# Collective Procrastination and Protest Cycles*

Germán Gieczewski† and Korhan Kocak‡

November 2023.

**Abstract**

This paper studies a model of collective action in which citizens face repeated opportunities to protest against a regime, and imperfectly observe the gains from protesting at each moment. We depart from the existing literature by assuming that citizens are partially altruistic, and hence may protest to change the outcome (i.e., be pivotal), even when protesting yields no private benefits. We show that "pivotal protesting" entails complex dynamic considerations. Indeed, the continuation value of the status quo influences the citizens' willingness to protest today. Thus, a mere change in expectations about the future may trigger a revolt. The same logic can induce a pattern of protest cycles, as well as introduce a novel source of inefficiency: a temptation to protest later rather than earlier when future protesting opportunities are mildly attractive leads to protests being inefficiently delayed. Thus, altruistic agents can fall prey to a form of collective procrastination.

## 1 Introduction

It should not be controversial that people, in particular in their capacity as protesters, think about the future, and that participation in mass protests is often spurred by

*A previous version was circulated under the title "Altruism in Protests." We are thankful to Carlo Horz, Federica Izzo, Dimitri Landa, Andrew Little, Mehdi Shadmehr, seminar participants at Princeton, Sabanci, and UCSD, and audience members at APSA, CPFT, and MPSA for helpful comments and suggestions. Kocak gratefully acknowledges financial support from the Center of Behavioral Institutional Design and Tamkeen under the NYU Abu Dhabi Research Institute Award CG005.

†Department of Politics, Princeton University.

‡Division of Social Science, NYU Abu Dhabi.

the perception that a successful protest may improve future welfare; conversely, that the state of society will deteriorate if nothing is done to change the status quo; and, more generally, that the country stands at a crossroads.

For example, the massive 2019–2020 protests in Hong Kong were triggered by the introduction of a *proposed* bill that would have allowed extraditions to mainland China. As crowds swelled, government rhetoric hardened, and police crackdowns intensified, protesters described being motivated by a sense that Hong Kong faced a do-or-die fight for its future, saying to the media: "If we don't succeed now, our freedom of speech, our human rights, all will be gone."[1] A similar display of forward thinking is encapsulated in the chilling slogan used by Taiwanese protesters: "Today's Hong Kong, tomorrow's Taiwan."[2] Similarly, the 2013-2014 protests in Ukraine that ultimately ousted Yanukovych were sparked by his postponing a promised integration agreement with the European Union, a move that signaled the country would instead seek closer ties with Russia.[3] Many other examples exist.[4]

Yet forward-looking considerations are conspicuously absent from most formal models of mass protest, even dynamic ones. The reason traces back to the fact that, as the literature on protests—formal and otherwise—is well aware, protests are a collective action problem (Olson, 1965; Lichbach, 1995): even large public benefits cannot induce selfish citizens to undertake private costs. The most popular formal approaches to circumventing this problem, so as to explain participation, postulate that citizens are motivated either by private benefits or by expressive warm-glow payoffs. As we discuss later, both assumptions yield models in which citizens behave *as if* myopic in equilibrium, *even* when assumed to be forward-looking.

In this paper, we study a model of repeated protests, in which citizens can attack the regime (protest, mobilize) in each of many periods, and receive information about the potential gain from doing so in each period. Our model uses the machinery of global games (Carlsson and Van Damme, 1993; Morris and Shin, 1998), and is in

[1] `https://www.reuters.com/article/us-hongkong-protests-radicals/now-or-never-hong-kong-protesters-say-they-have-nothing-to-lose-idUSKCN1VH2JT`

[2] `https://foreignpolicy.com/2014/08/19/todays-hong-kong-tomorrows-taiwan/`

[3] `https://www.nytimes.com/2013/11/22/world/europe/ukraine-refuses-to-free-ex-leader-raising-concerns-over-eu-talks.html`

[4] Consider the 2013 protests in Turkey, whose immediate trigger was the violent eviction of a sit-in at Gezi park, but which responded more broadly to "creeping political authoritarianism" (Özel, 2014), or the 2023 protests in Israel against a plan to weaken the judiciary's influence over policymaking.

many ways canonical, but it presumes a novel motivation for citizens to participate relative to the formal literature on repeated protests (Angeletos, Hellwig and Pavan, 2007; Little, 2017), one which—crucially—preserves a role for forward thinking in equilibrium. Namely, we assume that citizens are partially altruistic: they put some weight on the welfare of their fellow citizens. Under this assumption, the importance of public benefits in a citizen's benefit-cost calculation does not vanish as the population becomes large. Even as any citizen's chance of being pivotal goes to zero, the impact she can have on total welfare by tipping the outcome grows without bound. It follows that partially altruistic citizens retain *agency* in a large population and thus react to expectations about the future: since the public benefit from a successful protest is simply the gap between continuation values under regime change and the status quo, incentives to protest respond both to the "carrot" of a better post-revolutionary outcome and the "stick" of a status quo that is or will become oppressive. In particular, a mere change in expectations can trigger a protest.

But this logic carries further: the citizens realize that, by successfully overthrowing the regime today, they are forfeiting chances to instead do so in the future. A successful protest thus robs all citizens of the potential gain from future protests. Their calculus must account for this. As a result, they are more likely to attack today if conditions for an attack are good today *or if they are bad tomorrow*.

Our most surprising result, however, is that when agents are *imperfectly* altruistic—that is, they value others' welfare, but less than their own—the "opportunity cost" considerations induced by altruism can lead to excessive and inefficient delay in equilibrium, a form of *collective procrastination*. More precisely, giving citizens more opportunities to protest in the future, *even less attractive ones* than the current one, can lower their equilibrium welfare by inducing a sort of "paralysis of options." The logic of this result is related to the intuition behind equilibrium selection in all global games: in global games, agents can sometimes coordinate on an attack, but they need the state of the world to be somewhat better than the bare minimum needed to render an all-out attack profitable. A crowd in a global game thus behaves much like a person with low motivation or willpower to exert effort. Offering such a person an "out" in the form of a second chance can tempt her to procrastinate, leaving her worse off. Thus, an attack may come not when it is most profitable, but rather when there are no second chances left.

Due to the same logic, the equilibrium generally displays a pattern of *protest*

*cycles*: like waves crashing against the shoreline, citizens eventually coordinate on an attack, then—if unsuccessful—let several periods pass before trying again, and the process repeats. These waves are strategic and forward-looking: citizens attack when the *anticipated* delay until the next wave crashes is long enough that they become impatient.

It is important to note that our results do not hinge on altruism *per se*, but on the fact that, because of it, the citizens have *agency*: they do not see the aggregate consequences of their participation as negligible. As we show in an example, similar results arise *without* altruism in a finite population version of the game, where each agent's participation has a real chance of changing the outcome. Technically, adding altruism to the model allows us to recover the logic of pivotality even when the population is large. Substantively, it provides a way to account for other-regarding preferences, community, civic duty, and other moral concerns that have long been recognized by classical accounts of mass social movements (Lichbach, 1995; Wood, 2003) as motivating potential protesters, but which have received little attention in the formal literature. Indeed, to our knowledge, we are the first to study a dynamic formal model of protests with moral agents.[5]

# 2 Related Literature

The literature on social movements broadly considers two types of potential motivations for citizens to engage in costly collective action: private benefits (Olson, 1965; Tullock, 1971), that is, material or social benefits of regime change, at least some of which are exclusive to participants; and psychological rewards, such as frustration in response to relative deprivation (Gurr, 1970) and "pleasure in agency" (Wood, 2003).[6]

Many formal models of protest assume private benefits as the citizens' motivation (Casper and Tyson, 2014; Tyson and Smith, 2018; Bueno De Mesquita and Shadmehr, 2023). Under this assumption, each citizen's incentive to participate depends on the size of the available private (i.e., excludable) benefits, and the expected probability of success. Because each citizen's *marginal* contribution to the success probability of a large protest is negligible, it does not affect their decision.

[5]The only other paper in this literature that explicitly studies altruism is Shadmehr (2021), but focusing on a static setting.

[6]See Lichbach (1995) for a broad survey of protester motivations.

Many other models focus on non-tangible rewards—often taking a black-box approach and assuming "warm-glow" payoffs from expressing discontent (Persson and Tabellini, 2009; Little, Tucker and LaGatta, 2015; Egorov and Sonin, 2021). If these payoffs are obtained only when the protest succeeds ("pleasure in agency", e.g., in Morris and Shadmehr 2023), they operate similarly to private benefits; if obtained regardless of the outcome, even the overall probability of success becomes unimportant.

In either approach, there is no role for *marginal* success probabilities, and hence public benefits, if the population is large. Since both private benefits and warm glow payoffs are typically assumed to be independent, *e.g.*, of expectations about the future status quo, citizens behave myopically in equilibrium in existing models. (We revisit the comparison with other approaches in Section 7.)

The closest papers to ours, Angeletos et al. (2007) and Little (2017), both study repeated global games.[7] In both models, a population of agents—driven by private benefits—choose whether to attack a regime in each of many periods. (Little (2017) extends Angeletos et al. (2007), allowing the game to continue after a coup with a new regime.) In both models, the agents, though fully rational, behave myopically. Thus, in the first period, agents play as in a static global game, *regardless of continuation values*. Dynamics arise because—unlike in our setting—regime strength is assumed to be fixed, so survival today creates common knowledge tomorrow that the regime is strong enough to have survived, leading to equilibrium multiplicity. In particular, there is an equilibrium where no attack occurs after the first period. However, if new information arrives over time, repeated attacks are possible once the signal of past survival has lost its relevance. In this family of models, the logic of collective procrastination does not arise, and signals of *future* regime strength or payoffs cannot affect equilibrium behavior.

A broader literature on regime change models mass protests as global games (Carlsson and Van Damme, 1993), though usually assuming a single opportunity to attack. Papers in this literature often focus on how different information structures shape coordination, and how different groups interact. Hollyer, Rosendorff and Vreeland (2015) and Little (2012), for example, study how macroeconomic indicators

[7]In global games, first used by economists to study coordination games such as currency attacks (Morris and Shin, 1998), players obtain noisy information (e.g., about the stability of a regime) and then act simultaneously. The inability to coordinate behavior perfectly due to slight differences in information typically yields equilibrium uniqueness in static models.

and electoral results respectively can act as public signals that catalyze coordination. Such signals are generated endogenously in Casper and Tyson (2014): failed mass protests reveal anti-regime sentiment, inducing elites to attempt a coup. Boix and Svolik (2013) examine the role of information generated by power-sharing agreements in coordinating behavior by elites. In all of these papers, as here, actions are strategic complements. In Tyson and Smith (2018), the regime has both opponents and adherents; actions are strategic substitutes across groups. Another strand of the literature considers interventions by the regime to change payoffs or manipulate information (Angeletos, Hellwig and Pavan, 2006; Edmond, 2013).

Although not about regime change, Chassang and Padró i Miquel (2010) and Chassang (2010) do incorporate forward-looking concerns in a dynamic coordination game. Both papers study two-player[8] dynamic cooperation games with exit: when one player exits (*e.g.*, attacks the other) the game ends and both players receive terminal payoffs. Their model is related to a variant of ours, discussed in Appendix B, in which the game ends when the protest fails rather than when it succeeds. The assumption that the game ends when cooperation fails leads to different incentives and results—in particular, cycles and procrastination do not arise.

Another strand of the literature on dynamic attacks assumes a single attack which agents can join at different times, and studies intra-attack dynamics (Dasgupta, 2007; Shadmehr and Bernhardt, 2019). In these models, extremists may protest first, but all citizens are tempted to wait and join a protest later—to gain information about the state from others' actions and ensure they are not left as the lone protester. Thus, both free-riding and *bandwagoning* or *cascades* (Kuran, 1991; Lohmann, 1994) are possible. These effects do not appear in our model: since each period represents a different protest, there is no such thing as joining a protest "later."

In many ways, our assumption of altruistic citizens mirrors a literature that addresses the paradox of voting. Models of turnout with selfish voters are known to predict unrealistically low turnout (Feddersen, 2004; Blais, 2000). High turnout in large elections is better explained by models with a "civic duty" (Feddersen and Sandroni, 2006; Coate and Conlin, 2004) or altruistic motive, even if the weight placed on others' welfare is small (Edlin, Gelman and Kaplan, 2007; Jankowski, 2007; Fowler, 2006; and especially Myatt 2015). Similarly, private benefits rarely accrue to millions

[8]As discussed above and in the example below, forward-looking concerns can arise in dynamic coordination games, even without altruism, if the population is finite.

of protesters; a model of selfish protesters that accounts for this should predict very low turnout among non-insiders. But protesting, like voting, is a form of civic expression, and arguably *the* closest substitute for voting available in a non-democratic society, so it can plausibly be explained by similar motives as voting turnout. Moreover, the assumption of altruism (rather than a purely expressive "warm glow" payoff) preserves a role for (non-excludable) instrumental concerns, in the same way that turnout in models of civic or altruistic voting is responsive to election closeness, but purely expressive turnout is not.

# 3 Two-Player Example

To build intuition, we start with a simple two-player, two-period example before presenting our model with many players and altruism. Because each citizen truly can change the outcome when the population is small, pivotality concerns (and the attendant dynamic incentives our paper studies) arise even without altruism.

Two citizens choose whether to protest in each of two periods. Protesting costs $c$ for each citizen. If *both* citizens protest in a given period, the regime falls; otherwise, it survives. Regime change in period $t$ gives both players a payoff $\theta_t \sim N(\mu, \sigma_\theta^2)$ and ends the game. At the end of period 2, the game ends even if the regime survives. Future payoffs are discounted by $\delta < 1$.

Consider first the case of full information and no state uncertainty ($\sigma_\theta^2 = 0$; regime change payoff is $\mu$). If $\mu < c$, then full abstention is both socially optimal and the only (subgame perfect) equilibrium. On the other hand, if $\mu > c$, then having both citizens protest in both periods is both socially optimal and *an* equilibrium of the noncooperative game (though there are others, as is common in coordination games): deviating to abstention in period 2 lowers the deviator's payoff from $\mu - c$ to 0, and abstaining in period 1 lowers the deviator's payoff from $\mu - c$ to $\delta(\mu - c)$. These results do not change qualitatively if there is some state uncertainty ($\sigma_\theta^2 > 0$ small) but the state in each period is commonly observed by both players. Thus, with full information, collective procrastination is suboptimal and also need not arise in equilibrium.

Suppose now, however, that $\sigma_\theta^2 > 0$, and observations of the state are slightly noisy: at the beginning of each period, each citizen privately observes a signal $x_{it} = \theta_t + \epsilon_{it}$, where $\epsilon_{it} \sim N(0, \sigma_\epsilon^2)$; then they simultaneously choose whether to protest.

We will focus on limits of equilibria as $\sigma_\epsilon^2 \to 0$.

Now, in period 2, a unique[9] equilibrium is selected in which each player $i$ protests when $x_{i2}$ is above a threshold $x_2^* = 2c$. The reason is that, when $i$ sees a signal that makes her indifferent ($x_{i2} = x_2^*$), it is equally likely that the other citizen has seen a higher signal ($x_{j2} > x_2^*$, hence $j$ protests) or a lower one ($x_{j2} < x_2^*$, hence $j$ abstains), so $i$'s expected payoff from protesting is $\frac{\theta_2}{2} + \frac{0}{2} - c$ (where $\theta_2 \approx x_{i2} = x_2^*$, since signals are very precise), while her payoff from abstention is zero.

Now consider period 1. If $\mu < 2c$ and $\sigma_\theta^2$ is small, then in most cases both players will abstain in period 2 and, by the same argument, in period 1 as well. If $\mu > 2c$, then the protest equilibrium will be most likely selected in period 2. So $i$'s continuation payoff if the regime survives period 1 is approximately $\delta(\mu - c)$. At the threshold signal value $x_1^*$ that makes $i$ indifferent in period 1, her payoff if she attacks is now only approximately $\frac{\theta_1}{2} + \frac{\delta(\mu-c)}{2} - c \approx \frac{x_1^*}{2} + \frac{\delta(\mu-c)}{2} - c$, because there is about a 50% chance that the other citizen receives a lower signal and stays home. Her payoff from abstention is about $\delta(\mu - c)$. Then, if $i$ is indifferent,

$$\frac{x_1^*}{2} + \frac{\delta(\mu - c)}{2} - c \approx \delta(\mu - c) \implies x_1^* \approx \delta(\mu - c) + 2c,$$

which is higher than the second-period threshold, $2c$.

To summarize, when $\mu < c$, both citizens (efficiently) stay home. When $c \leq \mu < 2c$, both citizens (inefficiently) stay home, due to the inefficiency of risk-dominant equilibria.[10] When $\mu > \frac{2-\delta}{1-\delta}c$, both citizens likely protest in both periods.[11] But, when $2c < \mu < \frac{2-\delta}{1-\delta}c$, the citizens pass in period 1 and attack in period 2.

As the social planner's solution shows, waiting is inefficient: if it's ever optimal to protest, players should protest in both periods. But, when information is noisy, the opportunity to protest in period 2 makes it even harder to coordinate on protesting in period 1. In fact, when $2c < \mu < \frac{2-\delta}{1-\delta}c$, the citizens would be better off if protesting in period 2 were impossible: they would still protest only once, but at least the protest

[9] There is also a no participation equilibrium, but it disappears if a lone protester can overthrow the regime with any arbitrarily small yet positive probability.

[10] Intuitively, if $\mu$ is only slightly above $c$, protesting is optimal if both players do it, but being the lone protester is costlier than forgoing a profitable protest, so protesting is unwise if there is mutual uncertainty about what the other citizen will do. Notice that this static inefficiency arises even when there is only one period.

