Collective Procrastination and Protest Cycles[*]

Germán Gieczewski[†] and Korhan Kocak[‡]

June 2024.

Abstract

This paper studies a model of "pivotal protesting," in which citizens act in order to change the outcome rather than to collect private benefits. We show that, when citizens face repeated opportunities to protest against a regime, pivotal protesting entails complex dynamic considerations: 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 often induces a pattern of protest cycles, driven by a novel source of inefficiency: an expectation that a protest will take place tomorrow can excessively sap incentives to coordinate on protesting today. Thus, potential protests crowd each other out. This can lead to a form of collective procrastination: access to more opportunities to protest can lower the citizens' welfare, as collective action becomes inefficiently delayed.

Word count: 9698

[*] 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.

1 Introduction

The timing of mass protests often appears random and is hard to predict: outpourings of anger often manifest grievances that have simmered for years or decades. In some instances, protests arrive in cycles, repeatedly swelling and waning even with no major change in fundamentals that would explain the pattern. For example, the Chilean protests in 2006, 2008, 2011, and 2019 were sparked by minor events such as changes to subway fares, but in fact responded to long-standing issues of low funding for education, economic inequality and disenfranchisement, as reflected in the chant: *no son 30 pesos, son 30 años* ("it's not about 30 pesos, it's about 30 years") (Borzutzky and Perry, 2021).

Moreover, when patterns of protest *do* follow fundamentals, they often respond to such things as current well-being and the threat or promise of future changes, rather than swings in the private benefits available to participants, which rational choice models of collective action would consider paramount (Tullock, 1971; Olson, 1965; Lichbach, 1995). For instance, the 2019–2020 protests in Hong Kong were triggered by the introduction of a proposed bill that would have allowed extraditions to mainland China. Similarly, Ukraine's 2013-2014 Euromaidan protests were sparked by Yanukovych postponing a promised integration agreement with the European Union, and instead seeking closer ties with Russia. Protesters, when asked, articulate such forward-looking rationales: "If we don't succeed now, our freedom of speech, our human rights, all will be gone."[1]

In this paper, we present a novel theory that provides a unified explanation for these stylized facts. 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.[2] Our model uses the machinery of global games (Carlsson and Van Damme, 1993) and is in many ways canonical.[3] The key driver of our results is that, in our model, the citizens are not motivated by private benefits, but (at least partly) by their own agency: the probability that their participation will be decisive.

The model yields a constellation of intuitive results that resonate with the substantive literature but are hard to obtain in conventional formal models of mass protest. Citizens motivated by their own agency respond not to the *expected* probability that a protest will

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

[2] In keeping with the literature on protests, we will speak of a *regime* and protesters seeking *regime change*, but the model applies equally to movements seeking major policy changes in democracies by non-electoral means.

[3] In global games, first used to study coordination games such as currency attacks (Morris and Shin, 1998), players obtain noisy information (*e.g.*, about regime strength) and then act simultaneously. The inability to coordinate behavior perfectly due to slight differences in information often yields equilibrium uniqueness.

succeed but to the *marginal* probability that their participation will change the outcome. As a result, citizens may participate even if all benefits from protesting are public, while costs are private. Because public benefits are simply the gap between continuation values under regime chance and the status quo, protests respond both to the "carrot" of a better post-revolutionary outcome and the "stick" of an increase in deprivation under the regime (Gurr, 1970), be it current or expected, and material or, *e.g.*, representational. In particular, a mere change in expectations can trigger a protest.

Moreover, agency-driven protesting tends to feature *cycles of protest*, even when fundamentals are stable over time. The reason has to do with the fact that citizens display a bias towards inaction, for two reasons: they do not fully internalize the social benefits of their participation, and they cannot be sure that others will join them if they protest, since the citizens' signals are imperfectly aligned. Under this bias, an expectation that citizens will coordinate on protesting tomorrow saps incentives to coordinate on protesting today, over and above their baseline static reluctance to act. Thus, even when fundamentals are high enough in every period that a protest would occur today *if this were the last chance*, the equilibrium features some periods in which citizens coordinate on protesting, and others in between where the expectation of a near-enough protest in the future crowds out today's would-be protest, inducing a form of *collective procrastination* or *paralysis of options*. Collective procrastination can be socially inefficient, to the point where citizens would be better off if future chances to protest were taken away, as this would spur them to act today. In particular, even when the status quo is steadily deteriorating, protests may arrive not when they are most profitable or likely to succeed, but when there are no second chances left.

