only eliminate this equilibrium but yield a unique equilibrium where instead "all attack;"

[10] A noteworthy property from this figure is that, in all cases, there is a unique branch continuously connecting the unique (and stable) pure expression equilibrium $m_{\sigma=1}$ 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 $m_{\sigma=1} < m_{\sigma=0}$ and "all attack" if $m_{\sigma=1} > m_{\sigma=0}$. 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 supplementary material B.1, 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 3).

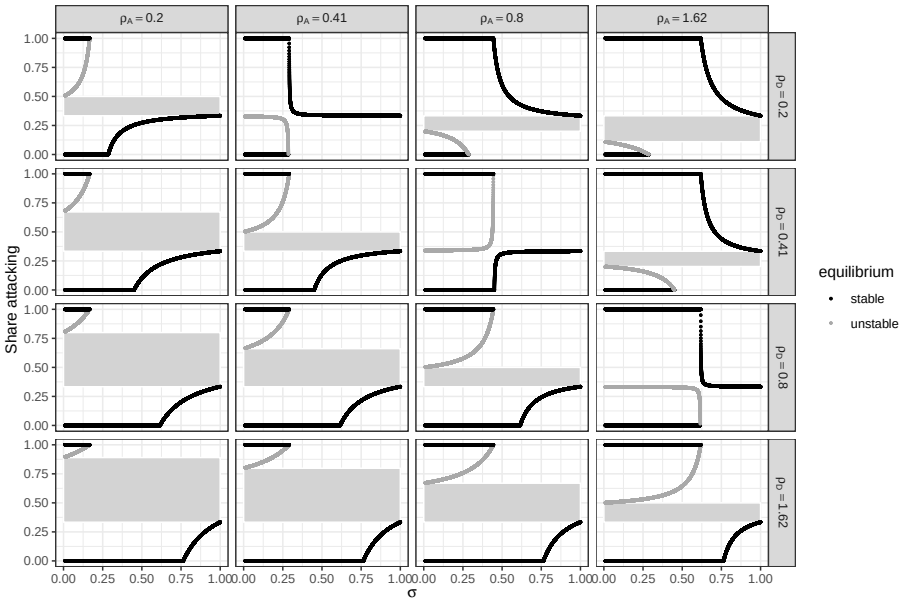

Figure 1: Equilibria in a linear-uniform model – black = stable, dark grey = unstable, no equilibria in light-grey-shaded rectangular areas. Each panel varies salience $\sigma \in [0, 1]$ for given sanction values (ρ_A, ρ_D), and these vary between panels (ρ_A increases rightwards, ρ_D increases downwards). In upper right (lower left) panels, where $\rho_A < 2\rho_D$ ($\rho_A > 2\rho_D$), the pure coordination equilibrium is lower (higher) than the expressive equilibrium and the interior equilibrium is decreasing (increasing) in σ.

- when $m_{\sigma=0} > m_{\sigma=1}$, an increase in salience from low to middling ranges renders "none attack" more and more resilient by pushing threat points away and may—even in the absence of any shock—turn it into the unique equilibrium.

While Figure 1 highlights effects of changing salience given sanctioning, Figure 2 illustrates the possible effects of changing 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), as a function of $\rho_D \in [0, 1]$, for different values of political salience σ.

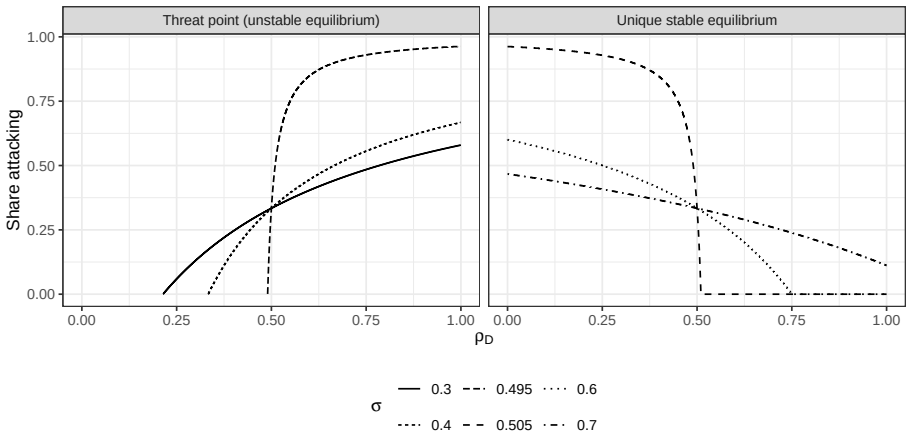

Figure 2: Illustration of (unique) interior equilibria as a function of $\rho_D \in [0, 1]$ when $\rho_A = 1$, for six different values of σ. The left panel has three values $\sigma < 0.5$, in which case this interior equilibrium is unstable, acts as a threat point to the coexisting stable "none attack" 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.