[11] When $\mu > \frac{2-\delta}{1-\delta}c$, we have $\mu > \delta(\mu - c) + 2c$, so that in most cases $\theta_1 > \delta(\mu - c) + 2c$, and hence $x_{i2} > \delta(\mu - c) + 2c \approx x_1^*$.

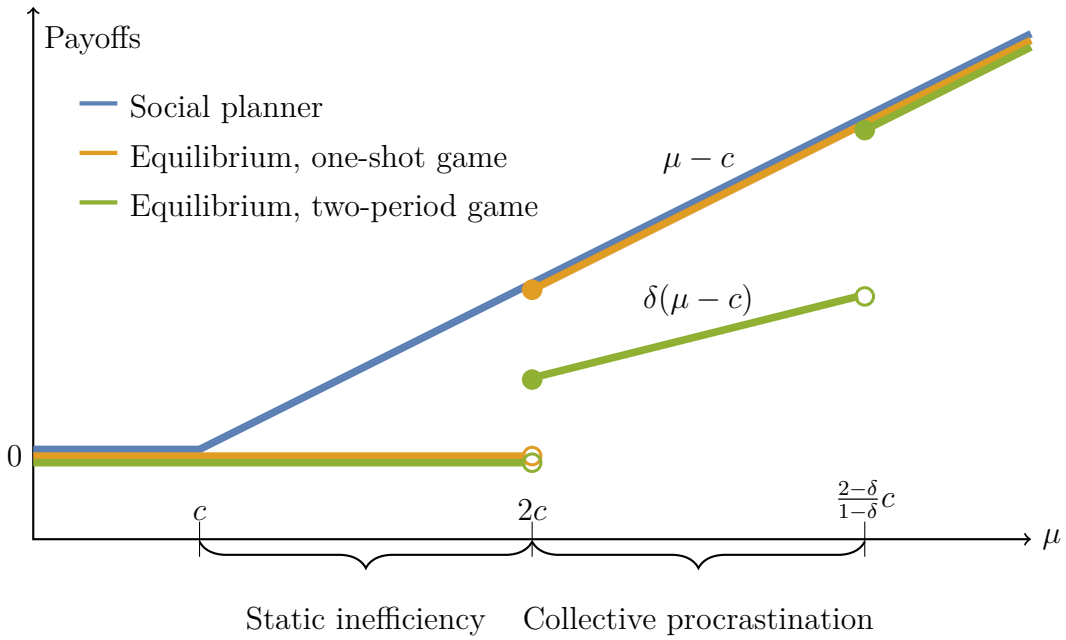


Figure 1: Ex ante payoffs in equilibrium, when there is either one or two opportunities to protest, and socially optimal payoff, as a function of $\mu$.

would not be delayed. This is visualized in Figure 1.

# 4 The Model

Now we extend the analysis to a set $N$ of players who repeatedly choose whether to attack or not. In our main specification, the set of players is a continuum: $N = [0, 1]$. To clarify some issues concerning the scaling of payoffs and pivotal probabilities as the population grows, we briefly discuss the case of a finite population in Section 5.

Time is discrete and finite: $t \in \{0, 1, \ldots, T\}$. The payoffs from a successful attack in period $t$ are governed by a parameter $\theta_t \sim N(\mu_t, \sigma_\theta^2)$, drawn independently across periods.

The information structure and timing of the game are as follows. At the beginning of each period $t$, if the game has not yet ended, Nature draws the value of $\theta_t$ and then reveals to each player $i$ a signal

$$x_{it} = \theta_t + \epsilon_{it},$$

where $\epsilon_{it} \sim N(0, \sigma_\epsilon^2)$ is independent across players and periods.

Each player $i \in N$ then simultaneously chooses to attack ($a_{it} = 1$) or abstain ($a_{it} = 0$). These actions result in the regime being overthrown with probability $f(l_t)$, where $l_t$ denotes the fraction of the population who attack in period $t$. If the regime falls, the players receive some terminal payoffs, described below, and the game ends. With probability $1 - f(l_t)$, the game continues in the next period. (At the end of period $T$, the game ends even if the regime survives.) We assume that $f$ is smooth, increasing and convex. More formally, we assume that $f$ is twice continuously differentiable; $0 \leq f(0) < f(1) \leq 1$; $f'(l) > 0$ and $f''(l) > 0$ for all $l$; and $0 < \inf_{l \in (0,1)} f''(l) \leq \sup_{l \in (0,1)} f''(1) < \infty$. A simple example is given by any quadratic function, $f(l_t) = b_0 + b_1 l_t + b_2 l_t^2$, with $b_0 \geq 0$, $b_1, b_2 > 0$, and $b_0 + b_1 + b_2 \leq 1$.

## Payoffs

We allow the players' preferences to reflect some degree of altruism, measured by a parameter $\alpha \in [0, 1]$. To make this explicit, we distinguish between each player $i$'s *hedonic* flow payoff in period $t$, $u_{it}$, and her flow *utility* in period $t$, $v_{it}$, defined as

$$v_{it} = u_{it} + \alpha \sum_{j \neq i} u_{jt}. \tag{1}$$

In other words, each player puts weight $\alpha$ on the well-being of each other player, and weight 1 on her own. Thus $\alpha = 0$ models completely self-interested players, while $\alpha = 1$ models fully altruistic players that consider the welfare of others just as important as their own, as a social planner would. (Note that Equation (1) only yields a well-defined utility function if the population is finite. However, the resultant expression for the players' *marginal* payoff from attacking—which is the key object of interest—extends in a natural way to the case of an infinite population. See Section 5 for details.)

The players have a common discount factor $\delta \in (0, 1)$. We denote $i$'s discounted hedonic payoffs from period $t$ onwards by $U_{it}$, defined as

$$U_{it} = \sum_{t \leq \tau \leq T} \delta^{\tau - t} u_{i\tau}.$$

Hedonic payoffs are as follows. Each agent $i$ who attacks in a period $t$ bears a flow cost of attacking $c > 0$ in that period. If the regime falls in period $t$, then all agents

also receive a one-time payoff $\theta_t$ defined above, and the game ends. If the regime survives in period $t$, all agents instead accrue a known *status quo* flow payoff $\nu_t$, and the game moves on to the next period.[12] Note that all agents receive either $\theta_t$ or $\nu_t$, as appropriate, *regardless* of whether they attacked in that period.

Our solution concept is Perfect Bayesian Equilibrium.

## Assumptions: Interpretation and Discussion

In many ways, our model takes after existing workhorse models of protests in the global games literature. We depart from the standard assumptions when necessary to obtain a model that clearly highlights the forces we are interested in. Some of these departures are worth discussing.

First, we assume that the benefits from a successful revolt are public. Although there is evidence that both private and public benefits matter in practice (Cantoni, Yang, Yuchtman and Zhang, 2019; Muller and Opp, 1986), models in this literature typically focus on private benefits (Angeletos et al., 2007; Edmond, 2013; Little, 2017)—an exception is Shadmehr (2021). In Section 6, we show that the general logic of our results survives if we allow for both private and public benefits.

Second, we allow for some degree of altruism. This assumption is what keeps public benefits relevant in the agents' benefit-cost calculation as the population becomes large and, hence, the probability of being pivotal goes to zero. Section 3 shows that our results arise even without altruism if the population is small—in general, they hold whenever there are multiple periods and the logic of pivotality is present.

Third, the payoff from revolution is affected by the state of the world, $\theta_t$, but the probability of a successful revolt, $f(l_t)$, is not *directly* affected by the state. A natural interpretation is that $\theta_t$ parameterizes the expected outcome after a revolution—for example, the ideology or competence of a *de facto* opposition leader—rather than the regime's ability to stave off protesters. This assumption is for simplicity; qualitatively similar results hold if there is uncertainty about the function $f$, or other payoff parameters such as $\nu_t$ or $c$.

[12] As written, the model assumes that, after period $T$, there are no more protesting opportunities and also no more status quo payoffs. We could, however, assume that status quo payoffs $\nu_{T+1}$, $\nu_{T+2}$, ... keep accruing forever if the regime survives through period $T$. Adding such "post-terminal" payoffs to the model is equivalent to bundling them into the period-$T$ status quo payoff, i.e., setting $\tilde{\nu}_T = \sum_{t \geq T} \delta^{t-T} \nu_t$.

It is worth comparing our setup to the two most popular payoff specifications in global games. In some models (Morris and Shin, 2003; Little, 2016), attackers receive $\theta+l-1$, while abstainers receive 0. Our model generates a similar specification for the *marginal* payoff from attacking. But these papers have no concept of an attack being "successful" or "unsuccessful" in a binary sense. In another variant (Morris and Shin, 1998; Dasgupta, 2007; Angeletos et al., 2007; Shadmehr, 2021; Little, 2017; Shadmehr and Bernhardt, 2019; Edmond, 2013), attackers receive $1-c$ if successful and $-c$ if unsuccessful, but only succeed if $l \geq 1-\theta$. This specification is inconvenient for our purposes because it only yields a supermodular game when pivotality concerns—which are central to our analysis—are absent.

Fourth, we assume that regime change ends the game. This assumption is less substantively restrictive than it might appear: the payoff $\theta_t$ represents the citizens' expected continuation utility from a new regime starting in period $t+1$. The new regime could itself face protests. Such possibilities are all captured by the payoff $\theta_t$.

Fifth, the probability of a successful revolt, $f(l_t)$, is strictly increasing and convex in the size of the protest. The convexity assumption guarantees supermodularity in the presence of pivotality concerns, by ensuring that the marginal impact of an additional protester is higher the more protesters there are.[13] This assumption best models settings in which overthrowing the regime is "hard" and requires a large mass of people to show up, whereas concavity of $f$ might be natural if even a moderate crowd is sufficient, and so there are diminishing returns when $l_t$ is large. The model is not intractable if we assume that $f$ is concave—leading to strategic substitutability, as found by Cantoni et al. (2019)—though the equilibrium strategies would involve some degree of mixing, and procrastination would no longer arise due to fears of miscoordination.[14]

Finally, we assume that the state of the world $\theta_t$ is drawn independently across periods. This contrasts with Angeletos et al. (2007) and Little (2017), in which the state is fixed over time. However, our model allows the *mean* of the state in each period to follow an arbitrary sequence $(\mu_t)_{t=0,1,\ldots,T}$. In Section 6 we show that persistent shocks can be accommodated, if any information about them is *commonly observed*; the key assumption keeping our model tractable is that the *idiosyncratic* uncertainty about $\theta_t$—which supports unique equilibrium selection—is transient. (In

[13] In a model with no pivotality concerns it is enough to assume that $f$ is increasing.

[14] Procrastination could still arise due to the free-riding effect discussed at the end of Section 5.

our analysis, we focus on the case of $\sigma_\epsilon^2$ small, so it is substantively unimportant whether the idiosyncratic shocks are persistent or transient.)

# 5 Analysis

We solve the game by backward induction from the last period. Suppose the regime has survived until the beginning of period $T$. What is left to play is a static coordination game, which can be solved using familiar techniques from the global games literature.

Let $\Delta_{iT}$ be $i$'s *marginal* payoff from attacking, given a signal observation $x_{iT}$ and the other players' equilibrium strategies. In equilibrium, $i$ must attack if $\Delta_{iT} > 0$ and abstain if $\Delta_{iT} < 0$.

To see how marginal payoffs should be calculated, it is instructive to consider the case of a large but finite population.[15] Suppose that $N = \{1, \ldots, n\}$. Then the marginal payoff from attacking is

$$-c + E\left[(1 + \alpha(n-1))\,(\theta_T - \nu_T)\left(f\left(\tilde{l}_T + \frac{1}{n}\right) - f(\tilde{l}_T)\right) \mid x_{iT}\right],$$

where $c$ is the cost of protesting, $\alpha$ is the altruism parameter, $n$ is population size, $\theta_T$ and $\nu_T$ are payoffs from regime change and status quo in period $T$ respectively, $f(l)$ is the probability of regime change when fraction $l$ of citizens attack, and $\tilde{l}_T \equiv \frac{1}{n}\sum_{j\neq i} a_{jT}$ is the fraction of the population who attacks, assuming $i$ abstains. As $n \to \infty$, both $\tilde{l}_t$ and $\tilde{l}_t + \frac{1}{n}$ converge to $l_t$, while

$$(1 + \alpha(n-1))\left(f\left(\tilde{l}_t + \frac{1}{n}\right) - f(\tilde{l}_t)\right) \to \alpha f'(l_t).$$

Note that, if the agents are even slightly altruistic (i.e., for any $\alpha > 0$), pivotality is taken into account, and public benefits matter, even as the population grows. This happens because the total gain from a successful revolt (approximately $\alpha n$) increases proportionally with population size, while an agent's probability of being pivotal (approximately $\frac{f'(l_t)}{n}$) decreases proportionally; these two forces offset each other.

[15] It is preferable not to work directly with a finite population in the main model, because in that case the distribution of signals would be random even conditional on the state, complicating the analysis.

With a continuous population, then, the net payoff of attacking is:

$$\Delta_{iT} = -c + E\left[\alpha(\theta_T - \nu_T)f'(l_T) \mid x_{iT}\right]. \tag{2}$$

Our first result characterizes the agents' equilibrium behavior in the last period.

**Lemma 1.** *Assume $\sigma_\epsilon > 0$ is small enough. Then the period-T subgame has a unique equilibrium. In this equilibrium, each player $i$ attacks if and only if $x_{iT}$ is weakly greater than a threshold $x_T^*(\sigma_\epsilon)$. Moreover, as $\sigma_\epsilon \to 0$, $x_T^*(\sigma_\epsilon)$ converges to a limit $x_T^*$, which equals*

$$x_T^* = \frac{c}{\alpha[f(1) - f(0)]} + \nu_T.$$

The equilibrium threshold yields intuitive comparative statics. Higher costs of protesting $c$ and better status quo payoffs $\nu_T$ both drive up the threshold $x_T^*$, discouraging protesting; greater "agency", $f(1) - f(0)$, and altruism $\alpha$ encourage it. That the unique equilibrium is in threshold strategies follows from familiar arguments for global games. Here is an intuitive derivation of the equilibrium threshold $x_T^*$. A citizen $i$ whose signal $x_{iT}$ equals $x_T^*$ must be indifferent, i.e., $\Delta_{iT}(x_T^*) = 0$. When $\sigma_\epsilon$ is small, $x_{iT}$ is a precise signal of the state, so $i$ believes that $\theta_T$ is close to $x_T^*$. On the other hand, $i$'s signal says very little about where it ranks *relative to other citizens' signals*; as is standard in global games (Morris and Shin, 2003), $i$ expects that the fraction of citizens with higher signals than her own is approximately uniformly distributed between 0 and 1. Because it is precisely those citizens who will attack, $l_t | x_{iT} = x_T^*$ is approximately uniform between 0 and 1. Substituting all this into Equation (2),

$$\Delta_{iT}(x_T^*) \approx -c + \alpha(x_T^* - \nu_T)\int_0^1 f'(l)dl = -c + \alpha(x_T^* - \nu_T)[f(1) - f(0)].$$

Setting this expression equal to zero yields the limit threshold from Lemma 1.[16]

Our next observation is that the full game can be solved using exactly the same approach, with one difference. Denote by $\overline{U}_{t+1}(\sigma_\epsilon, \sigma_\theta)$ the expected value of any agent's hedonic continuation payoffs at the beginning of period $t+1$, assuming the regime has survived until then. Then $i$'s marginal utility from attacking in period $t$

[16]Note that the equilibrium strategy in the limit is the Laplacian action, that is, the best response to a uniform prior over others' actions (Morris and Shin, 2003).

is

$$\Delta_{it} = -c + E\left[\alpha\left(\theta_t - \nu_t - \delta\overline{U}_{t+1}(\sigma_\epsilon, \sigma_\theta)\right) f'(l_t) \mid x_{it}\right],$$

because regime change attains the payoff $\theta_t$ but, in the process, forgoes both the current status quo payoff $\nu_t$ and the continuation payoff $\delta\overline{U}_{t+1}(\sigma_\epsilon, \sigma_\theta)$, which captures future payoffs from both protests and the status quo. Our next result traces out the consequences of this observation.

**Proposition 1.** *Assume $\sigma_\epsilon$ is small enough. Then the game has a unique equilibrium. In this equilibrium, each player $i$ attacks in period $t$ if and only if $x_{it}$ is weakly greater than a threshold $x_t^*(\sigma_\epsilon, \sigma_\theta)$. Moreover, as $\sigma_\epsilon \to 0$, we have $x_t^*(\sigma_\epsilon, \sigma_\theta) \to x_t^*(\sigma_\theta)$ and $\overline{U}_{t+1}(\sigma_\epsilon, \sigma_\theta) \to \overline{U}_{t+1}(\sigma_\theta)$. And as $\sigma_\theta \to 0$, we have $x_t^*(\sigma_\theta) \to x_t^*$, $\overline{U}_{t+1}(\sigma_\theta) \to \overline{U}_{t+1}$. The sequence of limit thresholds and continuation utilities $(x_0^*, \ldots, x_T^*; \overline{U}_0, \ldots, \overline{U}_T)$ is found by recursively solving the following system of equations for $t = T, T-1, \ldots, 0$:*

$$x_t^* = \frac{c}{\alpha[f(1) - f(0)]} + \nu_t + \delta\overline{U}_{t+1}; \tag{3}$$

$$\overline{U}_t = \begin{cases} -c + f(1)\mu_t + (1 - f(1))\left(\nu_t + \delta\overline{U}_{t+1}\right) & \text{if } \mu_t > x_t^* \\ f(0)\mu_t + (1 - f(0))\left(\nu_t + \delta\overline{U}_{t+1}\right) & \text{if } \mu_t < x_t^*, \end{cases} \tag{4}$$

*taking $\overline{U}_{T+1} = 0$.*

Per Equation (3) the equilibrium threshold in all periods is as in Lemma 1, but augmented to account for the continuation value $\delta\overline{U}_{t+1}$ of preserving the status quo. Note that, when $\sigma_\epsilon$ and $\sigma_\theta$ are both low, $x_{it}$ is close to $\mu_t$ for most citizens. Then, in periods where $\mu_t > x_t^*$, a mass protest takes place ($l_t \approx 1$) and the regime falls with probability close to $f(1)$. On the contrary, when $\mu_t < x_t^*$, almost nobody protests, and the regime falls with probability close to $f(0)$. This observation underpins Equation (4). Equation (3) then reveals that mass protests occur precisely in periods where $\mu_t - \nu_t > \frac{c}{\alpha[f(1)-f(0)]} + \delta\overline{U}_{t+1}$: a high current gain from regime change, $\theta_t - \nu_t \approx \mu_t - \nu_t$, encourages protests, but so does a low continuation value $\delta\overline{U}_{t+1}$.