On a technical level, the key difference between our model and existing models of protest is simply that we assume a finite population. When the population is finite, the participation of one additional individual citizen has a real—albeit small—probability of being decisive. In contrast, most models of protest assume a continuous population, in which each citizen necessarily acts as a pure "price-taker," who would never act in the absence of private benefits. Simply assuming any finite population size activates all of the channels that we focus on. Of course, if the relevant population numbers in the millions, the forces behind our results, though present, are very small, as each citizen is very unlikely to be pivotal. But, as we show in Section 6.1, pivotal protesting remains relevant for any population size if protesters display a modicum of altruism towards their fellow citizens, a form of civic-mindedness.[4]

[4] This point mirrors an observation in the voting literature that rational models of turnout predict unrealistically low turnout (Feddersen, 2004), but augmenting such models with a small degree of altruism towards fellow citizens (Myatt, 2015) or civic-duty motives (Feddersen and Sandroni, 2006) remedies this problem in a more satisfactory way than models of purely expressive voting, *e.g.*, by correctly predicting that turnout

2 Related Literature

Our paper contributes to a growing literature modeling mass protests as global games. In many such formal models (Casper and Tyson, 2014; Tyson and Smith, 2018; Bueno De Mesquita and Shadmehr, 2023), citizens are motivated by private benefits, available only to those who took part in a successful protest. Other models assume intangible "warm glow" payoffs from expressing discontent (Persson and Tabellini, 2009; Little, Tucker and LaGatta, 2015; Egorov and Sonin, 2021). If these payoffs accrue only when the protest succeeds (*e.g.* "pleasure in agency"; Wood 2003, Morris and Shadmehr 2023), they operate similarly to private benefits. In either case, the incentive to participate depends on the size of excludable benefits and the total (not marginal) probability of success.

In most of this literature, the population is infinite. Each citizen thinks herself powerless to change the outcome, so her action cannot shift the probability of receiving public benefits. Therefore, public benefits become irrelevant in the citizens' strategic calculus.

The upshot of this irrelevance becomes clear in models of *repeated* global games, which are closely related to our paper. In Angeletos, Hellwig and Pavan (2007) and Little (2017), an infinite number of agents—driven by private benefits—choose whether to attack a regime in each of many periods. The agents, though rational and forward-looking, behave myopically: information about future opportunities to attack, for example, has no impact on equilibrium play today. In particular, play in the first period is as in a static global game. The reason is that continuation payoffs if the game continues are an exogenous windfall from any individual agent's point of view. For the same reason, collective procrastination cannot arise.

This contrasts with our results, in which future threats and opportunities play a central role. A phenomenon reminiscent of protest cycles *can* arise in Angeletos et al. (2007) and Little (2017), though for different reasons: these models assume that regime strength does not change, so agents learn about it over time, both from new signals and from the very fact that the regime must have been relatively strong if it survived past attacks.

A broader formal literature studies games of regime change with a single opportunity to attack. The focus is often 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 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

will be higher in close elections (Blais, 2000).

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 manipulate payoffs or 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 dynamic cooperation games with exit: when one player exits (*e.g.*, attacks the other) the game ends. Their model is related to a variant of ours, discussed in Appendix A.9, 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.

In another strand of the literature, focused on intra-attack dynamics, there is a single attack which agents can join at different times (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 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."

3 The Model

We model a set $N = \{1, \ldots, n\}$ ($n \geq 2$) of citizens who repeatedly choose whether to "attack" (protest, mobilize) or not. Time is discrete and finite: $t \in \{1, \ldots, T\}$. 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 θ_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 $l_t = \frac{\sum_{i=1}^n a_{it}}{n}$ 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, 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 assume a common discount factor $\delta \in (0, 1)$. Letting u_{it} be player i's flow payoff in period t, we denote i's discounted payoffs from period t onwards by $U_{it} = \sum_{t\leq\tau\leq T} \delta^{\tau-t} u_{i\tau}$.

Flow payoffs are as follows. Each citizen 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 θ_t defined above, and the game ends. If the regime survives in period t, all agents instead accrue a known *status quo* flow payoff ν_t, and the game moves on to the next period.[5] Note that all agents receive either θ_t or ν_t, as appropriate, *regardless* of whether they attacked in that period.