The left panel illustrates three cases of low to middling salience, $\sigma < 0.5$: Then, whenever an interior equilibrium exists, it coexists with a stable "none attack" equilibrium and is the latter's threat point (unstable), which gets removed as ρ_D increases. 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, $\sigma > 0.5$: Then, such equilibrium is stable and the overall unique equilibrium, whose 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 support, respectively. For instance, consider the two corresponding cases of middling salience in Figure 2, $\sigma = 0.495$ in the left panel and $\sigma = 0.505$ in the right panel. In other ranges, 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 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 democratic sanctions are relatively weak, increases to middling levels of political salience can render democracies especially vulnerable. Increases in the 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 democratic equilibrium relies on continued bandwagoning by latent opponents. Since bandwagoning incentives are stronger for the anti-regime than the pro-regime equilibrium, when increased salience renders bandwagoning incentives relatively less important, 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 polarization put democracy under especially great risk of dramatically tipping in response to only small shocks.

In situations in which democratic sanctions are strong, increases in salience from low to middling levels have the opposite effect of rendering democracies more resilient. However, increases in salience at high levels also make it more difficult to keep opposition at bay. The intuition is that in a democracy-optimal equilibrium the indifferent agent is indifferent only because of the threatened sanction. On the basis of pure policy preferences she would support the insurrection. An increase in political salience thus shifts the agent to act against the regime.

By the same token, sanctions can have dramatic strategic effects on regime resilience as well as support depending on the level of salience. 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 safeguards for democracy and increase sanctions for its opponents.

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.

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

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.

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.

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.

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.

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

Appendix: Supplementary material

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

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

for some preference parameters $\overline{u} > 0$ (the political "bliss" value when $\widehat{v} = v_i$) and $\tau > 0$ (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 strictly 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,

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

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 $2\tau \cdot (v^A - v^D) \cdot (\underline{v} - w) < 0$ to a maximum of $2\tau \cdot (v^A - v^D) \cdot (\overline{v} - w) > 0$. Mapping any ideal point v_i into ϵ_i as

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

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 (2) 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 Additional examples and visualization

B.1 Multimode distribution

We imagine a distribution of preferences with many modes, in which p in linear, and in which $m_{\sigma=1} = 1 - F(0) < m_{\sigma=0} = \frac{\rho_D}{\rho_A + \rho_D} = \frac{1}{3}$.

Figure 3 then shows how equilibria change as σ changes, plotting μ against m. Fixed points are equilibria. The boundary cases of Lemma 1 can be seen at the extremes, with a multiplicity of equilibria possible for intermediate values of σ.

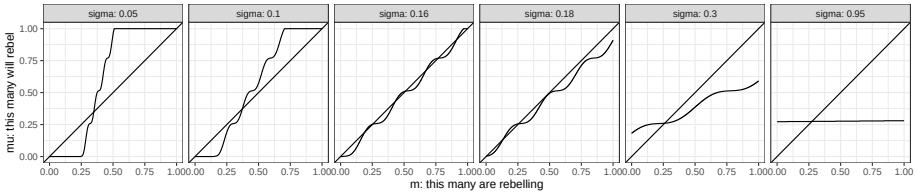

Figure 3: When σ is small there are generically three equilibria generated by a symmetric coordination game. One of these is the interior, unstable, "**pure coordination**" equilibrium. When σ is large there is a unique "**pure expression**" equilibrium, which is stable. At intermediate levels there can be many equilibria.