In fact, protests are always welfare-improving in equilibrium: whenever $\mu_t > x_t^*$, the net payoff of a mass protest, $-c + [f(1) - f(0)](\mu_t - \nu_t - \delta\overline{U}_{t+1})$ (as per Equation (4)) is at least $-c + \frac{c}{\alpha}$, which is positive whenever altruism is imperfect ($\alpha < 1$). Then the expectation that citizens will coordinate on a protest in period $t+1$ discourages protests in $t$ by increasing $\delta\overline{U}_{t+1}$, while the expectation that citizens will coordinate on abstention tomorrow spurs protests today. This leads to *cycles of protest*.

To illustrate, consider the example shown in Figure 2, where $f(l) = \frac{l+l^2}{4}$, $T = 5$, $c = \alpha = 0.1$, $\delta = 0.8$, and $\sigma_\epsilon$, $\sigma_\theta$ are both small, with $\sigma_\epsilon << \sigma_\theta$. We assume $\mu_t = 3$ and $\nu_t = 0$ for all $t$, so regime change and status quo payoffs are constant. Then there appears to be no reason to wait for a "better" moment (*i.e.*, higher $\theta_t$) to attack; attacks ought to make sense in every period, or never. Yet, in equilibrium, the citizens condition their actions today on expected future attacks, leading to cycling. Indeed, in period 5, $\mu - \nu = 3 > 2 = \frac{0.1}{0.1[0.5-0]} = \frac{c}{\alpha[f(1)-f(0)]}$, so there is a protest in period 5. But as a result, the continuation value in period 4, $\delta\overline{U}_5$, equals $0.8(-0.1+0.5\times 3) = 1.12$, a value high enough that it tempts the citizens to abstain in period 4, as $\frac{c}{\alpha[f(1)-f(0)]} + \delta\overline{U}_5 = 3.12 > 3$. In period 3, citizens are more impatient because they would have to wait two full periods for the next protest: $\delta^2\overline{U}_5 = \delta\overline{U}_4 = 0.8 \times 1.12 = 0.896$, so $\frac{c}{\alpha[f(1)-f(0)]} + \delta\overline{U}_4 = 2.896 < 3$, and a protest occurs in period 3. By similar logic, the citizens abstain in periods 1 and 2, and protest in period 0, having a 50% chance of success ($f(1) = 0.5$) with each attack.[17]

This example also reveals that having additional opportunities to protest can be harmful: for example, conditional on reaching period 4, the citizens' equilibrium utility is 1.12, but it would be 1.4 if protesting in period 5 were impossible, because they would then coordinate on attacking in period 4. Thus, the availability of future protest opportunities can induce *collective procrastination*. In general terms, changes to the environment which slightly increase the agents' payoffs *given any strategy profile*—but discourage them from protesting—may leave them worse off in equilibrium.

Because the expectation of an imminent attack discourages attacking today, it is generally true that, if the profitability of attacks is in an intermediate region, attacks arrive in waves separated by periods of apparent calm, even if the underlying fundamentals—the level of discontent, the state of the economy, and so on—remain stable. The following proposition formalizes this argument.

**Proposition 2.** *Suppose the status quo payoff $\nu_t$ equals $\nu$ for all periods $t < T$, with $\nu_T = \frac{\nu}{1-\delta}$.*[18] *Then there are thresholds $\mu_0 \leq \mu_* < \mu^*$ such that, for $\sigma_\epsilon << \sigma_\theta$ small enough:*

*(i) If $\mu_t = \mu > \mu^*$ for all $t$, then, in every period, almost everyone attacks.*

[17] Note that, because $\sigma_\theta$ is small, citizens effectively know when they will next coordinate on a protest. When $\sigma_\theta$ is substantial, a similar logic holds in fuzzier form.

[18] This amounts to assuming status quo payoffs of size $\nu$ for period $T$ and all periods thereafter (see Footnote 12), which keeps continuation values constant over time in case of no attacks.

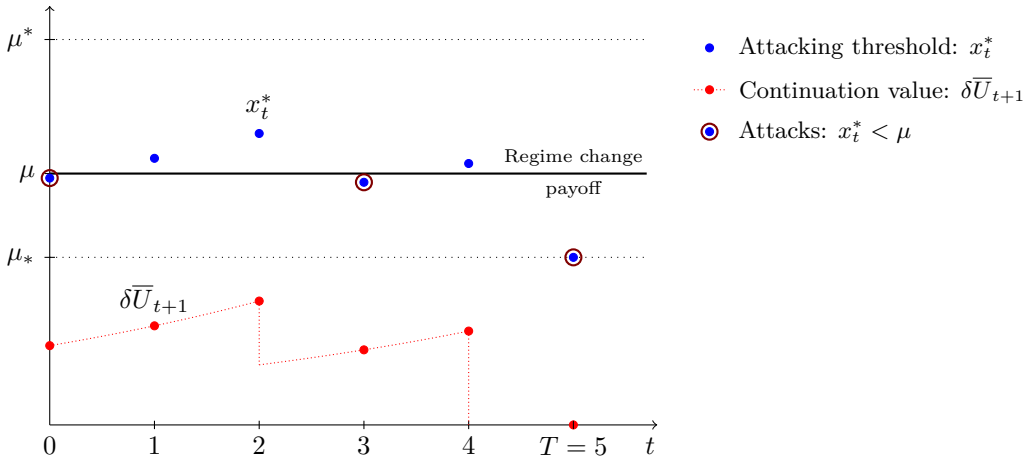


Figure 2: Pattern of attacks when the regime change payoff is intermediate: $\mu_t \equiv \mu$ between $\mu_*$ and $\mu^*$

*(ii)* *If $\mu_t = \mu < \mu_0$ for all $t$, then, in every period, almost everyone abstains.*

*(iii)* *Generically,*[19] *if there is $\eta > 0$ for which $\mu_t \in (\mu_* + \eta, \mu^* - \eta)$ for all $t$, there are protest cycles: for $T$ large enough, there are arbitrarily many periods in which most players attack, and arbitrarily many periods in which most players abstain.*

*Moreover*

$$\mu_0 = \frac{c}{\alpha[f(1) - f(0)]} + \frac{\nu}{1-\delta},$$
$$\mu_* = \frac{c}{\alpha[f(1) - f(0)]} + \frac{\delta c}{1-\delta}\frac{f(0)}{\alpha[f(1) - f(0)]} + \frac{\nu}{1-\delta},$$
$$\mu^* = \frac{c}{\alpha[f(1) - f(0)]} + \frac{\delta c}{1-\delta}\left[\frac{f(1)}{\alpha[f(1) - f(0)]} - 1\right] + \frac{\nu}{1-\delta}.$$

An important implication of Proposition 2 is that, as $\delta \to 1$, $\mu^* - \frac{\nu}{1-\delta}$ grows without bound. This means that, if citizens are very patient, cycles of protest are almost inevitable: protesting in every period becomes impossible unless the payoff to protesting is extremely high ($\mu_t > \mu^*$).

Finally, Proposition 3 characterizes the model's comparative statics in a limited sense. It shows the effects of a marginal change in the parameters—in particular, $\mu_{t'}$

[19] The statement is true except for a set of sequences $(\mu_t)_t$ of Lebesgue measure zero.

or $\nu_{t'}$—on the incentive to attack in any period $t \leq t'$, measured by changes in the equilibrium thresholds $x_t^*$.

**Proposition 3.** *Consider the generic case in which $\mu_t \neq x_t^*$ for all $t$. Assume $0 < f(0) < f(1) < 1$. Then:*

*(i) A marginal increase in the current or future status quo payoff increases the current threshold for attack: $\frac{\partial x_t^*}{\partial \nu_{t'}} > 0$ for all $t' \geq t$.*

*(ii) A marginal increase in the payoff of future regime change increases the current threshold for attack, but a change in the payoff of current regime change does not affect it: $\frac{\partial x_t^*}{\partial \mu_{t'}} > 0$ for all $t' > t$ but $\frac{\partial x_t^*}{\partial \mu_t} = 0$.*

Explicit formulas for the derivatives $\frac{\partial x_t^*}{\nu_{t'}}$, $\frac{\partial x_t^*}{\mu_{t'}}$ are given in the Appendix. The intuition behind the result is as follows: when the status quo payoff, $\nu_{t'}$, or the regime change payoff, $\mu_{t'}$, increases in some future period $t' > t$, it becomes more attractive to let the regime survive at time $t$, for a chance to receive this increased payoff at time $t'$. Then the incentive to attack in period $t$ decreases, and $x_t^*$ increases. Similarly, if $\nu_t$ increases, the players are incentivized to let the regime survive today. On the other hand, an increase in $\mu_t$ has no effect on $x_t^*$—but makes players more likely to attack at time $t$, since it increases $\theta_t$, and thus the players' signals $x_{it}$. The general message is that an attractive status quo always deters attacks, while an attractive regime change payoff today encourages attacks *now* while discouraging attacks *in previous periods*.

When information is precise, Proposition 3 characterizes only *latent* changes in the willingness to attack: for example, if $\mu_t < x_t^*$, then there will be no attack at time $t$, a conclusion left unaffected by any marginal parameter change. If a parameter changes enough, collective behavior eventually changes discontinuously, and perhaps simultaneously in multiple periods. For instance, as $\nu_{t'}$ increases, all the thresholds $x_t^*$ for $t < t'$ smoothly increase, up to the point when one of them crosses $\mu_t$ from below. At that point, the agents would suddenly switch from attacking in period $t$ to abstaining, and this expectation may in turn galvanize them to attack in a previous period $t''$, etc.

We finish our analysis with a discussion of the social planner's solution, as well as the equilibrium of the game, in the benchmark case of full information ($\sigma_\epsilon = 0$). This comparison highlights that it is the fear of miscoordination under noisy information that causes collective procrastination and cycles of protest.

*Remark* 1. Assume $\sigma_\epsilon = 0$. In the social planner's solution, all agents attack in period $t$ if $\theta_t$ is higher than a threshold

$$x_t^{\text{sp}} = \frac{c}{[f(1) - f(0)]} + \nu_t + \delta \overline{U}_{t+1}^{\text{sp}}, \tag{5}$$

and all agents abstain if $\theta_t < x_t^{\text{sp}}$. Moreover,

(i) Citizens' payoffs weakly increase if $\mu_t$ or $\nu_t$ increase for any $t$.

(ii) If regime change and status quo payoffs are constant ($\mu_t \equiv \mu$, $\sigma_\theta = 0$, $\nu_t \equiv \nu$ for all $t < T$ and $\nu_T = \frac{\nu}{1-\delta}$), and the regime never falls without a protest ($f(0) = 0$), then either there is an attack in every period (if $\mu > \frac{c}{[f(1)-f(0)]} + \frac{\nu}{1-\delta}$) or there are no attacks ($<$).

(iii) If $\alpha f'(1) \geq f(1) - f(0)$, the social planner's solution is also an equilibrium of the game with perfect signals ($\sigma_\epsilon = 0$).

Per Equation (5), the social planner uses essentially the same threshold for action that the agents would use in the equilibrium of our main model (cf. Equation (3)) *if* they were fully altruistic ($\alpha = 1$). Part (i) of Remark 1 implies that, in the social planner's solution, there is no procrastination: a higher continuation payoff is always weakly beneficial, as the social planner chooses to wait only when waiting is the best option. Part (ii) reveals that there are no spurious cycles: if fundamentals are stable, then the social planner has the agents always attack or never attack. Finally, part (iii) reveals that, if altruism $\alpha$ is not too low and the incentive to coordinate is relatively strong,[20] then the social planner's solution is also an equilibrium of the game when the state $\theta_t$ is commonly known in each period; it is the addition of noise, as in our main model, that makes this equilibrium unattainable.

It is worth noting that, in fact, the gap between the social planner's solution and the equilibrium with noise (compare Equations (3) and (5)) stems from two distinct forces. Besides the fear of miscoordination that we have highlighted, there is also a simple free-riding effect at work: if we expect to coordinate on protesting today, but a citizen $i$ deviates by abstaining, she may still see the regime fall thanks to the

[20]The expression $\frac{f'(1)}{f(1)-f(0)} = \frac{f'(1)}{\int_0^1 f'(l)dl}$ measures coordination motives: if $f$ is linear, then one's incentive to protest is independent of other citizens' participation, and $\frac{f'(1)}{f(1)-f(0)} = 1$; if $f$ is steeply convex, then $\frac{f'(1)}{f(1)-f(0)} \gg 1$.

efforts of others—a relevant temptation if citizens are imperfectly altruistic.[21] The condition $\alpha f'(1) \geq f(1) - f(0)$ in part (iii) of Remark 1 ensures that agents are altruistic enough, or it is imperative enough to participate in a burgeoning protest, that free riding does not preclude the socially optimal outcome in the absence of noise. When $\alpha f'(1) < f(1) - f(0)$, even the best equilibrium under full information may also feature procrastination and cycles, purely due to the free-riding effect.

# 6 Extensions

This Section presents two extensions. The first shows how the model may accommodate more general forms of uncertainty and informational shocks. The second shows how the results change if private benefits (*i.e.*, "club goods") are present in addition to public benefits. In Section B of the Online Appendix, we cover an alternative setting where, unlike in our main model, there is no hope of overthrowing the regime, but protests serve to keep a resistance alive and stave off permanent repression.

## 6.1 Generalized Uncertainty

We assumed for simplicity that the citizens face idiosyncratic uncertainty about their payoff from regime change, $\theta_t$. We could have obtained similar results by instead assuming that they face idiosyncratic uncertainty regarding their status quo payoff, $\nu_t$. In particular, it would still be true that, as information becomes precise, most citizens attack in period $t$ if

$$\mu_t > \frac{c}{\alpha[f(1) - f(0)]} + E[\nu_t] + \delta \overline{U}_{t+1},$$

and most abstain if the reverse strict inequality holds. Qualitatively similar results are obtained if we instead assume uncertain and time-varying costs of protesting, $c$.

Perhaps more importantly, we can allow for very general uncertainty and learning about *future* payoff parameters. For example, we can assume that, for each $t$, $\mu_t$ and $\nu_t$ are distributed according to some cumulative distribution functions $F_t$, $G_t$, with their realized values being fully revealed by the beginning of period $t$—but this information can arrive as a lump sum at time $t$, or in a previous period, or gradually

[21] Note that, in our two-player example, the temptation to free-ride is completely absent, because a protest backed by only one citizen cannot succeed.

over many periods, with all such signals being revealed publicly to all citizens. Because this uncertainty is resolved by time $t$, it makes no difference when characterizing the citizens' equilibrium strategies at time $t$. The only change to our analysis is that we must write a more complicated version of Equation (4), as the expected continuation value $\overline{U}_{t+1}$ must now take into account that the parameters $\mu_{t+1}$ and $\nu_{t+1}$, the players' equilibrium actions, and the next period's continuation value, $\overline{U}_{t+2}$, may take many possible values.

Adding this kind of uncertainty to the model allows us to think about the equilibrium response to information about future shocks. For example, let $f(l) = \frac{2l+l^2}{8}$, $c = 1$, $\delta = 0.8$, and $\alpha = \frac{4}{33}$. Assume that $\mu_t \equiv 1$, but $\nu_t$ depends on the *state* of the society, which may be *green*, *yellow*, or *red*. We can think of these as different stages of democratic backsliding, where green corresponds to the status quo, yellow to the introduction of bills that will entrench the incumbent in power, and red to after the bill has been ratified. Alternatively, these colors can capture different levels of social strife, where green is peaceful, yellow is tension, and red corresponds to conflict. Either way, while the state is green or yellow, $\nu_t = 0$, whereas $\nu_t = \underline{\nu} < 0$ in the red state. If the state is green at time $t$, then, at time $t+1$, it will still be green with probability 0.98; with probability 0.02, it will turn yellow. If the state turns yellow in period $t$, it remains in this state for three periods ($t$, $t+1$, $t+2$) and then becomes red forever. Note that the yellow state is not materially worse than the green one—it just denotes that citizens are *aware* of an imminent slide to the red state.