Our solution concept is Perfect Bayesian Equilibrium.

Assumptions: Interpretation and Discussion

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).[6] In Section 6, we show that the general logic of our results survives if we allow for both private and public benefits.

Second, the payoff from revolution is affected by the state of the world, θ_t, but the probability of a successful revolt, $f(l_t)$, is not *directly* affected by the state. A natural interpretation is that θ_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

[5] As written, the model assumes that, after period T, there are no more protesting opportunities nor status quo payoffs. We could instead assume that status quo payoffs ν_{T+1}, ν_{T+2}, . . . keep accruing forever if the regime survives through period T. Adding such "post-terminal" payoffs is equivalent to bundling them into the period-T status quo payoff, *i.e.*, setting $\tilde{\nu}_T = \sum_{t\geq T} \delta^{t-T} \nu_t$.

[6] An exception is Shadmehr (2021), which also considers altruism as in Section 6.1, albeit in a static model.

there is uncertainty about the function f, or other payoff parameters such as ν_t or c.

Third, the probability of a successful revolt, $f(l_t)$, is increasing and convex in the size of the protest. The convexity assumption guarantees that even in the presence of pivotality concerns actions are strategic complements: the marginal impact of an additional protester is higher the more protesters there are.[7] This assumption best models settings in which overthrowing the regime is "hard" and requires a large mass of protesters, whereas concavity of f might be natural if a moderate crowd is sufficient, so there are diminishing returns for l_t large. The model is not intractable if we assume that f is concave—leading to strategic substitutability—though the equilibrium strategies would involve some degree of mixing, and procrastination would no longer arise due to fears of miscoordination.[8]

It is worth comparing our setup to the most popular payoff specification in global games (Morris and Shin, 1998; Dasgupta, 2007; Angeletos et al., 2007; Shadmehr, 2021; Little, 2017; Shadmehr and Bernhardt, 2019; Edmond, 2013), in which 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.[9] However, much like the canonical framework, our setup yields a tractable expression for the marginal payoff of protesting, which is the key object of interest.

Fourth, we assume that regime change ends the game. This assumption is less substantively restrictive than it might appear: the payoff θ_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 θ_t.

Finally, we assume that the state of the world θ_t is drawn independently across periods. This contrasts with Angeletos et al. (2007) and Little (2017), in which the state is drawn once. However, our model allows the *mean* of the state in each period to follow an arbitrary sequence $(\mu_t)_{t=1,\dots,T}$. 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 merely that the *idiosyncratic* uncertainty about θ_t is transient. (In our analysis, we focus on the case of σ_ϵ^2 small, so it is substantively unimportant whether the idiosyncratic shocks are persistent or transient.)

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

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

[9] Technically, $f_\theta(l) = \mathbb{1}_{l \geq 1-\theta}$ is a step function, hence not convex in l.

