Political salience and regime resilience

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

07/12/2022

Abstract

We study a version of a canonical model of attacks against political regimes where agents have an expressive utility for taking political stances that is scaled by the *salience* of political decision-making. Increases in political salience can have divergent effects on regime stability depending on costs of being on the losing side. When regimes have weak sanctioning mechanisms, middling levels of salience can pose the greatest threat, as regime supporters are insufficiently motivated to act on their preferences and regime opponents are sufficiently motivated to stop conforming. Our results speak to the phenomenon of 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.

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* to assess how salience, in and of itself, can contribute to the vulnerability of political regimes.

In some accounts, disinterest in politics threatens democracy, the "slow slump in interest in politics and current events" according to Putnam et al. [2000], can be one source of vulnerability. Other work highlights the rising stakes of political decision-making. 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]

*Our thanks to Daniel Markovits, Ethan Bueno de Mesquita, and Haoyu Zhai for generous advice on this project.

[1]We leave aside the empirical question of whether political salience is rising or falling. In some accounts it is falling, as in Putnam et al. [2000]. In many journalistic accounts it is rising:

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, citizen action depends on political salience which in turn produces an intrinsic preference for action. Our first key variable is the strength of these preferences. Our second captures the regime's safeguards—the sanctions imposed on (failed) insurgents.

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 focus on how the regime-optimal equilibrium changes when politics becomes more salient, and in particular on how changing salience alters both the equilibrium size of protest groups and the size of collective deviations required to transition to a more threatening equilibrium.

The key findings relate to how changes in political salience and, hence, expressive concerns, affect regime resilience. When sanctions for siding with unsuccessful anti-regime movements are low relative to sanctions for siding with failing regimes, 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 middle 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 and then eliminate 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, latent opposition becomes active and gains strength from bandwagoning by others.

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.

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 Model

We examine a model in the spirit of the classic accounts of Granovetter [1978] and Kuran [1989] in which a collection of players trade 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 manipulation by regimes (Edmond [2013]) or 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. For exposition we will assume that the incumbent regime is democratic and describe citizens as attacking or defending democracy vis-à-vis an autocratic agitator.

Let $\epsilon_i \sim F$ on $[-1, \alpha]$ denote a player-specific payoff for attacking (relative to defending) democracy, with $0 < \alpha < \infty$; we assume that F is strictly increasing and differentiable. We interpret ϵ_i as reflecting preferences over policy outcomes. For our illustrations we generally assume that $F(0) > 0.5$, indicating that a majority favors democracy-defending behavior.We describe in the Supplementary Material how our model can be microfounded by a model in which democracy would select the median voter's ideal policy as outcome and confronts an attack by an alternative policy regime. However, no results depend on this interpretation, and ϵ_i could as well reflect general identification with different political systems.

Let $\rho_A > 0$ (resp., $\rho_D > 0$) denote punishments imposed by winning autocrats (resp., democrats) on citizens who have taken actions against them. Let $p(m)$ denote the probability that democracy is overthrown when m actors take actions against it; we assume $p(0) = 0$, $p'(m) > 0$ and $p(1) = 1$. Let "salience" $\sigma \in [0, 1]$ denote the relative importance of expressive policy concerns to punishment concerns. Substantively we think of σ as reflecting the importance placed on political action (relative to costs associated with sanctioning). Thus, similar to ideas in Riker and Ordeshook [1968], σ can capture a weight placed on civic duty, though here with heterogeneity regarding whether there is a duty to attack

or defend a regime. We note that σ can also be interpreted as reflecting political polarization—how different policy outcomes would be when different groups control government—and thus the stakes of political control (see also Chiopris et al. [2021] on platform divergence and attitudes to backsliding).[2]

The expected utility gain from supporting autocracy (rather than democracy) is then:

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

and i will support autocracy (resp., defend democracy) if this expected utility is positive (resp., negative). 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.[3] We call an equilibrium m^* "stable" if there exists some $\delta > 0$ such that $|\mu(m) - m^*| < |m - m^*|$ for all m with $0 < |m - m^*| < \delta$. Otherwise we call it unstable.

2.1 Results

Our interest is in how equilibria depend on σ. We begin by characterizing the boundary cases:

Proposition 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.

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.

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

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

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

[3] 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.