Suppose that the state turns yellow in period $t_0$. Using Equations (3) and (4), we can show that citizens attack in every red period (i.e., from $t_0 + 3$ onwards) if $\underline{\nu} < -10$. Moreover, if $\underline{\nu} < -20$, citizens also attack in the last yellow period, $t_0 + 2$; if $\underline{\nu} < -40$, they also attack in period $t_0 + 1$; and if $\underline{\nu} < -80$, they also attack in period $t_0$, that is, as soon as the state becomes yellow. Of course, citizens cannot "preemptively" attack in period $t_0 - 1$ because they do not know when the state will turn yellow until they see it; and they will not attack in the green state so long as $\underline{\nu} > -1800$. A crucial assumption underpinning this example is that $f(1) < 1$: because even an all-out attack is not guaranteed to topple the regime, and the red state is very costly, the citizens would do well to begin attacking ahead of time if the red state is approaching. The more costly this state is, the earlier they begin to attack. Only when $\underline{\nu}$ is extremely negative (in particular, $\underline{\nu} < -1800$), the citizens attack even in the *green* state, in an attempt to dodge a future red state that may

not ever materialize. A general lesson from this example is that, in our model, the citizens may respond proactively to a threat that the status quo will worsen in the future, or that future opportunities to protest may disappear.

## 6.2 Private and Public Benefits

For simplicity, in our main model there are *only* public benefits from protesting: any payoff from regime change benefits all citizens. We can instead allow for the coexistence of public and *private* benefits that are only obtained by participants in a successful attack. Suppose that a fraction $\rho$ of regime change benefits are private: if the regime falls at time $t$ protesters receive $\theta_t$ and abstainers receive only $(1-\rho)\theta_t$. ($\rho \in [0,1]$ is a commonly known parameter.) Then $i$'s marginal payoff from protesting at time $t$ becomes

$$\Delta_{it} = -c + E\left[\alpha((1-\rho)\theta_t + l_t\rho\theta_t - \nu_t - \delta\overline{U}_{t+1})f'(l_t) + \rho\theta_t f(l_t) \mid x_{it}\right], \tag{6}$$

where $\rho\theta_t f(l_t)$ is the expected private benefit received by $i$, and $\alpha(1-\rho)\theta_t f'(l_t)$ and $\alpha l_t \rho\theta_t f'(l_t)$ are, respectively, $i$'s valuation of others' public *and* private benefits generated by $i$'s own participation. Under the assumptions made in the main model, the game remains one of strategic complements, so the citizens attack when $x_{it}$ is above a threshold, converging to a limit $x_t^*$ when $\sigma_\epsilon^2$ is small enough, specifically:

$$x_t^* = \frac{c + \alpha[f(1) - f(0)](\nu_t + \delta\overline{U}_{t+1})}{(1-\rho)\alpha[f(1) - f(0)] + \rho(\alpha f(1) + (1-\alpha)\int_0^1 f(l)dl)}. \tag{7}$$

A derivation of Equation (7) can be found in the Appendix. Note that, when $\rho = 0$, this simplifies to Equation (3).

This extension yields two insights. First, it shows that adding private benefits does not fundamentally alter the strategic logic of the problem: so long as pivotality concerns are active (due to either altruism or a small population), continuation values matter in the citizens' strategic calculus (*i.e.*, $\delta\overline{U}_{t+1}$ appears in Equation (7)), and similar arguments as in our main model show that protest cycles and procrastination can result. This is true *even if all material benefits are private* ($\rho = 1$).

Second, it allows us to cleanly compare our model to canonical models of protest (Morris and Shin, 2003; Angeletos et al., 2007; Little, 2017) in which there is a con-

tinuum of citizens; private benefits are available; and there are no altruistic concerns. We can obtain a model with these properties within our framework by setting $\rho > 0$ and $\alpha = 0$. Equation (7) becomes

$$x_t^* = \frac{c}{\rho \int_0^1 f(l)dl}. \tag{8}$$

The continuation utility, $\overline{U}_{t+1}$, vanishes from the expression: non-atomic and selfish agents will act *as if* myopic, *even when they are forward-looking*, because the value of continuing the game matters in their strategic calculus only insofar as their participation might affect the probability of regime change, which it cannot. Therefore, information about the future becomes irrelevant in the selfish model of protests.[22]

Although selfish protesters are less motivated to act than a social planner would like in a static setting, they may be inefficiently slow *or* quick to act in a dynamic setting, precisely because they disregard the future in their calculations. In the context of an improving environment ($\mu_t$, $\nu_t$ increasing over time) selfish citizens might unnecessarily "jump the gun," chasing a short-term payoff. In contrast, they might fail to react to an approaching catastrophe ($\mu_t$, $\nu_t$ sharply decreasing) if current regime change payoffs are not tempting enough.

Finally, our simple "selfish model" contains no mechanism leading to protest cycles: the threshold in Equation (8) is constant over time, so if $\mu_t$ is constant, there will be attacks in all periods (if $\mu_t > x_t^*$) or none ($<$). Existing papers within this framework (Angeletos et al., 2007; Little, 2017) show that intermittent attacks are possible *if the state is hidden and persistent* (*i.e.*, $\theta$ is drawn only once and remains fixed). The resulting linkage across periods is that after a failed attack, citizens know that the regime was (and remains) strong enough to have survived. This negative signal discourages further attacks unless new information $x_{it}$ suggests that the regime is in fact not too strong. In other words, the logic behind waves of protest is informational and backward-looking, in contrast to the strategic, forward-looking logic driving Proposition 2.

[22] Again, this is true when the population is large. Otherwise, even selfish agents care about being pivotal, and engage in forward-looking behavior.

# 7 Discussion

Our model generates empirical predictions that do not follow from existing models of protests. For instance, public benefits can drive protest behavior, whether private benefits are present or not. "News shocks" can be impactful: threatening bills can cause protests, while a promise to hold elections can defuse them. And higher prosociality (higher $\alpha$) should increase participation, as well as make participation more sensitive to the aforementioned forces (e.g., news shocks). Some of these predictions have empirical support: Cantoni et al. (2019) find that more prosocial citizens in Hong Kong were more likely to protest, and Muller and Opp (1986) find that protest behavior responds more to "public goods incentives" than to selective incentives or psychological rewards. These and other predictions warrant further testing.

As a preliminary exercise, it is worth examining the key events of some recent protest movements through the lens of our model to see how its predictions can map to reality. The aforementioned Hong Kong protests, sparked by the February 2019 introduction of a proposed extradition bill,[23] are an example in which it is apparent that protesters were motivated by forward-looking considerations. In June, when the bill would have been discussed at the Legislative Council (Purbrick, 2019), the protests peaked, leading to clashes with police. Further protests followed, now against the bill, the police crackdowns, and the government's condemnation of the protests as riots. The bill was then suspended indefinitely.[24] Crowds peaked at as many as two million participants. Pro-democracy candidates, previously a minority, captured over 80% of District Council seats at the November 2019 local elections. The conflict echoed massive protests in 2003 against a proposed national security bill, as well as the 2014 Umbrella Revolution, which condemned a proposal to implement democratic elections but only between candidates selected by a pro-Beijing committee. These explosions of dissent punctuated a rising collective unease with the mainland's attempts to encroach on Hong Kong's autonomy, described as "the political ground simultaneously shifting and shrinking beneath their feet," all this against the backdrop of a ticking clock, as the terms of the 1997 Sino-British Joint Declaration that delineates the "one country, two systems" framework would formally expire in 2047.[25] Finally,

[23] https://www.nytimes.com/2019/06/09/world/asia/hong-kong-extradition-protest.html

[24] https://www.nytimes.com/2019/06/16/world/asia/hong-kong-protests.html

[25] https://time.com/5786776/hong-kong-joshua-wong-future-homeland/

in June of 2020, the mainland National People's Congress, bypassing the local government, imposed a national security law that criminalized dissent in Hong Kong. The law had an immediate chilling effect on protests and led to the disbandment of pro-democracy parties, raids on media offices, and mass arrests of activists and opposition politicians.[26]

It is worth highlighting three facts. First, the 2003, 2014, and 2019 protests all began in response to *proposed* bills or reforms, which had not yet had any material consequences but could be taken as *signs* that the future and autonomy of Hong Kong were quickly deteriorating. Thus, citizens demonstrated forward-looking protest participation. Second, even though there was common knowledge that the "one country, two systems" framework would expire in 2047, it took concrete threats that the system would be subverted ahead of schedule, and imminently, to spur them to act. Citizens thus showed signs of collective procrastination, as their strongest attempts at extracting concessions through mobilization took place when the government had already shifted towards a hard-line approach, with Xi Jinping having made bold moves to centralize authority and minimize internal dissent in China throughout the 2010s. Third, protests *grew* in response to police brutality, which arguably showed that the worst was yet to come, and so the time to act was now.

Through the lens of our model, we can thus see the protests as an increasingly desperate resistance which responded to future threats but only when the prospect of disaster became imminent. In contrast, to explain the 2019 surge in protest behavior, models of "selfish protesting" would need to allege an increase in private benefits from success; a decrease in the cost of protesting; or a weakening of the regime which made it a tempting target. While private benefits may drive the behavior of leaders and activists, it cannot reasonably explain the participation of millions of people, and the other explanations run counter to the facts (as the protesters faced a regime that had dug in its heels and was willing to respond with violence). Another possible rationalization is that, in a coordination game, any idiosyncratic event could trigger collective action by shifting the players' "focal point"—but this explanation would chalk the consistent response to threatening legislation up to coincidence. In particular, in these models, continuation utilities have no role to play, *even if* individual citizens are rational and forward-looking.

[26] `https://www.nytimes.com/2021/01/05/world/asia/hong-kong-arrests-national-security-law.html`

Finally, one may plausibly explain the observed protest behavior as an emotional response to grievances (Passarelli and Tabellini, 2017; Gibilisco, 2021; Correa, Nandong and Shadmehr, 2021), yet this explanation is incomplete without a model of *why* and *when* certain events aggrieve people.[27]

The 2014 Euromaidan revolution in Ukraine is another recent example. After years of negotiations with the European Union and promises of European integration, the Yanukovych administration announced in November of 2013 that it was suspending plans to sign a broad political and economic association agreement with the EU, only a week before the scheduled signing. Instead, Ukraine would seek closer ties with Russia, which had threatened retaliatory trade sanctions in response to the EU deal. Protesters gathered the same day, angry that their hopes to finally escape the Russian sphere of influence—to no longer live in "a post-Soviet barrack temporarily repainted in yellow and blue"—were quickly vanishing.[28] Ukraine failed to sign the EU agreement as scheduled, even as both sides claimed that a deal was still on the table.[29] The protests grew in number and scope of demands, and turned into riots after the government responded with a violent crackdown.[30] The situation worsened further after the government passed a package of draconian anti-protest laws in January,[31] and reached its nadir in February, with over 100 protesters being killed by police amid escalating clashes. Soon, widespread desertion among demoralized police forces forced Yanukovych to flee to Russia.[32] Again, a pattern emerges of citizens protesting in response to signals of a worsening future (or a dashed hope of improvement), after years of inaction despite a bleak outlook, and strengthening their resolve in the face of violence and draconian measures signaling a turn towards dictatorship. These and other examples do not readily fit an image of self-interested protesters opportunistically chasing the spoils of victory, or responding in knee-jerk

[27] A viable approach, which yields similar results to ours, would be to assume a form of "pleasure in agency" that accounts for the collective gain in net present value from a successful protest; $\Delta_{it}$ would then equal $E[f(l_t)(\theta_t - \nu_t - \delta\overline{U}_{t+1}) \mid x_{it}]$. We thank Mehdi Shadmehr for this observation.

[28] https://www.nytimes.com/2013/11/27/world/europe/protests-continue-as-ukraine-leader-defends-stance-on-europe.html

[29] https://www.reuters.com/article/us-ukraine-eu/eu-says-door-remains-open-to-ukraine-as-unity-cracks-idUSBRE9BE05120131216

[30] https://www.nytimes.com/2013/12/02/world/europe/thousands-of-protesters-in-ukraine-demand-leaders-resignation.html

[31] https://www.washingtonpost.com/world/in-ukraine-protesters-appear-to-be-preparing-for-battle/2014/01/20/904cdc72-81bd-11e3-9dd4-e7278db80d86_story.html

[32] https://www.nytimes.com/2014/02/24/world/europe/as-his-fortunes-fell-in-ukraine-a-president-clung-to-illusions.html?_r=1

fashion to material deprivation.

# 8 Conclusions

In this paper we develop a dynamic model of protests in which citizens act *as if* they are somewhat likely to be pivotal, and hence may act even if the benefits from regime change accrue to all citizens, including non-participants. The strategic calculus we uncover arises if the population is small, or irrespective of population size if citizens are altruistic. We show that, in a dynamic context, the willingness to engage in "pivotal protesting" responds not just to contemporaneous benefits and costs, but also to the future ramifications of regime change or its absence. In particular, altruistic citizens may protest in response to an event that increases the cost of protesting if it also paints a bleak picture of the future, as is the case with police crackdowns or authoritarian measures. Because an expectation of future collective action makes present collective action less urgent, and vice versa, spikes in social turmoil are self-limiting and may arrive in waves, even if the underlying material and social conditions are stable over time. Moreover, because *partially* altruistic citizens act only when collective action is socially beneficial by a wide enough margin, the mere existence of future chances to act may tempt them to drag their feet today, leaving them worse off.

The dynamic encouragement and discouragement effects that are central to our analysis are absent from models of repeated mass protests driven by private benefits. Within our theory, they are the source of novel predictions which, in our view, translate into more natural conceptualizations of many protest processes. (Though we discussed two prominent examples, the emergence of protests in response to negative expectations and state violence appears to be a general phenomenon.) In assuming limited altruism, we aim to anchor the formal literature on social movements closer to the substantive literature, capturing notions such as public-mindedness, grievances, and other moral considerations that undeniably play a role in civic behavior.

The model is flexible and allows many extensions besides the ones covered in the paper. One salient question concerns government manipulation: if indeed collective action is vulnerable to a form of collective "limited willpower," how would a government shape payoffs or beliefs over time to defuse protests? For example, the government may increase clientelistic transfers when the threat of revolt spikes, or promise to hold new elections as an alternative to immediate resignation.

A more challenging direction is to enrich the informational environment. For instance, the government may have private information about its strength or willingness to repress dissent, while citizens may have private information about their level of discontent. Signaling concerns would then arise: citizens may mobilize to communicate, rather than just to overthrow the government, and the government may repress to show strength or resolve.

Another possible avenue for future work is to model other instances of civic behavior under our partial altruism framework. While altruism has been proposed as an explanation for turnout, other activities that have received less attention, such as campaigning for a candidate, are plausibly motivated by civic-mindedness and also involve dynamic coordination between potential supporters.

# References


**Angeletos, George-Marios, Christian Hellwig, and Alessandro Pavan**, "Signaling in a Global Game: Coordination and Policy Traps," *Journal of Political Economy*, 2006, *114* (3), 452–484.

**_ , _ , and _** , "Dynamic global games of regime change: Learning, multiplicity, and the timing of attacks," *Econometrica*, 2007, *75* (3), 711–756.

**Blais, André**, *To Vote or Not to Vote?: The Merits and Limits of Rational Choice Theory*, University of Pittsburgh Press, 2000.

**Boix, Carles and Milan W Svolik**, "The foundations of limited authoritarian government: Institutions, commitment, and power-sharing in dictatorships," *The Journal of Politics*, 2013, *75* (2), 300–316.

**Bueno De Mesquita, Ethan and Mehdi Shadmehr**, "Rebel Motivations and Repression," *American Political Science Review*, 2023, *117* (2), 734–750.

**Cantoni, Davide, David Y. Yang, Noam Yuchtman, and Y. Jane Zhang**, "Protests as strategic games: experimental evidence from Hong Kong's antiauthoritarian movement," *The Quarterly Journal of Economics*, 2019, *134* (2), 1021–1077.

**Carlsson, Hans and Eric Van Damme**, "Global games and equilibrium selection," *Econometrica*, 1993, pp. 989–1018.

**Casper, Brett Allen and Scott A Tyson**, "Popular Protest and Elite Coordination in a Coup détat," *The Journal of Politics*, 2014, *76* (2), 548–564.

**Chassang, Sylvain**, "Fear of miscoordination and the robustness of cooperation in dynamic global games with exit," *Econometrica*, 2010, *78* (3), 973–1006.

**_ and Gerard Padró i Miquel**, "Conflict and deterrence under strategic risk," *The Quarterly Journal of Economics*, 2010, *125* (4), 1821–1858.

**Coate, Stephen and Michael Conlin**, "A Group Rule-Utilitarian Approach to Voter Turnout: Theory and Evidence," *American Economic Review*, 2004, *94* (5), 1476–1504.

**Correa, Sofía, Gaetan Nandong, and Mehdi Shadmehr**, "Crises, Catharses, and Boiling Frogs: Path Dependence in Collective Action," *Available at SSRN 3906282*, 2021.

**Dasgupta, Amil**, "Coordination and delay in global games," *Journal of Economic Theory*, 2007, *134* (1), 195–225.

**DeGroot, Morris H.**, *Optimal Statistical Decisions*, McGraw-Hill, 1970.

**Edlin, Aaron, Andrew Gelman, and Noah Kaplan**, "Voting as a Rational Choice: Why and How People Vote To Improve the Well-Being of Others," *Rationality and Society*, 2007, *19* (3), 293–314.

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

**Egorov, Georgy and Konstantin Sonin**, "Elections in Non-Democracies," *The Economic Journal*, 2021, *131* (636), 1682–1716.

**Feddersen, Timothy and Alvaro Sandroni**, "A Theory of Participation in Elections," *American Economic Review*, 2006, *96* (4), 1271–1282.

**Feddersen, Timothy J.**, "Rational Choice Theory and the Paradox of Not Voting," *Journal of Economic Perspectives*, March 2004, *18* (1), 99–112.

**Fowler, James H**, "Altruism and Turnout," *The Journal of Politics*, 2006, *68* (3), 674–683.

**Gibilisco, Michael**, "Decentralization, Repression, and Gambling for Unity," *The Journal of Politics*, 2021, *83* (4), 1353–1368.