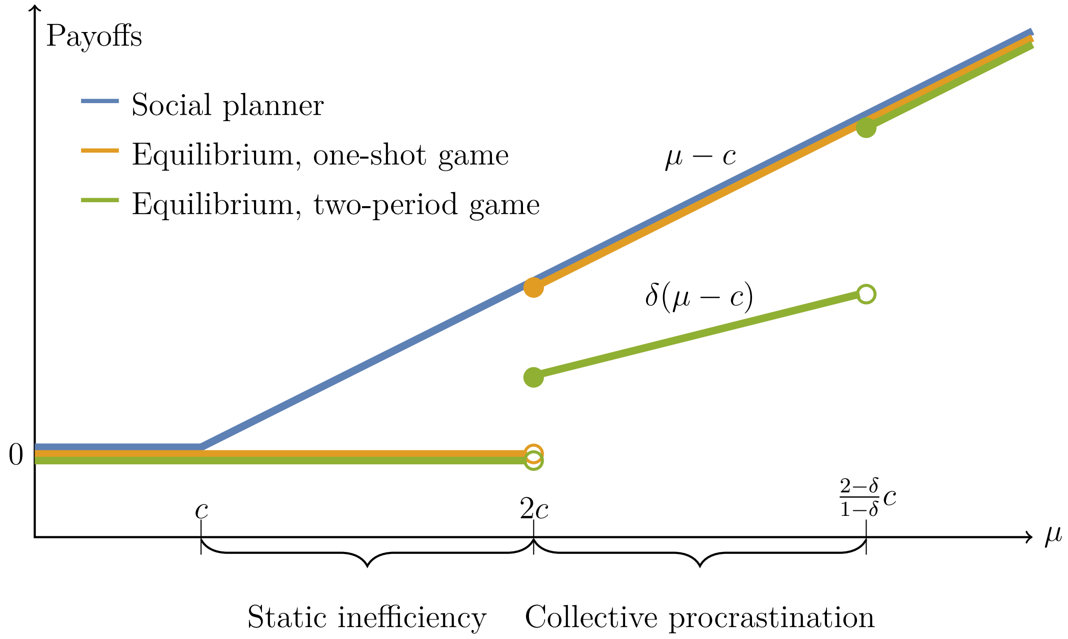

Figure 1: Equilibrium payoffs, when there is either one or two opportunities to protest, and socially optimal payoff, as a function of μ.

4 Two-Player Example

To build intuition, we start with a two-player, two-period example ($n = 2$, $T = 2$), with no status quo payoffs ($\nu_t \equiv 0$) and constant revolution payoffs in expectation ($\mu_1 = \mu_2 = \mu$). Regime change requires the participation of *both* citizens: $f(1) = 1$, $f(0.5) = f(0) = 0$. There is a small amount of state uncertainty, and signals are very slightly noisy: σ_θ, σ_ϵ are small, with $\sigma_\theta >> \sigma_\epsilon > 0$. We begin by characterizing the social planner's solution.

Remark 1. If $\mu < c$, then the social planner's solution is full abstention with probability going to 1 as $\sigma_\theta \to 0$. On the other hand, if $\mu > c$, then it is socially optimal to have both citizens protest in both periods with probability going to 1 as $\sigma_\theta \to 0$.

The result is intuitive: if $\theta_t < c$, then protesting is socially wasteful. If $\theta_t > c$, then protesting generates a net payoff of $\theta_t - c > 0$ per capita. Moreover, since $\mu_1 = \mu_2$ and σ_θ is small, it is almost always better to protest in the first period than to delay until the second period, as $\theta_1 - c \approx \mu - c > \delta(\mu - c) \approx \delta(\theta_2 - c)$. The per capita payoff induced by the social planner's solution is thus approximately $\max(\mu - c, 0)$, as illustrated in blue in Figure 1.

We now consider the noncooperative equilibrium of the game.

Remark 2. For σ_ϵ small enough, there is a unique[10] equilibrium in which each player i protests at time t if and only if $x_{it} \geq x_t^*$. As $\sigma_\theta \to 0$, the equilibrium outcome (with probability going

[10] There is also a non-participation equilibrium, but its existence is knife-edge: it disappears if we make $f(0.5)$ arbitrarily small but positive. We set $f(0.5) = 0$ to keep the algebra simple.

to 1) is full abstention if $\mu < 2c$; full protesting in both periods if $\mu > \frac{2-\delta}{1-\delta}c$; and protesting only in period 2 if $2c < \mu < \frac{2-\delta}{1-\delta}c$.

The existence of a unique equilibrium in threshold strategies is standard from global games. To find the equilibrium thresholds, consider the problem faced by i in period 2 if she sees a marginal signal $x_{i2} = x_2^*$. When σ_ϵ is very small, it is equally likely that the other citizen has a higher signal than hers ($x_{j2} > x_2^*$, hence j protests) or not ($x_{j2} < x_2^*$, hence j abstains), so i's expected payoff from protesting is $\frac{\theta_2}{2} + \frac{0}{2} - c \approx \frac{x_2^*}{2} + \frac{0}{2} - c$, while her abstention payoff is zero. Since i must be indifferent at the threshold, $x_2^* \approx 2c$. Thus, in period 2, the players obtain the socially optimal outcome only for $\mu > 2c$, as illustrated in orange in Figure 1.