The full set of equilibria over the range of σ is shown for this example in Figure 4. We see again the equilibria identified by Lemma 1 at the boundaries. A grey block marks the region where there are no equilibria (Proposition 1 (ii)). Note also that below this region (where fewer are rebelling than would want to, absent sanctions), stable equilibria are increasing in σ and unstable equilibria are decreasing, indicating a generally higher risk of rebellion, locally. Above this region (where more are rebelling than would want to, absent sanctions), stable equilibria are decreasing in σ and unstable equilibria are increasing,

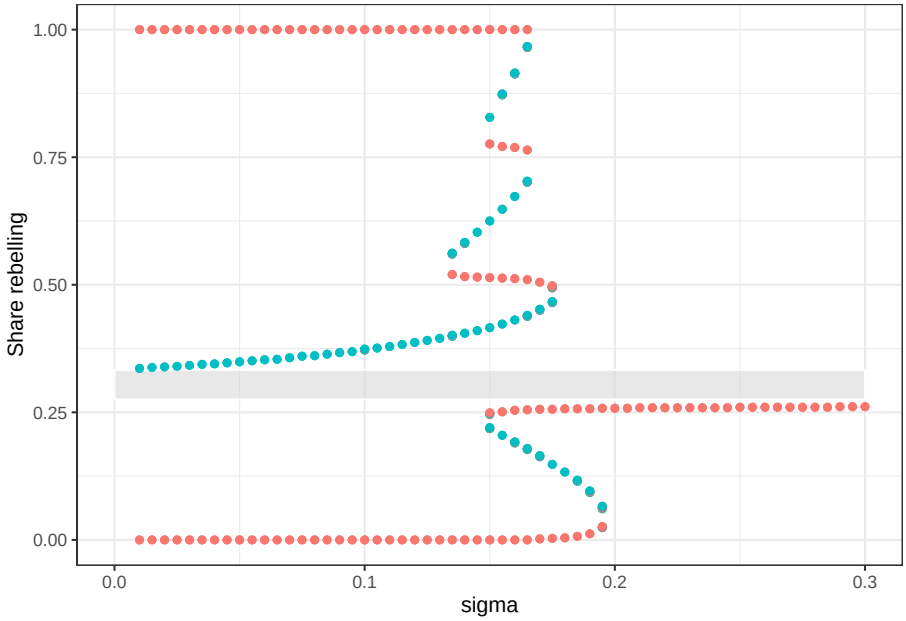

Figure 4: Summary of equilibria as a function of σ for the same parameters as in Figure 3.

B.2 Beta distribution

In Figure 5 we illustrate equilibria for a setting with Beta-distributed preferences, assuming $\alpha = 1$, and success probabilities that are convex in participation ($p(m) = m^{3/2}$).

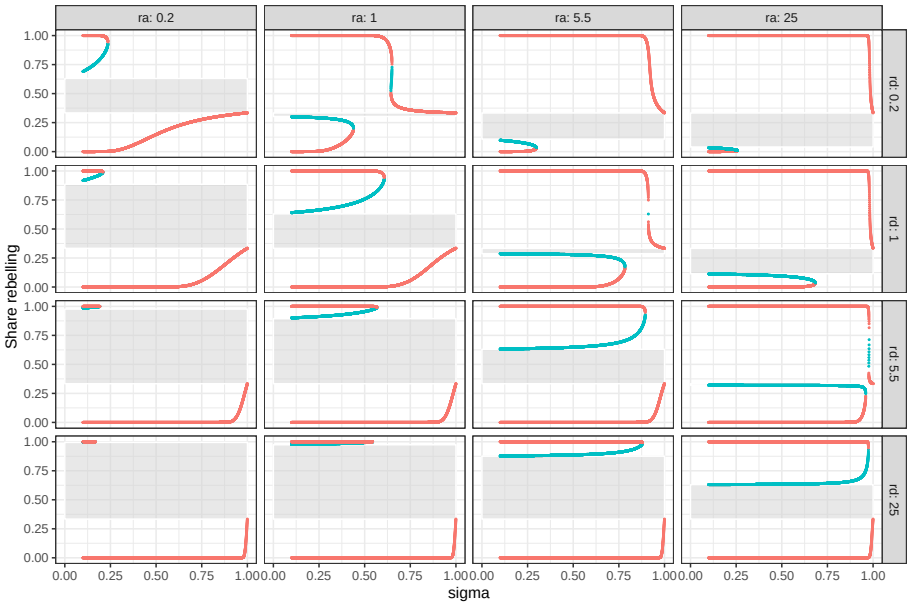

Figure 5: Equilibria in a quadratic/Beta model (pink = stable). No equilibria in grey areas. In upper right panels, the pure coordination equilibrium is lower than the expressive equlibrium. In the lower left panels, the pure coordination equilibrium is higher than the expressive equlibrium.