**Greene, William H.**, *Econometric Analysis*, fifth ed., Prentice Hall, 2003.

**Gurr, Ted Robert**, *Why Men Rebel*, Princeton University Press, 1970.

**Hollyer, James R, B Peter Rosendorff, and James Raymond Vreeland**, "Transparency, protest, and autocratic instability," *American Political Science Review*, 2015, *109* (4), 764–784.

**Jankowski, Richard**, "Altruism and the Decision to Vote: Explaining and Testing High Voter Turnout," *Rationality and Society*, 2007, *19* (1), 5–34.

**Kuran, Timur**, "Now out of Never: The Element of Surprise in the East European Revolution of 1989," *World Politics*, 1991, *44* (1), 748.

**Lichbach, Mark Irving**, *The Rebel's Dilemma*, University of Michigan Press, 1995.

**Little, Andrew T**, "Elections, fraud, and election monitoring in the shadow of revolution," *Quarterly Journal of Political Science*, 2012, *7* (3), 249–283.

**Little, Andrew T.**, "Communication technology and protest," *The Journal of Politics*, 2016, *78* (1), 152–166.

_ , "Coordination, learning, and coups," *Journal of Conflict Resolution*, 2017, *61* (1), 204–234.

**Little, Andrew T, Joshua A Tucker, and Tom LaGatta**, "Elections, protest, and alternation of power," *The Journal of Politics*, 2015, *77* (4), 1142–1156.

**Lohmann, Susanne**, "The Dynamics of Informational Cascades: The Monday Demonstrations in Leipzig, East Germany, 1989-91," *World Politics*, 1994, *47* (1), 42–101.

**Milgrom, Paul and John Roberts**, "Rationalizability, Learning, and Equilibrium in Games with Strategic Complementarities," *Econometrica*, 1990, pp. 1255–1277.

**Morris, Stephen and Hyun Song Shin**, "Unique equilibrium in a model of self-fulfilling currency attacks," *American Economic Review*, 1998, pp. 587–597.

_ **and** _ , "Global games: Theory and applications," in "Advances in Economics and Econometrics: Theory and Applications, Eighth World Congress, Volume 1" Cambridge University Press 2003, pp. 56–114.

_ **and Mehdi Shadmehr**, "Inspiring regime change," *Journal of the European Economic Association*, 2023, p. jvad023.

**Muller, Edward N. and Karl-Dieter Opp**, "Rational Choice and Rebellious Collective Action," *The American Political Science Review*, 1986, *80* (2), 471–488.

**Myatt, David P**, "A Theory of Voter Turnout," 2015. http://dpmyatt.org/uploads/turnout-2015.pdf.

**Olson, Mancur**, *The Logic of Collective Action*, Cambridge University Press, 1965.

**Özel, Soli**, "A Moment of Elation: The Gezi Protests/Resistance and the Fading of the AKP Project," in Umut Özkrml, ed., *The Making of a Protest Movement in Turkey: #occupygezi*, London: Palgrave Macmillan UK, 2014, pp. 7–24.

**Passarelli, Francesco and Guido Tabellini**, "Emotions and Political Unrest," *Journal of Political Economy*, 2017, *125* (3), 903–946.

**Persson, Torsten and Guido Tabellini**, "Democratic Capital: The Nexus of Political and Economic Change," *American Economic Journal: Macroeconomics*, 2009, *1* (2), 88–126.

**Purbrick, Martin**, "A Report of The 2019 Hong Kong Protests," *Asian Affairs*, 2019, *50* (4), 465–487.

**Shadmehr, Mehdi**, "Protest Puzzles: Tullock's Paradox, Hong Kong Experiment, and the Strength of Weak States," *Quarterly Journal of Political Science*, 2021, *16* (3), 245–264.

_ **and Dan Bernhardt**, "Vanguards in revolution," *Games and Economic Behavior*, 2019, *115*, 146–166.

**Tullock, Gordon**, "The paradox of revolution," *Public Choice*, 1971, pp. 89–99.

**Tyson, Scott A and Alastair Smith**, "Dual-layered coordination and political instability: Repression, co-optation, and the role of information," *The Journal of Politics*, 2018, *80* (1), 44–58.

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

# A Appendix

*Proof of Lemma 1.* The general strategy of the proof follows four steps:

(i) Show that the game is supermodular in actions, that is, if others' strategies increase in the sense of attacking at more signal realizations, then any player's incentive to attack also increases.

(ii) Show that the best response to a symmetric threshold strategy profile is a threshold strategy. Using standard arguments from the supermodular games literature, conclude that the game has extremal equilibria in symmetric threshold strategies.

(iii) Show that the game has a unique equilibrium in symmetric threshold strategies, hence a unique equilibrium.

(iv) Characterize the equilibrium threshold, in particular as $\sigma_\epsilon \to 0$.

The proof follows standard approaches for global and, more generally, supermodular games. There are two complications, however, that make the proof less than standard. First, the game is not supermodular in the traditional sense; we show instead that a closely related game (with the same set of equilibria) is supermodular.[33] Second, for the purpose of proving Proposition 1, we need a stronger result than stated in this Lemma: not only do we need to show the existence of a threshold $\overline{\sigma}_\epsilon > 0$ such that if $\sigma_\epsilon < \overline{\sigma}_\epsilon$ then the equilibrium is unique (and in threshold strategies), but we also need $\overline{\sigma}_\epsilon$ to be uniformly bounded away from zero as parameters vary (in particular as $\nu_T$ varies), because the game in periods $t < T$ has an a priori uncertain continuation value $\nu_t + \delta \overline{U}_{t+1}$ that is itself a function of $\sigma_\epsilon$.

**(i) Supermodularity.** Formally, denote $j$'s strategy in period $T$ by $A_{jT}$, the set of realizations of $x_{jT}$ for which $j$ attacks. We will impose as an additional technical condition that a strategy $A_T = (A_{jT})_{j\in[0,1]}$ can only be compatible with equilibrium if $A_T$ is Lebesgue measurable as a subset of $\mathbb{R} \times [0, 1]$. (Otherwise, objects such as the fraction of attackers in equilibrium may not be well defined.)

[33] Lemma 2.3 in Morris and Shin (2003) deals with a similar failure of supermodularity to the one discussed below, but their result assumes a uniform prior and does not rule out equilibria that are not in threshold strategies.

Let $(A_{jT})_{j\in[0,1]}$, $(\tilde{A}_{jT})_{j\in[0,1]}$ be two strategy profiles such that $A_{jT} \subseteq \tilde{A}_{jT}$ for all $j$. The standard approach would be to show that $\Delta_{iT}(x_{iT}) \leq \tilde{\Delta}_{iT}(x_{iT})$ for all $i$, $x_{iT}$, where $\Delta_{it}(x_{iT})$ is as defined in Equation (2) and $\tilde{\Delta}_{iT}(x_{iT})$ is the analogous object when other players instead use the strategy $\tilde{A}_T$. However, this inequality does not necessarily hold. Intuitively, if $A_T$ and $\tilde{A}_T$ differ only in that agents attack more under $\tilde{A}_T$ when their signal realizations are very *low*, then an agent who expects others to play according to $\tilde{A}_T$ may be less willing to attack, because she is afraid that $f'(l_T)$—hence the effect of her participation—will be higher precisely when $\theta_T - \nu_T$ is negative, a case in which she would prefer *not* to topple the regime.

We then need a more careful argument. Our strategy is to argue that (a) agents never want to attack when their signals are relatively low, no matter what they expect others to do, and (b) when restricting attention to strategies that respect this constraint, supermodularity does hold.

*Remark* 2. (DeGroot, 1970, Theorem 9.5.1) $\theta_T|x_{iT} \sim N\left(\frac{\sigma_\epsilon^2\mu_T+\sigma_\theta^2 x_{iT}}{\sigma_\theta^2+\sigma_\epsilon^2}, \frac{\sigma_\theta^2\sigma_\epsilon^2}{\sigma_\theta^2+\sigma_\epsilon^2}\right)$.

*Remark* 3. Let $X \sim N(\mu, \sigma^2)$. Then $E(X|X > a) \leq \max\left(a + \sqrt{\frac{2}{\pi}}\sigma, \mu + \sqrt{\frac{2}{\pi}}\sigma\right)$.

*Proof.* Follows immediately from the inverse Mills ratio formula (see Greene (2003), p. 759). □

**Lemma 2.** *There is $\overline{\sigma}_\epsilon^2 > 0$ such that, if $\sigma_\epsilon^2 < \overline{\sigma}_\epsilon^2$, then no agent with a signal below $\frac{c}{2\alpha f'(1)}+\nu_T$ ever attacks. Moreover, we can take $\overline{\sigma}_{\epsilon 1}^2 = \min\left(\left(\frac{c}{4\alpha f'(1)}\right)^2, \sigma_\theta^2 \frac{c}{4\alpha f'(1)|\mu_T-\nu_T|}\right)$.*

*Proof.* Note that, for any $x_{iT} \leq \frac{c}{2\alpha f'(1)} + \nu_T$,

$$\begin{aligned}
\Delta_{iT}(x_{iT}) &= -c + \alpha \int_{-\infty}^{\infty} (\theta_T - \nu_T) f'(l_T(\theta_T)) g(\theta_T|x_{iT}) d\theta_T \leq \\
&\leq -c + \alpha \int_{\nu_T}^{\infty} (\theta_T - \nu_T) f'(l_T(\theta_T)) g(\theta_T|x_{iT}) d\theta_T \leq \\
&\leq -c + \alpha f'(1) \int_{\nu_T}^{\infty} (\theta_T - \nu_T) g(\theta_T|x_{iT}) d\theta_T \leq \\
&\leq -c + \alpha f'(1)(E(\theta_T|x_{iT}, \theta_T \geq \nu_T) - \nu_T) \leq \\
&\leq -c + \alpha f'(1) \left[\max\left(\frac{\sigma_\epsilon^2\mu_T + \sigma_\theta^2 x_{iT}}{\sigma_\theta^2+\sigma_\epsilon^2}, \nu_T\right) + \sqrt{\frac{2}{\pi}} \frac{\sigma_\theta\sigma_\epsilon}{\sqrt{\sigma_\theta^2+\sigma_\epsilon^2}} - \nu_T\right] \leq \\
&\leq -c + \alpha f'(1) \left[\max\left(\frac{\sigma_\epsilon^2(\mu_T - \nu_T) + \sigma_\theta^2 \frac{c}{2\alpha f'(1)}}{\sigma_\theta^2+\sigma_\epsilon^2}, 0\right) + \sigma_\epsilon\right]
\end{aligned}$$

where $g(\theta_T|x_{iT})$ is the posterior density of the state given $i$'s signal $x_{iT}$.

There are two cases. If $\sigma_\epsilon^2(\mu_T-\nu_T)+\sigma_\theta^2\frac{c}{2\alpha f'(1)} \leq 0$, then the above expression equals $-c+\alpha f'(1)\sigma_\epsilon$, which is negative whenever $\sigma_\epsilon < \frac{c}{\alpha f'(1)}$. If $\sigma_\epsilon^2(\mu_T-\nu_T)+\sigma_\theta^2\frac{c}{2\alpha f'(1)} > 0$, then the expression equals

$$\begin{aligned}
&-c+\alpha f'(1)\left[\frac{\sigma_\epsilon^2(\mu_T-\nu_T)+\sigma_\theta^2\frac{c}{2\alpha f'(1)}}{\sigma_\theta^2+\sigma_\epsilon^2}+\sigma_\epsilon\right] \leq \\
\leq &-c+\alpha f'(1)\left[\frac{\sigma_\epsilon^2}{\sigma_\theta^2}(\mu_T-\nu_T)+\frac{c}{2\alpha f'(1)}+\sigma_\epsilon\right] = -\frac{c}{2}+\alpha f'(1)\left[\frac{\sigma_\epsilon^2}{\sigma_\theta^2}(\mu_T-\nu_T)+\sigma_\epsilon\right]
\end{aligned}$$

which is at most $-\frac{c}{4}+\alpha f'(1)\frac{\sigma_\epsilon^2}{\sigma_\theta^2}(\mu_T-\nu_T)$ if $\sigma_\epsilon \leq \frac{c}{4\alpha f'(1)}$. This expression is negative whenever $\sigma_\epsilon^2 < \sigma_\theta^2\frac{c}{4\alpha f'(1)|\mu_T-\nu_T|}$. □

Now assume $\sigma_\epsilon^2 < \overline{\sigma}_{\epsilon 1}^2$ and consider a modified game in which, when an agent $i$ receives a signal $x_{iT} < \frac{c}{2\alpha f'(1)}+\nu_T$, she is forced mechanically to abstain, while for $x_{iT} \geq \frac{c}{2\alpha f'(1)}+\nu_T$ she is allowed to choose an action as usual. This game clearly has the same set of equilibria as the original. Next, we argue that it is supermodular for $\sigma_\epsilon^2$ small enough.

**Lemma 3.** *Assume that* $\sigma_\epsilon^2 < \overline{\sigma}_{\epsilon 2}^2 = \min\left(\overline{\sigma}_{\epsilon 1}^2, \sigma_\theta^2\frac{c}{4\alpha f'(1)}\frac{1}{|\mu_T-\nu_T-\frac{c}{4\alpha f'(1)}|}, \left(\frac{c}{4\alpha f'(1)}\right)^2\frac{1}{\ln(\overline{f''})-\ln(\underline{f''})}\right)$, *where* $\overline{f''} = \sup_{l\in(0,1)} f''(l)$, $\underline{f''} = \inf_{l\in(0,1)} f''(l)$. *Then, in the restricted game where actions are chosen only when* $x_{iT} \geq \frac{c}{2\alpha f'(1)}+\nu_T$, $\Delta_{iT}(x_{iT})$ *is weakly increasing in* $A_T$.

*Proof.* Consider two strategy profiles $A_T$, $\tilde{A}_T \subseteq [\frac{c}{2\alpha f'(1)}+\nu_T, \infty)\times[0,1]$, such that $A_{jT} \subseteq \tilde{A}_{jT}$ for all $j$. For any $i$ and any $x_{iT} \geq \frac{c}{2\alpha f'(1)}+\nu_T$, we will compare $\Delta_{iT}(x_{iT})$ to $\tilde{\Delta}_{iT}(x_{iT})$. To simplify notation, we will drop the $T$ indices. We then have

$$\begin{aligned}
\tilde{\Delta}_i(x_i)-\Delta_i(x_i) =&\alpha\int_{-\infty}^{\infty}(\theta-\nu)[f'(\tilde{l}(\theta))-f'(l(\theta))]g(\theta|x_i)d\theta \\
=&\alpha\int_{-\infty}^{\nu}(\theta-\nu)[f'(\tilde{l}(\theta))-f'(l(\theta))]g(\theta|x_i)d\theta+ \\
&+\alpha\int_{\nu}^{\infty}(\theta-\nu)[f'(\tilde{l}(\theta))-f'(l(\theta))]g(\theta|x_i)d\theta \\
\geq&\alpha\overline{f''}\int_{-\infty}^{\nu}(\theta-\nu)[\tilde{l}(\theta)-l(\theta)]g(\theta|x_i)d\theta+ \\
&+\alpha\underline{f''}\int_{\nu}^{\infty}(\theta-\nu)[\tilde{l}(\theta)-l(\theta)]g(\theta|x_i)d\theta.
\end{aligned}$$

It is enough to show that this last expression is at least zero. Next, we note that, for all $\theta$,

$$\tilde{l}(\theta) - l(\theta) = \int_{-\infty}^{\infty} \lambda(x) \frac{1}{\sigma_\epsilon} \phi\left(\frac{x-\theta}{\sigma_\epsilon}\right) dx,$$

where $\lambda(x)$ is the additional fraction of players who attack when seeing a signal $x$ under $\tilde{A}$ relative to $A$ (i.e., $\lambda(x) = |\tilde{A}(x)| - |A(x)|$) and $\phi$ is the standard normal density function. Then it is enough to show that, for any $x \geq \frac{c}{2\alpha f'(1)} + \nu$,

$$\overline{f''} \int_{-\infty}^{\nu} (\theta - \nu) \phi\left(\frac{\theta - x}{\sigma_\epsilon}\right) g(\theta|x_i) d\theta + \underline{f''} \int_{\nu}^{\infty} (\theta - \nu) \phi\left(\frac{\theta - x}{\sigma_\epsilon}\right) g(\theta|x_i) d\theta \geq 0. \quad (9)$$

Next, we argue that the "tightest" case is when $x$ and $x_i$ are as low as possible—that is, if we show the result for $x = x_i = \frac{c}{2\alpha f'(1)} + \nu$ then it will automatically follow for all other $x$, $x_i$. The reason is that, if Equation (9) holds, then

$$\overline{f''} \int_{-\infty}^{\nu} (\theta - \nu) \phi\left(\frac{\theta - x}{\sigma_\epsilon}\right) g(\theta|x_i) \gamma(\theta) d\theta + \underline{f''} \int_{\nu}^{\infty} (\theta - \nu) \phi\left(\frac{\theta - x}{\sigma_\epsilon}\right) g(\theta|x_i) \gamma(\theta) d\theta \geq$$
$$\overline{f''} \int_{-\infty}^{\nu} (\theta - \nu) \phi\left(\frac{\theta - x}{\sigma_\epsilon}\right) g(\theta|x_i) \gamma(\nu) d\theta + \underline{f''} \int_{\nu}^{\infty} (\theta - \nu) \phi\left(\frac{\theta - x}{\sigma_\epsilon}\right) g(\theta|x_i) \gamma(\nu) d\theta \geq 0$$

for any function $\gamma(\theta)$ that is positive and weakly increasing. Moreover, by standard properties of the normal distribution, $\phi\left(\frac{\theta - x}{\sigma_\epsilon}\right)$ and $g(\theta|x_i) = \frac{\sqrt{\sigma_\theta^2 + \sigma_\epsilon^2}}{\sigma_\theta \sigma_\epsilon} \phi\left(\frac{\theta - \frac{\sigma_\epsilon^2 \mu + \sigma_\theta^2 x_i}{\sigma_\theta^2 + \sigma_\epsilon^2}}{\frac{\sigma_\theta \sigma_\epsilon}{\sqrt{\sigma_\theta^2 + \sigma_\epsilon^2}}}\right)$ are both MLRP-increasing in $x$ and $x_i$, respectively (i.e., $\frac{g(\theta|x_i')}{g(\theta|x_i)}$ is increasing in $\theta$ for $x_i' > x_i$, and $\frac{\phi\left(\frac{\theta - x'}{\sigma_\epsilon}\right)}{\phi\left(\frac{\theta - x}{\sigma_\epsilon}\right)}$ is increasing in $\theta$ for $x' > x$).