Consider now period 1. If $\mu < 2c$, the same logic from period 2 dictates that the players abstain. If $\mu > 2c$, a player i with marginal signal $x_{i1} = x_1^*$ again believes that the other player will protest only with probability 0.5. But now, the expected payoff from regime survival is $\delta(\mu - c)$ rather than 0, as the regime would most likely fall tomorrow. Then i's abstention payoff is $\delta(\mu - c)$, while her protest payoff is $\frac{\theta_1}{2} + \frac{\delta(\mu-c)}{2} - c \approx \frac{x_1^*}{2} + \frac{\delta(\mu-c)}{2} - c$. Since i must be indifferent at the threshold, $x_1^* \approx \delta(\mu - c) + 2c > 2c$. For players to protest in period 1 with high probability, their signals must exceed this threshold, *i.e.*, we need $\mu \approx \theta_1 \approx x_{i1} \geq x_1^* \approx \delta(\mu - c) + 2c$, which implies $\mu \geq \frac{2-\delta}{1-\delta}c$.

To summarize, when regime change payoffs are very high ($\mu > \frac{2-\delta}{1-\delta}c$), the social optimum is still achieved. When they are moderate ($c < \mu < 2c$), both citizens (inefficiently) stay home, due to the familiar fear of miscoordination that arises even in the one-shot game when signals are slightly noisy. But, in an intermediate region ($2c < \mu < \frac{2-\delta}{1-\delta}c$), the citizens pass in period 1 and attack in period 2.

This is collective procrastination. It reflects how an expected successful protest tomorrow saps incentives to coordinate today. It is inefficient: as seen by comparing the green and orange payoffs in Figure 1, the players are worse off than if protesting in period 2 were impossible—as in that case they would coordinate on protesting today instead of tomorrow.

The appearance of collective procrastination hinges on the players' mutual fear of miscoordination.[11] Our insight is that this well-known source of inefficiencies in static global games has compounding effects in a dynamic game.

[11] To see why, note that if the state were commonly known in each period ($\sigma_\epsilon = 0$), then it would be an equilibrium for the players to follow the social planner's optimal strategy, which features no procrastination.

5 Analysis

We solve the general 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 Δ_{iT} be i's *marginal* payoff from attacking, given a signal observation x_{iT} and the other players' equilibrium strategies:

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

where c is the cost of protesting, n is population size, θ_T and ν_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. In equilibrium, i must attack if $\Delta_{iT} > 0$ and abstain if $\Delta_{iT} < 0$. 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 $x_T^*(\sigma_\epsilon)$. As $\sigma_\epsilon \to 0$, $x_T^*(\sigma_\epsilon)$ converges to a limit x_T^*, which equals*

$$x_T^* = \frac{cn}{f(1) - f(0)} + \nu_T.$$

Some properties of the equilibrium threshold are intuitive: higher costs of protesting c and better status quo payoffs ν_T both drive x_T^* up, discouraging protesting. On the other hand, x_T^* is decreasing in $\frac{f(1)-f(0)}{n}$, which is a measure of the citizen's "agency," *i.e.*, the likelihood that her participation will be decisive.

That the unique equilibrium is in threshold strategies follows from familiar arguments for global games. Here is an intuitive derivation of the threshold x_T^*. A citizen i whose signal x_{iT} equals x_T^* must be indifferent, *i.e.*, $\Delta_{iT}(x_T^*) = 0$. When σ_ϵ is small, x_{iT} is a precise signal of the state, so i believes that θ_T is close to x_T^*. On the other hand, as typically happens in global games, i's signal says very little about where it ranks *relative to other citizens' signals*; indeed, i expects that the fraction of citizens with higher signals than her own is approximately equally likely to be 0, $\frac{1}{n-1}$, $\ldots$, $\frac{n-2}{n-1}$, or 1.[12] Because it is precisely those citizens who will attack, $\tilde{l}_T | x_{iT} = x_T^*$ may equal 0, $\frac{1}{n}$, $\ldots$, or $\frac{n-1}{n}$, each with probability

[12] This is a discrete version of the well-known result that, in global games with an infinite population, the fraction of agents with signals higher than one's own is uniformly distributed between 0 and 1 (Morris and Shin, 2003).

approximately equal to $\frac{1}{n}$. Substituting all this into Equation (1),

$$\Delta_{iT}(x_T^*) \approx -c + (x_T^* - \nu_T) \sum_{j=0}^{n-1} \frac{f\left(\frac{j+1}{n}\right) - f\left(\frac{j}{n}\right)}{n} = -c + (x_T^* - \nu_T) \frac{f(1) - f(0)}{n}.$$