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

The attack response function is then:

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

Note that μ is differentiable, with μ' positive at any interior equilibrium $m^* \in (0,1)$.Stability of equilibrium is then equivalent to $\mu'(m^*) < 1$.

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, as well as the case that the pure coordination and the pure expression equilibria exactly coincide.

Assumption 1 [Genericity]: If $\mu(m) = m$ then $\mu'(m) \neq 1$, 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 are then given in the next proposition.

Proposition 2. *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 $\tilde{\sigma} < \rho_D/\alpha$.

2. "All attack" is an equilibrium if and only if $\tilde{\sigma} \leq \rho_A$. It is stable if $\tilde{\sigma} < \rho_A$.

3. If $\tilde{\sigma} > \max\{\rho_A, \rho_D/\alpha\}$, then there exists a stable interior equilibrium.

(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 $m^* < m_{\sigma=1}$ (resp., $m^* > m_{\sigma=1}$) is stable if and only if $\frac{\partial m^*}{\partial \tilde{\sigma}}$ is positive (resp., negative), and it is unstable if and only if $\frac{\partial m^*}{\partial \tilde{\sigma}}$ is negative (resp., positive).

A lesson of (i) is that increases in σ can remove both "all attack" and "none attack" equilibria (with the former disappearing first if $\rho_A \leq \rho_D/\alpha$). A lesson of (ii) and (iii) is that the direction of equilibrium effects of an increase in salience depends on the relative attack sizes in the pure coordination and pure expression equilibria.

2.2 Dynamic considerations

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, 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 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 changing 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.

Applying this notion to a stable "none attack" equilibrium, Proposition 2'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 $m_{\sigma=1}$.*

A special case of this result applies when $\tilde{\sigma} = \sigma = 0$, for which Proposition 1's (i) characterizes equilibrium. The stable "none attack" equilibrium has threat point $m^* = m_{\sigma=0}$. In view of Proposition 2's (ii) and (iii), a marginal increase in $\tilde{\sigma}$ does not change the "none attack" equilibrium directly, since $\rho_D/\alpha > 0 = \tilde{\sigma}$, yet moves the threat point closer if $m_{\sigma=0} < m_{\sigma=1}$ and further away if $m_{\sigma=0} > m_{\sigma=1}$ (given (ii) and continuity, there is only one direction to move).

More generally, the Corollary implies, for instance, that when salience is low and ρ_D is relatively low, an increase in salience reduces resiliency by lowering the threat point for a "none attack" equilibrium. When ρ_D is relatively high, however, it increases resiliency by raising the threat point.

2.3 Illustration

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

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 $\tilde{\sigma}$ 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 2's (iii), interior equilibrium m^* is therefore stable if and only if either regime sanctioning is relatively high ($\rho_A < 2\rho_D$) and there is *pro*-regime bandwagoning ($m^* < \frac{1}{3}$) or opposition sanctioning is relatively high ($\rho_A > 2\rho_D$) and there is *anti*-regime bandwagoning ($m^* > \frac{1}{3}$).

Equilibria are illustrated in Figure 1. The figure confirms:

1. Low salience always yields three equilibria, the two stable "none attack" and "all attack" equilibria as well as the unstable coordination equilibrium; increases in salience eliminate these equilibria and result ultimately—when $\sigma = 1$—in a unique (stable) pure expression equilibrium with partial attacks
2. Greater salience can increase or reduce risks of attack. In particular:
    - when $m_{\sigma=0} < m_{\sigma=1}$, an increase in salience from low to middling range can yield a unique equilibrium with all attacking
    - when $m_{\sigma=0} > m_{\sigma=1}$, an increase in salience from low to middling ranges can eliminate threat points and produce a unique equilibrium with none attacking

While Figure 1 highlights effects of changing salience given sanctioning, Figure 2 illustrates the possibly dramatic effects of changing sanctions given salience, showing how equilibrium depends on ρ_D when $\rho_A = 1.2$ and $\sigma = 0.55$. (For these values there is always a unique equilibrium.) Critically, with this intermediate