**Lemma 4.** *If* $\sigma_\epsilon^2 \leq \sigma_\theta^2 \frac{c}{4\alpha f'(1)} \frac{1}{|\nu + \frac{c}{4\alpha f'(1)} - \mu|}$, *then* $\frac{\sigma_\epsilon^2 \mu + \sigma_\theta^2 x_i}{\sigma_\theta^2 + \sigma_\epsilon^2} \geq \nu + \frac{c}{4\alpha f'(1)}$ *whenever* $x_i \geq \nu + \frac{c}{2\alpha f'(1)}$.

*Proof.* Taking $x_i = \nu + \frac{c}{2\alpha f'(1)}$, we want

$$\begin{aligned}
\sigma_\epsilon^2 \mu + \sigma_\theta^2 \left(\nu + \frac{c}{2\alpha f'(1)}\right) &\geq (\sigma_\theta^2 + \sigma_\epsilon^2) \left(\nu + \frac{c}{4\alpha f'(1)}\right) \\
\iff \sigma_\epsilon^2 \mu + \sigma_\theta^2 \frac{c}{4\alpha f'(1)} &\geq \sigma_\epsilon^2 \left(\nu + \frac{c}{4\alpha f'(1)}\right) \\
\iff \sigma_\theta^2 \frac{c}{4\alpha f'(1)} &\geq \sigma_\epsilon^2 \left(\nu + \frac{c}{4\alpha f'(1)} - \mu\right).
\end{aligned}$$

Then it is enough to take $\sigma_\epsilon^2 \leq \sigma_\theta^2 \frac{c}{4\alpha f'(1)} \frac{1}{\nu + \frac{c}{4\alpha f'(1)} - \mu}$ if $\nu + \frac{c}{4\alpha f'(1)} - \mu > 0$ and any value of $\sigma_\epsilon^2$ works otherwise. □

Now, using our previous results and Lemma 4, it is enough to show that

$$\overline{f''} \int_{-\infty}^{\nu} (\theta - \nu) \phi\left(\frac{\theta - x_0}{\sigma_\epsilon}\right) \phi\left(\frac{\theta - x_0}{\frac{\sigma_\theta \sigma_\epsilon}{\sqrt{\sigma_\theta^2 + \sigma_\epsilon^2}}}\right) d\theta +$$
$$\underline{f''} \int_{\nu}^{\infty} (\theta - \nu) \phi\left(\frac{\theta - x_0}{\sigma_\epsilon}\right) \phi\left(\frac{\theta - x_0}{\frac{\sigma_\theta \sigma_\epsilon}{\sqrt{\sigma_\theta^2 + \sigma_\epsilon^2}}}\right) d\theta \geq 0,$$

where $x_0 = \nu + \frac{c}{4\alpha f'(1)}$. In turn, it is enough to show that, for each $r \geq 0$, the value of the first integrand at $\theta = \nu - r$ is dominated by the value of the second integral at $\theta = \nu + \frac{c}{4\alpha f'(1)} + r$, i.e., it is enough to show

$$\overline{f''} r \phi\left(\frac{-r - \frac{c}{4\alpha f'(1)}}{\sigma_\epsilon}\right) \phi\left(\frac{-r - \frac{c}{4\alpha f'(1)}}{\frac{\sigma_\theta \sigma_\epsilon}{\sqrt{\sigma_\theta^2 + \sigma_\epsilon^2}}}\right) \leq \underline{f''} \left(r + \frac{c}{4\alpha f'(1)}\right) \phi\left(\frac{r}{\sigma_\epsilon}\right) \phi\left(\frac{r}{\frac{\sigma_\theta \sigma_\epsilon}{\sqrt{\sigma_\theta^2 + \sigma_\epsilon^2}}}\right)$$

for all $r \geq 0$. Rearranging, and since $r + \frac{c}{4\alpha f'(1)} \geq r$, it is enough to show that

$$e^{-\frac{1}{2\sigma_\epsilon^2}\left[\left(r + \frac{c}{4\alpha f'(1)}\right)^2 - r^2\right] - \frac{\sigma_\theta^2 + \sigma_\epsilon^2}{2\sigma_\theta^2 \sigma_\epsilon^2}\left[\left(r + \frac{c}{4\alpha f'(1)}\right)^2 - r^2\right]} \leq \frac{\underline{f''}}{\overline{f''}}$$

and hence enough to show

$$e^{-\frac{1}{\sigma_\epsilon^2}\left[\left(r + \frac{c}{4\alpha f'(1)}\right)^2 - r^2\right]} \leq \frac{\underline{f''}}{\overline{f''}}.$$

Since the left-hand side is decreasing in $r$, it is enough to show

$$e^{-\frac{1}{\sigma_\epsilon^2}\left(\frac{c}{4\alpha f'(1)}\right)^2} \leq \frac{\underline{f''}}{\overline{f''}} \iff -\frac{1}{\sigma_\epsilon^2}\left(\frac{c}{4\alpha f'(1)}\right)^2 \leq \ln(\underline{f''}) - \ln(\overline{f''}),$$

which holds whenever $\sigma_\epsilon^2 \leq \left(\frac{c}{4\alpha f'(1)}\right)^2 \frac{1}{\ln(\overline{f''}) - \ln(\underline{f''})}$. □

It follows that, when $\sigma_\epsilon < \overline{\sigma}_{\epsilon 2}$, both Lemma 2 and Lemma 3 apply, and the game (with restricted strategy space) is supermodular in actions, which implies the existence of a greatest equilibrium and a smallest equilibrium between which all other

equilibria are bounded (Milgrom and Roberts, 1990). Lemma 2 already implies the existence of a lower dominance region. We can similarly show the existence of an upper dominance region:

**Lemma 5.** *Assume that* $\sigma_\epsilon^2 < \overline{\sigma}_{\epsilon 3}^2 = \min\left(\overline{\sigma}_{\epsilon 2}^2, \sigma_\theta^2 \frac{c}{\alpha f'(0)} \frac{1}{|\nu_T + \frac{c}{\alpha f'(0)} - \mu_T|}\right)$. *Then any agent with a signal* $x_{iT} > 2\frac{c}{\alpha f'(0)} + \nu_T$ *always attacks.*

*Proof.* By Lemma 3, when $\sigma_\epsilon^2 < \overline{\sigma}_{\epsilon 2}^2$, an agent $i$'s incentive to attack is lowest if other agents never attack. In that case

$$\begin{aligned}\Delta_{iT}(x_{iT}) =& -c + \alpha f'(0)\left(E(\theta_T | x_{iT} - \nu_T)\right) \geq \\ \geq& -c + \alpha f'(0)\left(\frac{\sigma_\epsilon^2 \mu_T + \sigma_\theta^2\left(\frac{2c}{\alpha f'(0)} + \nu_T\right)}{\sigma_\theta^2 + \sigma_\epsilon^2} - \nu_T\right) = \\ =& -c + \frac{\sigma_\epsilon^2}{\sigma_\theta^2 + \sigma_\epsilon^2}\alpha f'(0)(\mu_T - \nu_T) + 2c\frac{\sigma_\theta^2}{\sigma_\theta^2 + \sigma_\epsilon^2}.\end{aligned}$$

This is positive whenever $\sigma_\epsilon^2 \alpha f'(0)(\mu_T - \nu_T) + 2c\sigma_\theta^2 > c(\sigma_\theta^2 + \sigma_\epsilon^2)$, or equivalently $\sigma_\epsilon^2(c - \alpha f'(0)(\mu_T - \nu_T)) < \sigma_\theta^2 c$ or $\sigma_\epsilon^2\left(\frac{c}{\alpha f'(0)} - \mu_T + \nu_T\right) < \sigma_\theta^2 \frac{c}{\alpha f'(0)}$. □

**(ii) Best response to symmetric threshold strategy is threshold strategy.** Because the extremal equilibria can be obtained by infinitely iterating the agents' best-response functions (starting with a strategy profile in which everyone always attacks, or no one ever does, both of which are symmetric and in threshold strategies), they will necessarily be symmetric threshold strategy profiles if we can show that the best response to a symmetric threshold strategy profile is another symmetric threshold strategy profile. In other words, we want to show that if all agents $j \neq i$ attack iff $x_{jT} \geq x^*$, then $i$'s incentive to attack is strictly increasing in $x_{iT}$.

More formally, let $\Delta_{iT}(x, x', \sigma)$ be the marginal payoff from attacking for agent $i$ when she observes $x_{iT} = x$; all other agents $j$ attack iff $x_{jT} \geq x'$; and $\sigma_\epsilon = \sigma$. Then we want to show the following:

**Lemma 6.** $\Delta_{iT}(x, x', \sigma)$ *is strictly increasing in* $x$ *for all* $x, x' \geq \frac{c}{2\alpha f'(1)} + \nu_T$ *and* $\sigma^2 \in (0, \overline{\sigma}_{\epsilon 3}^2)$.

*Proof.* Recall that

$$\Delta_{iT}(x, x', \sigma) = -c + \alpha \int_{-\infty}^{\infty} (\theta - \nu_T) f'\left(l_T(\theta)\right) g(\theta|x) d\theta.$$

Note that $(\theta - \nu_T) f'\left(l_T(\theta)\right)$ is negative for $\theta < \nu_T$, and positive and strictly increasing in $\theta$ for $\theta > \nu_T$ (because both $\theta - \nu_T$ and $l_T(\theta) = \Phi\left(\frac{\theta - x'}{\sigma_\epsilon}\right)$ are strictly increasing in $\theta$, and $f'$ is strictly increasing). It follows that $\Delta_{iT}(x, x', \sigma)$ is increasing in $x$ if (a) $g(\theta|x)$ is FOSD-increasing in $x$ as a function of $\theta$, and (b) for each $\theta_0 < \nu_T$, $g(\theta_0|x)$ is decreasing in $x$ for all $x \geq \frac{c}{2\alpha f'(1)} + \nu_T$. (a) follows from Remark 2. (b) follows from the facts that $\phi(z)$ is increasing in $z$ for $z < 0$, and $g(\theta_0|x) = \frac{\sqrt{\sigma_\theta^2 + \sigma_\epsilon^2}}{\sigma_\theta \sigma_\epsilon} \phi\left(\frac{\theta_0 - \frac{\sigma_\epsilon^2 \mu_T + \sigma_\theta^2 x}{\sigma_\theta^2 + \sigma_\epsilon^2}}{\frac{\sigma_\theta \sigma_\epsilon}{\sqrt{\sigma_\theta^2 + \sigma_\epsilon^2}}}\right)$, where $\frac{\sigma_\epsilon^2 \mu_T + \sigma_\theta^2 x}{\sigma_\theta^2 + \sigma_\epsilon^2} \geq \nu_T + \frac{c}{4\alpha f'(1)} \geq \nu_T > \theta_0$ by Lemma 4.

□

Moreover, Lemmas 2 and 5 imply that any such $x$ must be bounded between $\frac{c}{2\alpha f'(1)} + \nu_T$ and $\frac{2c}{\alpha f'(0)} + \nu_T$.

**(iii) Unique equilibrium in threshold strategies.** Finally, we show that there is a unique symmetric threshold strategy equilibrium, which implies that the greatest and smallest equilibria coincide, and hence that there are no other equilibria (Milgrom and Roberts, 1990). Formally, what we will show is that, for $\sigma_\epsilon$ small enough, $\Delta_{iT}(x, x, \sigma_\epsilon)$ is continuous and strictly increasing in $x$, so there must be a unique $x^*(\sigma_\epsilon)$ for which $\Delta_{iT}(x, x, \sigma_\epsilon) = 0$, as required.

Dropping the index $iT$ to economize on notation, we can write

$$\Delta(x, x, \sigma_\epsilon) = -c + \alpha \int_{-\infty}^{\infty} (\theta - \nu) f'\left(\Phi\left(\frac{\theta - x}{\sigma_\epsilon}\right)\right) \frac{\sqrt{\sigma_\theta^2 + \sigma_\epsilon^2}}{\sigma_\theta \sigma_\epsilon} \phi\left(\frac{\theta - \frac{\sigma_\epsilon^2 \mu + \sigma_\theta^2 x}{\sigma_\theta^2 + \sigma_\epsilon^2}}{\frac{\sigma_\theta \sigma_\epsilon}{\sqrt{\sigma_\theta^2 + \sigma_\epsilon^2}}}\right) d\theta.$$

Applying the change of variable $z = \Phi\left(\frac{\theta - x}{\sigma_\epsilon}\right)$, so $dz = \phi\left(\frac{\theta - x}{\sigma_\epsilon}\right) \frac{1}{\sigma_\epsilon} d\theta$, and denoting

$\psi = \Phi^{-1}$, so $\frac{\theta - x}{\sigma_\epsilon} = \psi(z)$ and $\theta = x + \sigma_\epsilon \psi(z)$, we can rewrite the integral as:

$$\Delta(x, x, \sigma_\epsilon) = -c + \alpha \int_0^1 (x - \nu + \sigma_\epsilon \psi(z)) f'(z) \frac{\sqrt{\sigma_\theta^2 + \sigma_\epsilon^2}}{\sigma_\theta} \frac{\phi\left( \frac{\theta - \frac{\sigma_\epsilon^2 \mu + \sigma_\theta^2 x}{\sigma_\theta^2 + \sigma_\epsilon^2}}{\frac{\sigma_\theta \sigma_\epsilon}{\sqrt{\sigma_\theta^2 + \sigma_\epsilon^2}}} \right)}{\phi(\psi(z))} dz$$

$$\Delta(x, x, \sigma_\epsilon) = -c + \alpha \int_0^1 (x - \nu + \sigma_\epsilon \psi(z)) f'(z) \frac{\sqrt{\sigma_\theta^2 + \sigma_\epsilon^2}}{\sigma_\theta} \frac{\phi\left( \frac{(x-\mu)\sigma_\epsilon}{\sigma_\theta \sqrt{\sigma_\theta^2 + \sigma_\epsilon^2}} + \psi(z) \frac{\sqrt{\sigma_\epsilon^2 + \sigma_\theta^2}}{\sigma_\theta} \right)}{\phi(\psi(z))} dz.$$

Now note that the expression $(x - \nu + \sigma_\epsilon \psi(z)) \frac{\sqrt{\sigma_\theta^2 + \sigma_\epsilon^2}}{\sigma_\theta} \frac{\phi\left( \frac{(x-\mu)\sigma_\epsilon}{\sigma_\theta \sqrt{\sigma_\theta^2 + \sigma_\epsilon^2}} + \psi(z) \frac{\sqrt{\sigma_\epsilon^2 + \sigma_\theta^2}}{\sigma_\theta} \right)}{\phi(\psi(z))}$ defines a function $h$ of its arguments $x$, $z$, $\mu$, $\nu$, $\sigma_\epsilon$, $\sigma_\theta$ that is well defined and $C^\infty$ over all $x, \mu, \nu \in \mathbb{R}$, $z \in (0, 1)$, $\sigma_\theta > 0$ and, importantly, all $\sigma_\epsilon \in \mathbb{R}$, *including zero* (and negative values). Moreover, we can show that the integrand $h(\cdot) f'(z)$ is uniformly bounded by an integrable function for all $\sigma_\epsilon$ below a threshold. Indeed, $f'$ is bounded. Using that $\Phi(y) \leq e^y$ for $y < 0$, we obtain $z \leq e^{\psi(z)}$, or $\psi(z) \geq \ln(z) \implies |\psi(z)| \leq |\ln(z)|$ for $z < 0.5$. Using that $\Phi(y) \leq \frac{\phi(y)}{|y|}$ for $y < 0$, we obtain $z|\psi(z)| \leq \phi(\psi(z)) = \frac{1}{\sqrt{2\pi}} e^{\frac{-1}{2}\psi(z)^2}$ for $z < 0.5$, so

$$\begin{aligned}
\frac{\phi\left( \frac{(x-\mu)\sigma_\epsilon}{\sigma_\theta \sqrt{\sigma_\theta^2 + \sigma_\epsilon^2}} + \psi(z) \frac{\sqrt{\sigma_\epsilon^2 + \sigma_\theta^2}}{\sigma_\theta} \right)}{\phi(\psi(z))} &= e^{-\frac{1}{2}\left[ \left( \frac{(x-\mu)\sigma_\epsilon}{\sigma_\theta \sqrt{\sigma_\theta^2 + \sigma_\epsilon^2}} + \psi(z) \frac{\sqrt{\sigma_\epsilon^2 + \sigma_\theta^2}}{\sigma_\theta} \right)^2 - \psi(z)^2 \right]} \\
&\leq e^{\frac{1}{2}\left[ \left( \frac{(x-\mu)}{\sigma_\theta^2} \right)^2 \sigma_\epsilon^2 + 2 \frac{(x-\mu)}{\sigma_\theta^2} |\psi(z)| \frac{\sigma_\epsilon + \sigma_\theta}{\sigma_\theta} \sigma_\epsilon + \psi(z)^2 \left( \frac{\sigma_\epsilon^2}{\sigma_\theta^2} + 2 \frac{\sigma_\epsilon}{\sigma_\theta} \right) \right]} \\
&\leq e^{A\sigma_\epsilon^2 + |\psi(z)|(B\sigma_\epsilon^2 + C\sigma_\epsilon) + \psi(z)^2(D\sigma_\epsilon^2 + \sigma_\epsilon)}
\end{aligned}$$

for some $A$, $B$, $C$, $D$, $E > 0$ independent of $z$ and $\sigma_\epsilon^2$. For $z$ low enough that $|\psi(z)| > 1$, this expression is bounded above by

$$\begin{aligned}
e^{\psi(z)^2((A+B+D)\sigma_\epsilon^2 + (C+E)\sigma_\epsilon} &\leq \left( \frac{1}{\sqrt{2\pi} z |\psi(z)|} \right)^{2\left[(A+B+D)\sigma_\epsilon^2 + (C+E)\sigma_\epsilon\right]} \\
&\leq \left( \frac{1}{z} \right)^{2\left[(A+B+D)\sigma_\epsilon^2 + (C+E)\sigma_\epsilon\right]}.
\end{aligned}$$

Hence the left end of the integral is of the form $\frac{|\ln(z)|}{z^\beta}$, and is well behaved for any

$\sigma_\epsilon$ such that the exponent $\beta$ is less than 1, e.g., $\sigma_\epsilon < \frac{1}{2(A+B+C+D+E)}$. An analogous bound can be given for $z$ close to 1. It follows by the dominated convergence theorem that $\Delta$ is a continuous function of its arguments, in particular at $\sigma_\epsilon = 0$, where

$$\Delta(x, x, 0) = -c + \alpha \int_0^1 (x - \nu) f'(z) dz = -c + \alpha(f(1) - f(0))(x - \nu).$$

But we need to go a step further. To prove that $\Delta$ is strictly increasing in $x$ for $\sigma_\epsilon$ small, we will show that $\frac{\partial \Delta}{\partial x}(x, x, \sigma_\epsilon)$ converges uniformly to $\frac{\partial \Delta}{\partial x}(x, x, 0) \equiv \alpha(f(1) - f(0)) > 0$ as $\sigma_\epsilon \to 0$.