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

Our next observation is that game in all periods can be solved using exactly the same approach, with one difference. Let $\overline{U}_{t+1}(\sigma_\epsilon, \sigma_\theta)$ denote each citizen's 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 is

$$\Delta_{it} = -c + E\left[\left(\theta_t - \nu_t - \delta \overline{U}_{t+1}(\sigma_\epsilon, \sigma_\theta)\right)\left(f\left(\tilde{l}_T + \frac{1}{n}\right) - f(\tilde{l}_T)\right) \mid x_{it}\right], \tag{2}$$

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

Proposition 1. *For σ_ϵ small enough, the game has a unique equilibrium. In it, each citizen i attacks in period t if and only if x_{it} is weakly greater than a threshold $x_t^*(\sigma_\epsilon, \sigma_\theta)$.*

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 $\sigma_\theta \to 0$, $x_t^*(\sigma_\theta) \to x_t^*$, $\overline{U}_{t+1}(\sigma_\theta) \to \overline{U}_{t+1}$. The sequence of limit thresholds $x_1^*, \ldots, x_T^*$ and continuation utilities $\overline{U}_1, \ldots, \overline{U}_T$) is found by recursively solving the following system of equations for $t = T$, $T-1$, $\ldots$, 1:

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

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

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

Per Equation (3) the equilibrium threshold in all periods is as in Lemma 1, but now accounting for the continuation value $\delta \overline{U}_{t+1}$ of preserving the status quo. Note that, when σ_ϵ and σ_θ are both low, x_{it} is close to μ_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 $\mu_t - \nu_t > \frac{cn}{f(1)-f(0)} + \delta \overline{U}_{t+1}$. Thus a high potential 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, $-c + [f(1) - f(0)](\mu_t - \nu_t - \delta \overline{U}_{t+1})$ (per Equation (4)) is at least $-c + cn > 0$. 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 $n = 10$, $T = 6$, $f(l) = \frac{l+l^2}{4}$, $c = 0.1$, $\delta = 0.8$, and σ_ϵ, σ_θ are both small, with $\sigma_\epsilon << \sigma_\theta$. We assume $\mu_t = 3$ and $\nu_t = 0$ for all t: regime change and status quo payoffs are constant. Then there should be no reason to wait for a "better" moment (*i.e.*, higher θ_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 cycles. Indeed, in period 6, $\mu - \nu = 3 > 2 = \frac{0.1 \times 10}{[0.5-0]} = \frac{cn}{f(1)-f(0)}$, so there is a protest. But as a result, the continuation value in period 5, $\delta \overline{U}_6$, equals $0.8(-0.1 + 0.5 \times 3) = 1.12$, a value high enough that it tempts the citizens to abstain in period 5, as $\frac{cn}{f(1)-f(0)} + \delta \overline{U}_6 = 3.12 > 3$. In period 4, citizens are more impatient because they would have to wait two full periods for the next protest: $\delta^2 \overline{U}_6 = \delta \overline{U}_5 = 0.8 \times 1.12 = 0.896$, so $\frac{cn}{f(1)-f(0)} + \delta \overline{U}_5 = 2.896 < 3$, so a protest occurs. By similar logic, the citizens abstain in periods 2 and 3, and protest in period 1, having a 50% chance of success ($f(1) = 0.5$) with each attack.[13] In every period of abstention—$t = 2$, 3 and 5—the citizens fall victim to collective procrastination.

Just as in our two-period example from Section 4, having additional opportunities to protest can be harmful. For example, conditional on reaching period 5, the citizens' equilibrium utility is 1.12, but it would be 1.4 if protesting in period 6 were impossible, because they would then coordinate on attacking in period 5. More generally, 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 ν_t equals ν for all periods $t < T$, with $\nu_T =$*