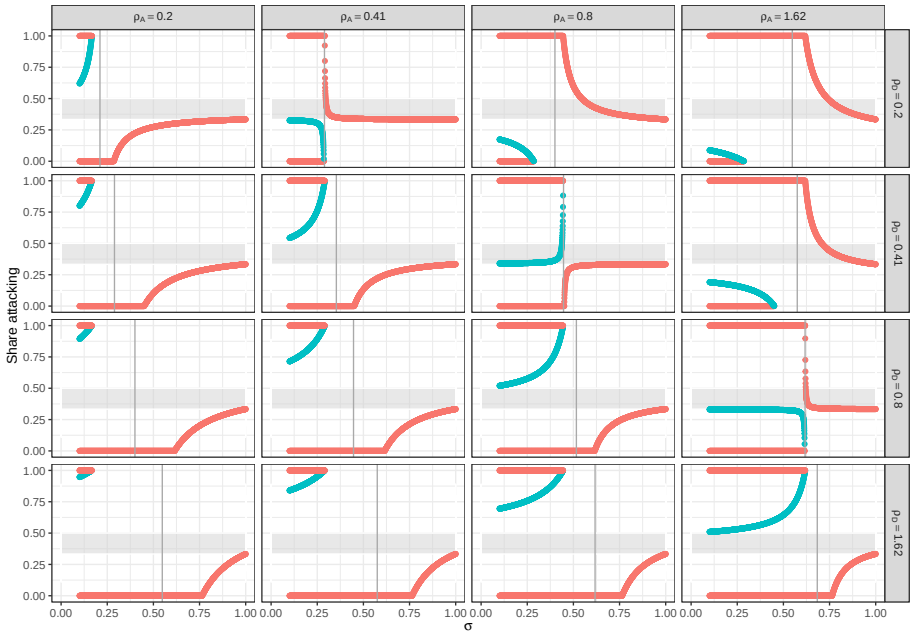

Figure 1: Equilibria in a linear-uniform model (pink = stable). No equilibria in grey areas. In upper right (lower left) panels, the pure coordination equilibrium is lower (higher) than the expressive equlibrium and the interior equilibrium is decreasing (increasing) in σ.

value of salience a small change in democratic sanctions can have dramatic strategic effects on bandwagoning and the level of system support.

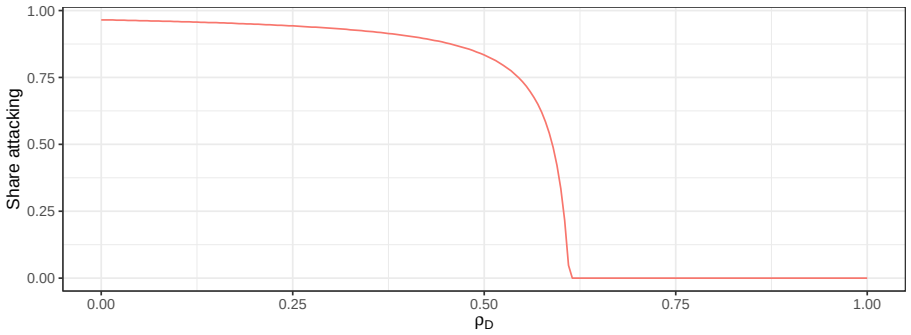

Figure 2: Illustration of (unique) equilibria as a function of ρ_D when $\rho_A = 1.2$ and $\sigma = 0.55$.

3 Conclusion

We study a 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 a 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 problem of democracy, our results suggest that when democratic sanctions are relatively weak, increases to middling levels of political salience can render democracies especially vulnerable. The intuition is that they rely on bandwagoning by latent opponents, and bandwagoning incentives are stronger for the anti-regime than the pro-regime equilibrium.

In situations in which democratic sanctions are strong, 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 support depending on the level of salience. This finding has bearing on contemporaneous threats to the 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.

4 Appendices (proofs)

4.1 Proof of Proposition 1

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.

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

4.2 Proof of Proposition 2

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.

2. Equilibrium at $m = 1$ for $\tilde{\sigma} \leq \rho_A$, and stability for $\tilde{\sigma} < \rho_A$: 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.

□

4.3 Proof of Corollary 1

Proof. Given the arguments preceding the corollary, threat point m^* is an unstable interior equilibrium, and by Proposition 2'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 $m^* > m_{\sigma=1}$. □

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.

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. Technical report, Working Paper, 2021.

Chris Edmond. Information manipulation, coordination, and regime change. *The 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.

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.

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. *The 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 et al. *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. *The American Political Science Review*, 105(4):829–851, 2011.

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