We can use a similar argument. Denoting $\frac{(x-\mu)\sigma_\epsilon}{\sigma_\theta \sqrt{\sigma_\theta^2 + \sigma_\epsilon^2}} + \psi(z) \frac{\sqrt{\sigma_\epsilon^2 + \sigma_\theta^2}}{\sigma_\theta} = w$, and using that $\phi'(x) = -x\phi(x)$, note that

$$\frac{\partial f'(z) h(\cdot)}{\partial x} = f'(z) \frac{\sqrt{\sigma_\theta^2 + \sigma_\epsilon^2}}{\sigma_\theta} \frac{\phi(w)}{\phi(\psi(z))} + (x - \nu + \sigma_\epsilon \psi(z)) f'(z) \frac{\sigma_\epsilon}{\sigma_\theta^2} \frac{\phi(w)}{\phi(\psi(z))} w.$$

Using the same bounds as before, the first term is bounded by an expression of the form $\frac{1}{z^\beta}$ for $z$ close to zero, while the second is bounded by an expression of the form $\frac{\ln(z)^2}{z^\beta}$ for $z$ close to zero, where $\beta < 1$ if $\sigma_\epsilon$ is small. Hence this expression is bounded (uniformly for $x$, $\mu$, $\nu$, $\sigma_\theta$, and $\sigma_\epsilon$ in any closed intervals, with $\sigma_\theta$ strictly positive) by an integrable function. The Leibniz integral rule then implies that $\frac{\partial \Delta}{\partial x}(x, x, \sigma_\epsilon) \equiv \alpha \int_0^1 \frac{\partial f'(z) h(\cdot)}{\partial x} dz$. Moreover, for any convergent sequence $Y_k = (x_k, \mu_k, \nu_k, \sigma_{\theta k}, \sigma_{\epsilon k})$ with limit $Y_\infty$, we have that $\frac{\partial \Delta}{\partial x}(Y_k) \xrightarrow[k \to \infty]{} \frac{\partial \Delta}{\partial x}(Y_\infty)$ by the dominated convergence theorem, since the integrand $f'(z)h(\cdot)$ is obviously continuous in the argument $Y$ and so converges pointwise. But then $\frac{\partial \Delta}{\partial x}(Y)$ is a continuous function of $Y$. Within any compact set, then, it must be uniformly continuous by the Heine-Cantor theorem. In particular, we can take a rectangle where $\sigma_\epsilon \in [0, 1]$ and the other variables lie in any closed interval (with $\min \sigma_\theta^2 > 0$). Then, by the uniform continuity, there is $\overline{\sigma}_\epsilon$ such that, if $\sigma_\epsilon \in (0, \overline{\sigma}_\epsilon)$ and the other variables lie in their respective intervals, $\frac{\partial \Delta}{\partial x}(x, \mu, \nu, \sigma_\theta, \sigma_\epsilon) - \frac{\partial \Delta}{\partial x}(x, \mu, \nu, \sigma_\theta, 0) < \frac{\alpha}{2}(f(1) - f(0))$, whence $\frac{\partial \Delta}{\partial x}(x, \mu, \nu, \sigma_\theta, \sigma_\epsilon) > 0$. In particular, taking the range of $x$ to contain $\left[\underline{\nu} + \frac{c}{2\alpha f'(1)}, \overline{\nu} + \frac{2c}{\alpha f'(0)}\right]$, this argument guarantees that there is $\overline{\sigma}_\epsilon$ such that, for all $\sigma_\epsilon \in (0, \overline{\sigma}_\epsilon)$, $\frac{\partial \Delta}{\partial x}$ is strictly increasing at every $x$ between the dominance regions, which yields the uniqueness.

**(iv) Equilibrium threshold as $\sigma_\epsilon \to 0$.** Our previous argument implies that, as $\sigma_\epsilon \to 0$, $x^*(\sigma_\epsilon) \to \frac{c}{\alpha[f(1) - f(0)]} + \nu$; indeed, if not, there would be $\eta_0 > 0$ and a

sequence $\sigma_k \to 0$ such that either $x^*(\sigma_k) \geq \frac{c}{\alpha[f(1)-f(0)]} + \nu + \eta_0$ for all $k$ or $x^*(\sigma_k) \leq \frac{c}{\alpha[f(1)-f(0)]} + \nu - \eta_0$ for all $k$. But our formula for $\Delta(x, x, 0)$ and the continuity of $\Delta$ would imply that, for $k$ high enough, $\Delta(x, x, \sigma_k) > 0$ at any $x \geq \frac{c}{\alpha[f(1)-f(0)]} + \nu + \eta_0$, and $\Delta(x, x, \sigma_k) < 0$ at any $x \leq \frac{c}{\alpha[f(1)-f(0)]} + \nu - \eta_0$, a contradiction.

□

*Proof of Proposition 1.* The marginal payoff from attacking in period $t$ is given by the expression

$$\Delta_{it} = -c + E\left[\alpha(\theta_t - \nu_t - \delta\overline{U}_{t+1})f'(l_t)|x_{it}\right].$$

By the same argument as in Lemma 1, for $\sigma_\epsilon$ small enough, this game has a unique equilibrium, which is symmetric and in threshold strategies. In fact, this game is equivalent to the game from period $T$, if we denote $\nu_t + \delta\overline{U}_{t+1} \equiv \nu_T$. Note that the proof of Lemma 1 yields the uniqueness result in this Proposition only because we showed that a threshold $\overline{\sigma}_\epsilon$ can be found below which uniqueness is guaranteed, *regardless of the value* that other parameters (in particular, $\nu$) take, as long as they lie in a compact interval. Indeed, in general the equilibrium in periods $t+1$ and onwards depends on the value of $\sigma_\epsilon$; hence the continuation value $\delta\overline{U}_{t+1}$ is a function of $\sigma_\epsilon$. Thus, for periods $t < T$, we need to show that there is a threshold $\overline{\sigma}_\epsilon$ such that, for all $\sigma_\epsilon < \overline{\sigma}_\epsilon$, the game with (endogenous) status quo payoff $\nu = \nu_t + \delta\overline{U}_{t+1}(\sigma_\epsilon)$ has a unique equilibrium. Our proof from Lemma 1 guarantees that we can find a threshold $\overline{\sigma}_\epsilon$ that works whenever $\nu$ lies, for instance, in $[\nu_t + \delta\underline{u}, \nu_t + \delta\overline{u}]$, where $\underline{u}$, $\overline{u}$ are the infimum and supremum of the game's possible continuation payoffs across all feasible strategy profiles. This interval is guaranteed to contain $\nu_t + \delta\overline{U}_{t+1}(\sigma_\epsilon)$.

Because of the continuity of $\Delta$ (in particular with respect to both $\sigma_\epsilon$ and $\nu$), our proof of Lemma 1 also implies that, as $\sigma_\epsilon \to 0$, $x_t^*(\sigma_\epsilon, \sigma_\theta) \to x_t^*(\sigma_\theta)$, where

$$x_t^*(\sigma_\theta) = \frac{c}{\alpha[f(1) - f(0)]} + \nu_t + \delta\overline{U}_{t+1}(\sigma_\theta),$$

where $\overline{U}_{t+1}(\sigma_\theta) = \lim_{\sigma_\epsilon \to 0} \overline{U}_{t+1}(\sigma_\epsilon, \sigma_\theta)$. The convergence of $\overline{U}_{t+1}(\sigma_\epsilon, \sigma_\theta)$ and $x_t^*(\sigma_\epsilon, \sigma_\theta)$ can be shown by backward induction from $T$, using that if $x_{t+1}^*$ converges, then $\overline{U}_{t+1}$ converges, and so $x_t^*$ does as well.

As for Equation (4), for general values of $\sigma_\epsilon$ and $\sigma_\theta$, let $U_t(x, \sigma_\epsilon, \sigma_\theta)$ be the expected continuation hedonic utility in equilibrium of an agent $i$ starting at time $t$, conditional

on seeing $x_{it} = x$, and $\overline{U}_t(\sigma_\epsilon, \sigma_\theta)$ be $i$'s expected continuation hedonic utility before seeing $x_{it}$ (both of which, by symmetry, are the same for all agents). Then we have

$$U_t(x, \sigma_\epsilon, \sigma_\theta) = -c\mathbb{1}_{\{x \geq x_t^*(\sigma_\epsilon, \sigma_\theta)\}} + E\left[\left(\theta_t - \nu_t - \delta\overline{U}_{t+1}(\sigma_\epsilon, \sigma_\theta)\right) f(l_t(\theta_t))|x\right] + \nu_t + \delta\overline{U}_{t+1}(\sigma_\epsilon, \sigma_\theta)$$

$$\begin{aligned}\overline{U}_t(\sigma_\epsilon, \sigma_\theta) = &-c\Phi\left(\frac{\mu_t - x_t^*(\sigma_\epsilon, \sigma_\theta)}{\sqrt{\sigma_\epsilon^2 + \sigma_\theta^2}}\right) + \\ &E\left[\left(\theta_t - \nu_t - \delta\overline{U}_{t+1}(\sigma_\epsilon, \sigma_\theta)\right) f(l_t(\theta_t))\right] + \nu_t + \delta\overline{U}_{t+1}(\sigma_\epsilon, \sigma_\theta)\end{aligned}$$

$$\begin{aligned}\overline{U}_t(\sigma_\epsilon, \sigma_\theta) = &-c\Phi\left(\frac{\mu_t - x_t^*(\sigma_\epsilon, \sigma_\theta)}{\sqrt{\sigma_\epsilon^2 + \sigma_\theta^2}}\right) + \\ &E\left[\left(\theta_t - \nu_t - \delta\overline{U}_{t+1}(\sigma_\epsilon, \sigma_\theta)\right) f\left(\Phi\left(\frac{\theta_t - x_t^*(\sigma_\epsilon, \sigma_\theta)}{\sigma_\epsilon}\right)\right)\right] + \nu_t + \delta\overline{U}_{t+1}(\sigma_\epsilon, \sigma_\theta).\end{aligned}$$

As $\sigma_\epsilon \to 0$, $\overline{U}_t(\sigma_\epsilon, \sigma_\theta)$ converges to

$$\overline{U}_t(\sigma_\theta) = -c\Phi\left(\frac{\mu_t - x_t^*(\sigma_\theta)}{\sigma_\theta}\right) + E\left[\left(\theta_t - \nu_t - \delta\overline{U}_{t+1}(\sigma_\theta)\right) f\left(\mathbb{1}_{\{\theta_t > x_t^*(\sigma_\theta)\}}\right)\right] + \nu_t + \delta\overline{U}_{t+1}(\sigma_\theta).$$

As $\sigma_\theta \to 0$, $\overline{U}_t(\sigma_\theta)$ converges to

$$\overline{U}_t = -c\mathbb{1}_{\{\mu_t > x_t^*\}} + \left(\mu_t - \nu_t - \delta\overline{U}_{t+1}\right) f\left(\mathbb{1}_{\{\mu_t > x_t^*\}}\right) + \nu_t + \delta\overline{U}_{t+1},$$

and $x_t^*(\sigma_\theta)$ converges to

$$x_t^* = \frac{c}{\alpha[f(1) - f(0)]} + \nu_t + \delta\overline{U}_{t+1},$$

as we wanted. □

*Proof of Proposition 2.* For part (i), assume that $\mu_t = \mu$ for all $t$, with $\mu < \mu_0$. Then, using Equation (3), we can calculate

$$x_T^* = \frac{c}{\alpha[f(1) - f(0)]} + \frac{\nu}{1 - \delta}.$$

Since $\mu < \mu_0$, as $\sigma_\theta$ goes to zero, for $\sigma_\epsilon(\sigma_\theta)$ small enough, we are in the limit equilibrium characterized in Proposition 1 in the case $\mu_t < x_t^*$, in which $\theta_t < x_t^*$ with

probability going to one, and $l_t$ converges in probability to zero. Hence

$$\overline{U}_T = f(0)\mu + (1 - f(0))\frac{\nu}{1-\delta}.$$

We can then calculate

$$x^*_{T-1} = \frac{c}{\alpha[f(1) - f(0)]} + \nu + \delta f(0)\mu + \delta(1 - f(0))\frac{\nu}{1-\delta}.$$

There are now two cases. If $\mu \in \left(\frac{\nu}{1-\delta}, \mu_0\right)$, then automatically $x^*_{T-1} > x^*_T > \mu$, so that almost no one attacks in period $T-1$ either. By backward induction, we obtain that

$$\overline{U}_t = f(0)\frac{1 - \delta^{T-t+1}(1 - f(0))^{T-t+1}}{1 - \delta(1 - f(0))}\mu + \left[1 - f(0)\frac{1 - \delta^{T-t+1}(1 - f(0))^{T-t+1}}{1 - \delta(1 - f(0))}\right]\frac{\nu}{1-\delta}$$
$$x^*_{t-1} = \frac{c}{\alpha[f(1) - f(0)]} + \nu + \delta\overline{U}_t,$$

whence $\overline{U}_t > \overline{U}_{t+1}$ and $x^*_t > x^*_{t+1} > \ldots > \mu$ for all $t$, and almost no one ever attacks in equilibrium. On the other hand, if $\mu \leq \frac{\nu}{1-\delta}$, then $\overline{U}_t$ and $x^*_{t-1}$ obey the same equations, but now $x^*_t > \mu$ instead follows from the fact that $x^*_t > \nu + \delta\overline{U}_{t+1}$ which is a convex combination of $\mu$ and $\frac{\nu}{1-\delta}$, hence higher than $\mu$.

For part (ii), suppose that $\mu_t = \mu > \mu^*$ for all $t$. Then, from Equation (4), we know that, if $x^*_t < \mu$ for all $t \geq t_0$, then for all $t$ between $t_0$ and $T-1$,

$$\overline{U}_t = -c + f(1)\mu + (1 - f(1))(\nu + \delta\overline{U}_{t+1}),$$

with $\overline{U}_T = -c + f(1)\mu + (1 - f(1))\frac{\nu}{1-\delta}$. Equivalently, for $t \geq t_0$,

$$\overline{U}_t = \frac{1 - \delta^{T-t+1}(1 - f(1))^{T-t+1}}{1 - \delta(1 - f(1))}(-c + f(1)\mu) + \left[1 - f(1)\frac{1 - \delta^{T-t+1}(1 - f(1))^{T-t+1}}{1 - \delta(1 - f(1))}\right]\frac{\nu}{1-\delta}.$$

This is a convex combination of $\mu - \frac{c}{f(1)}$ and $\frac{\nu}{1-\delta}$, with the weight on the first term decreasing in $t$. Since

$$\mu^* \geq \frac{c}{f(1)} + \frac{\nu}{1-\delta},$$

with equality iff $\alpha = 1$ and $f(0) = 0$, and $\mu > \mu^*$, we know that $\mu - \frac{c}{f(1)} > \frac{\nu}{1-\delta}$,

so $\overline{U}_{t_0} > \ldots > \overline{U}_T > \frac{\nu}{1-\delta}$ and $x^*_{t_0-1} > \ldots > x^*_T$. For most players to attack in equilibrium at time $t_0 - 1$, we need $x^*_{t_0-1} < \mu$.