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

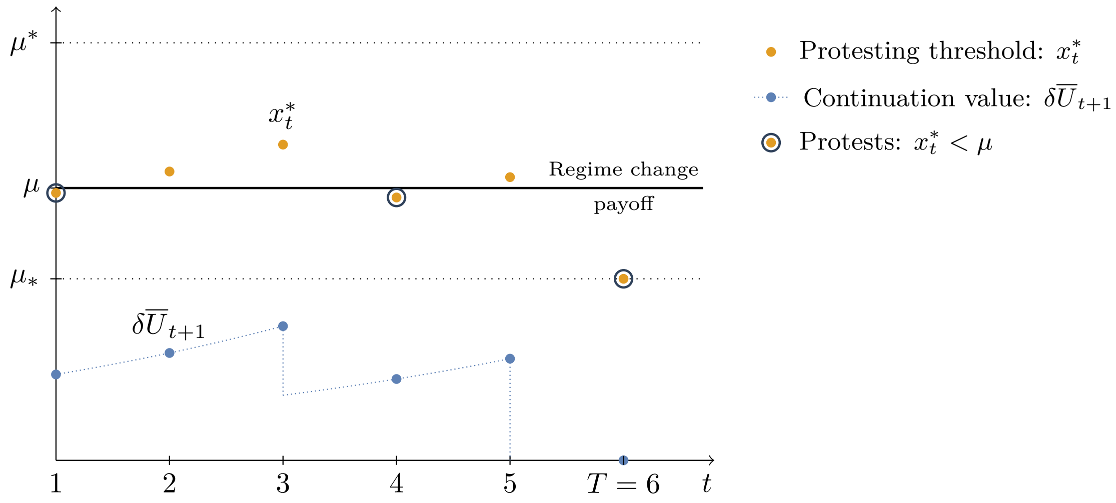

Figure 2: Pattern of attacks when the regime change payoff is intermediate: $\mu_t \equiv \mu$ between μ_* and μ^*

$\frac{\nu}{1-\delta}$.[14] *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, expected protest participation is close to 100%.

(ii) If $\mu_t = \mu < \mu_0$ for all t, then, in every period, expected participation is close to 0%.

(iii) Generically,[15] *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 with expected participation close to 100%, and arbitrarily many with expected participation close to 0%.*

Moreover

$$\begin{aligned}
\mu_0 &= \frac{cn}{f(1)-f(0)} + \frac{\nu}{1-\delta}, \\
\mu_* &= \frac{cn}{f(1)-f(0)} + \frac{\delta c}{1-\delta}\frac{nf(0)}{f(1)-f(0)} + \frac{\nu}{1-\delta}, \\
\mu^* &= \frac{cn}{f(1)-f(0)} + \frac{\delta c}{1-\delta}\left[\frac{nf(0)}{f(1)-f(0)} + n - 1\right] + \frac{\nu}{1-\delta}.
\end{aligned}$$

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:

[14] This amounts to assuming status quo payoffs of size ν for period T and all periods thereafter (see Footnote 5), which keeps continuation values constant over time in case of no attacks.

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

protesting in every period becomes impossible unless the payoff to protesting is extremely high ($\mu_t > \mu^*$).

Finally, Proposition 3 partially characterizes the model's comparative statics. It shows the effects of a marginal change in the parameters—in particular, $\mu_{t'}$ 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 a future period $t' > t$, it becomes preferable to let the regime survive at time t, for a chance to receive this increased payoff at time t'. The incentive to attack in period t decreases, and x_t^* increases. Similarly, if ν_t increases, the players are incentivized to let the regime survive today. On the other hand, an increase in μ_t has no effect on x_t^*—but makes players more likely to attack at time t, since it increases θ_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, until one of them crosses μ_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 an earlier period, etc.

We finish our analysis with a comparison of the equilibrium with the social planner's solution. We show that, in contrast to equilibrium behavior, the social planner only delays protests if justified by fundamentals (*e.g.*, μ_t increasing in t).

Remark 3. In the welfare-maximizing strategy profile, each citizen i attacks in period t if x_{it} is higher than a threshold, which for σ_ϵ small enough converges to

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

and abstains otherwise. Moreover,

(i) Citizens' payoffs weakly increase if μ_t or ν_t increase for any t.

(ii) If regime change and status quo payoffs are constant ($\mu_t \equiv \mu$; $\nu_t \equiv \nu$ for all $t < T$ and $\nu_T = \frac{\nu}{1-\delta}$; σ_θ and σ_ϵ small), 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 ($<$).

Per Equation (5), the social planner uses the same threshold for action as is used in the equilibrium of our main model (cf. Equation (3)) *if* there were a single citizen ($n = 1$). Part (i) of Remark 3 implies that, in the social planner's solution, there is no inefficient 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 citizens always attack or never attack.