Iterating, to prove the result we need to show that $x^*_t < \mu$ for all $t$ with the thresholds calculated as above, i.e., under the assumption that all agents will attack in future periods. Because the sequence is decreasing in $t$, it is enough to show that $\mu > \lim_{t\to-\infty} x^*_t$, i.e.,

$$\begin{aligned}
\mu &> \frac{c}{\alpha[f(1)-f(0)]}+\nu+\delta\frac{-c+f(1)\mu}{1-\delta(1-f(1))}+\delta\frac{(1-\delta)(1-f(1))}{1-\delta+\delta f(1)}\frac{\nu}{1-\delta} \\
\iff \frac{1-\delta}{1-\delta+\delta f(1)}\mu &> \frac{c}{\alpha[f(1)-f(0)]}-\frac{\delta c}{1-\delta+\delta f(1)}+\frac{\nu}{1-\delta+\delta f(1)} \\
\iff \mu &> \frac{c}{\alpha[f(1)-f(0)]}\left(1+\frac{\delta f(1)}{1-\delta}\right)-\frac{\delta c}{1-\delta}+\frac{\nu}{1-\delta}=\mu^*.
\end{aligned}$$

Finally, for part (iii), it is convenient to relabel time periods as follows: set $T = 0$ and assume the game is played beginning at any integer $t < 0$. Let $(x^*_t)_{t\in\mathbb{Z}_{\leq 0}}$ be the sequence of equilibrium attack thresholds for this game, as characterized in Proposition 1, for $\sigma_\theta \to 0$ with $\sigma_\epsilon$ small enough. We will show that, generically, there are infinitely many values of $t$ for which $x^*_t > \mu_t$ and infinitely many for which $x^*_t < \mu_t$. (We will discard the non-generic case in which $\mu_t = x^*_t$ for any $t$. Note that, given values of $\mu_{t+1}, \ldots, \mu_0$, and the other parameters satisfying this constraint, the value of $\overline{U}_{t+1}$ is uniquely pinned down, and hence so is $x^*_t$, by Equation (3), so there is a single real value of $\mu_t$ that is being ruled out.)

Suppose the former statement is not true, so that $x^*_t \leq \mu_t$ for all $t \leq t_0$ for some $t_0$. By our genericity assumption, we must then have $x^*_t < \mu_t$ for all $t \leq t_0$, and

$$\overline{U}_t = -c + f(1)\mu_t + (1-f(1))(\nu + \delta\overline{U}_{t+1}) \tag{10}$$

for all $t \leq t_0$. Let $\underline{\mu} = \liminf_{t\to-\infty}\mu_t$. Let $(t_s)_{s\in\mathbb{N}}$ be a subsequence such that $\underline{\mu} = \lim_{s\to\infty}\mu_{t_s}$. Then, taking the limit of the inequality $x^*_{t_s} < \mu_{t_s}$ as $s \to \infty$, we must have $x^* \leq \underline{\mu}$ for any $x^*$ that the $x^*_{t_s}$ accumulate to. In particular, $\liminf x^*_t \leq \underline{\mu}$, or equivalently

$$\frac{c}{\alpha[f(1)-f(0)]} + \nu + \delta\liminf\overline{U}_t \leq \underline{\mu}.$$

Equation (10) implies that $\overline{U}_t$, and $\overline{U}_{t'}$ for all $t' < t$, are increasing functions of $\mu_t$. Hence $\liminf\overline{U}_t$ is bounded below by a hypothetical $\tilde{U}$ calculated under the

assumptions that everyone always attacks and that $\mu_t = \underline{\mu}$ for all $t$, i.e.,

$$\liminf \overline{U}_t \geq \tilde{U} = \frac{-c + f(1)\underline{\mu}}{1 - \delta + \delta f(1)} + \frac{(1 - f(1))\nu}{1 - \delta + \delta f(1)},$$

calculating $\tilde{U}$ as in part (ii).

Then it must be that

$$\frac{c}{\alpha[f(1) - f(0)]} + \nu + \delta\frac{-c + f(1)\underline{\mu}}{1 - \delta + \delta f(1)} + \delta\frac{(1 - f(1))\nu}{1 - \delta + \delta f(1)} \leq \underline{\mu}$$
$$\iff \mu^* \leq \underline{\mu}.$$

Indeed, by construction, $\mu^*$ is the threshold value of $\underline{\mu}$ which would make this inequality hold with equality. But, since $\mu_t \leq \mu^* - \eta$ for all $t$, $\underline{\mu} \leq \mu^* - \eta < \mu^*$, a contradiction.

The proof for the latter part of the claim is similar. Suppose that $x_t^* \geq \mu_t$ for all $t$ below some $t_0$. By our genericity assumption, we must have $x_t^* > \mu_t$ for all $t \leq t_0$, so

$$\overline{U}_t = f(0)\mu_t + (1 - f(0))(\nu + \delta\overline{U}_{t+1}) \tag{11}$$

for all $t \leq t_0$. Letting $\overline{\mu} = \limsup_{t \to -\infty} \mu_t$, we must have $\limsup x_t^* \geq \overline{\mu}$, or equivalently

$$\frac{c}{\alpha[f(1) - f(0)]} + \nu + \delta \limsup \overline{U}_t \geq \overline{\mu}.$$

In turn $\overline{U}_t$ is bounded above by a hypothetical $\hat{U}$ calculated under the assumption that no one attacks in the future and $\mu_t = \overline{\mu}$ for all $t$, i.e.,

$$\limsup \overline{U}_t \leq \hat{U} = \frac{f(0)\overline{\mu}}{1 - \delta + \delta f(0)} + \frac{(1 - f(0))\nu}{1 - \delta + \delta f(0)}.$$

Then we must have

$$\frac{c}{\alpha[f(1) - f(0)]} + \nu + \delta\frac{f(0)\overline{\mu}}{1 - \delta + \delta f(0)} + \delta\frac{(1 - f(0))\nu}{1 - \delta + \delta f(0)} \geq \overline{\mu}$$
$$\iff \mu_* \geq \overline{\mu}.$$

But by assumption $\overline{\mu} \geq \mu_* + \eta > \mu_*$, a contradiction. □

*Proof of Proposition 3.* By Equation (4), $\frac{\partial x_t^*}{\partial \nu_t} = 1$. For $t' > t$, assuming a marginal change that does not change the equilibrium actions, $x_t^*$ only depends on $\nu_{t'}$ through $\overline{U}_{t+1}$, which only depends on $\nu_{t'}$ through $\overline{U}_{t+2}, \ldots$, which only depends on $\nu_{t'}$ through $\overline{U}_{t'}$. So

$$\frac{\partial x_t^*}{\partial \nu_{t'}} = \delta \frac{\partial \overline{U}_{t+1}}{\partial \nu_{t'}} = \delta \prod_{s=1}^{t'-t-1} \frac{\partial \overline{U}_{t+s}}{\partial \overline{U}_{t+s+1}} \frac{\partial \overline{U}_{t'}}{\partial \nu_{t'}} = \delta^{t'-t} \prod_{s=1}^{t'-t} \left(1 - f(\mathbb{1}_{\mu_{t+s} > x_{t+s}^*})\right) \geq 0,$$

with equality only if $f(1) = 1$ and $\mu_{t+s} > x_{t+s}^*$ for some $s$ between 1 and $t' - t$. As for changes in $\mu_t$, by Equation (4), $\frac{\partial x_t^*}{\partial \mu_t} = 0$. However, $\frac{\partial \overline{U}_t}{\partial \mu_t} = f(\mathbb{1}_{\mu_{t+s} > x_{t+s}^*})$. Hence, for $t' > t$,

$$\frac{\partial x_t^*}{\partial \mu_{t'}} = \delta \prod_{s=1}^{t'-t-1} \frac{\partial \overline{U}_{t+s}}{\partial \overline{U}_{t+s+1}} \frac{\partial \overline{U}_{t'}}{\partial \mu_{t'}} = \delta^{t'-t} \prod_{s=1}^{t'-t-1} \left(1 - f(\mathbb{1}_{\mu_{t+s} > x_{t+s}^*})\right) f(\mathbb{1}_{\mu_{t'} > x_{t'}^*}) \geq 0,$$

with equality only if $f(1) = 1$ and $\mu_{t+s} > x_{t+s}^*$ for some $s$ between 1 and $t' - t - 1$, or $f(0) = 0$ and $\mu_{t'} < x_{t'}^*$. □

*Proof of Remark 1.* The social planner aims to maximize the sum (or integral) of the citizens' (expected) ex ante utilities, $\int_0^1 \sum_{t=0}^{\infty} \delta_t u_{it} di$. The threshold in Equation (5) then follows from the following calculation: in period $t$, given that the planner would have the citizens play optimally from period $t + 1$ onwards (thus generating payoff $\overline{U}_{t+1}^{\text{sp}}$), she can generate an aggregate payoff of

$$f(0)\theta_t + (1 - f(0))(\nu_t + \delta \overline{U}_{t+1}^{\text{sp}})$$

by having nobody protest in period $t$, or an aggregate payoff of

$$-c + f(1)\theta_t + (1 - f(1))(\nu_t + \delta \overline{U}_{t+1}^{\text{sp}})$$

by having everybody protest. The latter expression dominates the former precisely when $\theta_t \geq x_t^{\text{sp}}$. Having a fraction of protesters protest is strictly worse than at least one of these two options, because $f$ is strictly convex.

(i) follows from the fact that the planner's payoff from any fixed strategy profile weakly increases as $\mu_t$ or $\nu_t$ increases; since the planner has full control over the players' strategies, her optimal payoff must increase by at least as much as if strategies

are held fixed (if anything, re-optimizing given the new parameters might yield further gains).

For (ii), note that, if $\mu_t \equiv \mu$; $\nu_t \equiv \nu$ for $t < T$, $\nu_T = \frac{\nu}{1-\delta}$; $f(0) = 0$; and $\sigma_\theta = 0$, then, if $\mu < \frac{c}{[f(1)-f(0)]} + \frac{\nu}{1-\delta}$, then there is no protest in period $T$ by Equation (5). Then, since $f(0) = 0$, $\overline{U}_T = \frac{\nu}{1-\delta}$, so $x^{\text{sp}}_{T-1} = \frac{c}{[f(1)-f(0)]} + \nu + \delta\frac{\nu}{1-\delta} = x^{\text{sp}}_T$. By backward induction, $x^{\text{sp}}_t$ is constant in $t$, and lower than $\mu$ for all $t$.

On the other hand, if $\mu > \frac{c}{[f(1)-f(0)]} + \frac{\nu}{1-\delta} = \frac{c}{f(1)} + \frac{\nu}{1-\delta}$, then everybody protests in period $T$, again by Equation (5). Then $\overline{U}_T = -c + f(1)\mu + (1 - f(1))\frac{\nu}{1-\delta}$. Plugging this into Equation (5), we obtain

$$\begin{aligned} x^{\text{sp}}_{t-1} &= \frac{c}{f(1)} + \nu + \delta\left(-c + f(1)\mu + (1 - f(1))\frac{\nu}{1-\delta}\right) \\ &= \left[\frac{c}{f(1)} + \frac{\nu}{1-\delta}\right](1 - \delta f(1)) + \delta f(1)\mu, \end{aligned}$$

which is less than $\mu$ whenever $\mu > \frac{c}{f(1)} + \frac{\nu}{1-\delta}$. By backward induction, we obtain that $\overline{U}_t$ is as calculated in part (ii) of Proposition 1, and is then a convex combination of $\mu - \frac{c}{f(1)}$ and $\frac{\nu}{1-\delta}$, so that $x^{\text{sp}}_t$ is between $\frac{c}{f(1)} + \frac{\nu}{1-\delta}$ and $\frac{c}{f(1)} + \nu + \delta\left(\mu - \frac{c}{f(1)}\right)$, and hence less than $\mu$, for all $t$.

For (iii), note that, whenever the social planner would have nobody attacking, the marginal incentive to attack for a citizen $i$ (if others are presumed to follow the social planner's strategy profile) is

$$\Delta_{it} = -c + \alpha f'(0)(\theta_t - \nu_t - \delta\overline{U}^{\text{sp}}_{t+1}),$$

which is always less than $\Delta^{\text{sp}}_t = -c + [f(1) - f(0)](\theta_t - \nu_t - \delta\overline{U}^{\text{sp}}_{t+1})$ because $\alpha \leq 1$ and $f$ is strictly convex (so $f'(0) < f(1) - f(0)$). Since the social planner has the citizens attack whenever $\Delta^{\text{sp}}_t > 0$, it must be that $\Delta^{\text{sp}}_t \leq 0$, which implies $\Delta_{it} < 0$.[34]

On the other hand, when the social planner has everybody attack, the marginal incentive to attack for a citizen $i$ (if, again, others follow the social planner's profile) is

$$\Delta_{it} = -c + \alpha f'(1)(\theta_t - \nu_t - \delta\overline{U}^{\text{sp}}_{t+1}),$$

which is at least as high as $\Delta^{\text{sp}}_t$ (hence at least zero) if $\alpha f'(1) \geq f(1) - f(0)$. □

[34]Note that, if $\theta_t - \nu_t - \delta\overline{U}^{\text{sp}}_{t+1} < 0$, the ordering between $\Delta_{it}$ and $\Delta^{\text{sp}}_t$ may flip, but in this case both the social planner and the citizen agree that there should be no attack anyway.

*Derivation of Equation* (7). By analogous arguments to those used in the proof of Proposition 1, $x_t^*(\sigma_\epsilon, \sigma_\theta)$ is the unique value of $x$ that solves the equation

$$\begin{aligned}
0 =& \Delta_{it} = -c + E\left[\alpha((1-\rho)\theta_t + l_t\rho\theta_t - \nu_t - \delta\overline{U}_{t+1})f'(l_t) + \rho\theta_t f(l_t)|x_{it} = x\right] \\
=& -c + \alpha(1-\rho)E(\theta_t f'(l_t)|x) - \alpha(\nu_t + \delta\overline{U}_{t+1})E(f'(l_t)|x) + \\
& + \alpha\rho E(l_t\theta_t f'(l_t)|x) + \rho E(f(l_t)\theta_t|x) \\
\xrightarrow[\sigma_\epsilon \to 0]{} & -c + \alpha(1-\rho)x[f(1) - f(0)] - \alpha(\nu_t + \delta\overline{U}_{t+1})[f(1) - f(0)] + \\
& + \alpha\rho E(\theta_t l_t f'(l_t)|x) + \rho E(\theta_t f(l_t)|x)
\end{aligned}$$

As shown in Proposition 1, as $\sigma_\epsilon \to 0$, $\theta_t|x$ converges to $x$, while $l_t|x$ is asymptotically uniformly distributed between 0 and 1. Then $E(\theta_t l_t f'(l_t)|x)$ converges to $x\int_0^1 l f'(l)dl = x(f(1) - \int_0^1 f(l)dl)$, and $E(\theta f(l_t)|x)$ converges to $x\int_0^1 f(l)dl$. Substituting these identities into the above and rearranging yields Equation (7). □

# B A Model of Fighting to Survive

This extension demonstrates the flexibility of our framework by considering a variant of the model with the following properties. Suppose now that, while the movement survives, the agents receive flow payoffs $\theta_t$ in *every* period. If the movement is crushed in period $t$, there are no more opportunities to demonstrate in the future, and agents receive a lump sum $\nu_t$ *once* and the game ends. (Of course, $\nu_t$ can represent a discounted sum of payoffs.) Demonstrating still costs $c$ and we make the same assumptions as before regarding altruism. The probability that the movement survives period $t$ is $f(l_t)$.

Then the net payoff of demonstrating for the marginal agent is

$$-c + E\left[\alpha(\theta_t + \delta\overline{U}_{t+1} - \nu_t)|x_{it} = x_t^*(\sigma_\epsilon)\right],$$

where $\overline{U}_{t+1}$ is the continuation payoff from arriving at $t+1$ with the movement still active. Hence, the limit equilibrium cutoff as $\sigma_\epsilon \to 0$ is now

$$x_t^* = \frac{c}{\alpha[f(1) - f(0)]} + \nu_t - \delta\overline{U}_{t+1}. \tag{12}$$

As in the main model, agents are reluctant to protest relative to the social planner's solution (because they do not fully internalize the benefits), which means that a marginal change in the future parameters which shifts the equilibrium from not attacking to attacking in a future period will discontinuously increase the players' payoffs. But, in this variant of the model, such an increase in continuation utilities will actually **encourage** more protests today, since the citizens are more likely to accrue that higher continuation utility precisely if they do protest today. (Mechanically, this appears in Equation (12) as a negative sign in front of the term $\delta\overline{U}_{t+1}$: an increasing continuation utility from survival lowers the threshold $x_t^*$ for protesting today.) More generally, expectations of future agitation reinforce, rather than discourage, incentives to fight today.

The logic leading to intermittent protests in the main model is then reversed, leading instead to bang-bang solutions. For example, then, if we assume $\nu_t \equiv 0$, instead of there being a range $[\mu_*, \mu^*]$ of protest payoffs leading to intermittent protests, there is a single threshold $\mu^* = \frac{c}{\alpha[f(1)-f(0)]}$ such that, if $\mu_t < \mu^*$ for all $t$, almost nobody protests in each period, while if $\mu_t > \mu^*$ for all $t$, most citizens protest in each period.

Chassang (2010) studies a closely related model, with two players who must both cooperate for the relationship (analogously, the protest) to survive, and a stationary environment (which does not allow free variation over time of $\mu_t$ or $\nu_t$) but with an infinite horizon. In the infinite-horizon case, the dynamic complementarity discussed in the previous two paragraphs is still present, but we can no longer backward induct from a last period to find a unique equilibrium. Within his model, Chassang provides an elegant characterization of (potentially multiple) infinite-horizon equilibria that are Markovian in a certain sense.