The gap between the social planner's solution and the noncooperative equilibrium (compare Equations (3) and (5)) stems from two distinct forces: *fears of miscoordination* and *free riding*. That free riding might play a role is not surprising, since each citizen does not internalize the benefits her participation bestows on other citizens. But free riding is not always the main culprit. To see why, recall our two-player example from Section 4, in which $f(1) = 1$ and $f(0.5) = f(0) = 0$. There, if there is full information ($\sigma_\epsilon = 0$), the social planner's solution is also a noncooperative equilibrium. Indeed, since full participation is required to win ($f(0.5) = 0$), free riding is impossible: a citizen cannot gain from abandoning her partner when both are supposed to protest. Thus, the source of equilibrium inefficiency when $\sigma_\epsilon > 0$ is solely that each citizen cannot know for sure if her partner will join her.

More generally, the degree of collective action and resulting inefficiency in i) the social planner's solution, ii) the equilibrium under full information, and iii) the equilibrium with noisy information stem from the players' effective agency in each case, which compare as follows:

$$\underbrace{f(1) - f(0)}_{\text{social planner's agency}} \quad \geq \quad \underbrace{f(1) - f\left(\frac{n-1}{n}\right)}_{\text{citizen's agency, full information}} \quad \geq \quad \underbrace{\frac{f(1) - f(0)}{n}}_{\text{citizen's agency, noisy information}} \tag{6}$$

Indeed, the planner can shift participation from 0 to 1 if she chooses. Each citizen can only shift it by $\frac{1}{n}$. But, under full information, a marginal citizen can know that others are participating, so she effectively chooses between $l_t = \frac{n-1}{n}$ and $l_t = 1$, while under noisy signals the marginal citizen is much more uncertain about others' behavior.

When f is very convex, so that $f\left(\frac{n-1}{n}\right)$ is close to $f(0)$, the citizens can do as well as the planner under full information, so that fears of miscoordination under noisy information are the main source of equilibrium inefficiency. On the contrary, if f is close to linear, then $f(1) - f\left(\frac{n-1}{n}\right)$ is close to $\frac{f(1)-f(0)}{n}$, and most inefficiency is caused by free riding. Intuitively, when f is linear, there is no coordination motive; on the contrary, the temptation to let the other citizens "handle the problem" is high.

6 Extensions

This Section presents three extensions. The first shows that, if the citizens are even slightly altruistic towards each other, then the incentive to engage in "pivotal protesting" remains large for any population size. The second shows that our qualitative results continue to hold if private benefits (*i.e.*, "club goods") are available in addition to public ones. The third adds to the model a notion of changes in expectations, and shows how the citizens' protest behavior today responds to new information about the future. Finally, in the 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 Altruism and Agency

As can be seen from Equation (3), the equilibrium thresholds x_t^* increase without bound as $n \to \infty$. Thus, holding all other parameters constant, protesting becomes impossible if the population is large enough. This result reflects the well-known insight that collective action in large populations is undone by free riding if all benefits are public (Olson, 1965).

In light of this, can pivotal protesting truly serve as a model of mass protests? We argue that it can, if the citizens in question exhibit even slight *other-regarding*, or altruistic, preferences towards their fellow citizens. Formally, suppose that each citizen i's objective function V_{it} at time t puts a weight 1 on her "hedonic" utility and a weight $\alpha \in [0, 1]$ on the hedonic utility of each other citizen. That is, $V_{it} = U_{it} + \alpha \sum_{j \neq i} U_{jt}$, where $U_{jt} = \sum_{t \leq \tau \leq T} \delta^{\tau - t} u_{i\tau}$ is j's continuation hedonic utility from t onwards, as defined in Section 3.

This model coincides with our baseline model when $\alpha = 0$. In contrast, if $\alpha = 1$, each citizen would behave as a social planner, putting equal weight on each player's welfare (including her own). For general α, i's marginal payoff from attacking becomes

$$\Delta_{it} = -c + [1 + \alpha(n-1)]\, E\left[\left(\theta_t - \nu_t - \delta\overline{U}_{t+1}(\sigma_\epsilon, \sigma_\theta)\right)\left(f\left(\tilde{l}_T + \frac{1}{n}\right) - f(\tilde{l}_T)\right) \mid x_{it}\right], \quad (7)$$

because whenever her participation changes the outcome, she (weight 1) as well as $n-1$ others (weight α on each) receive the windfall $\theta_t - \nu_t - \delta\overline{U}_{t+1}(\sigma_\epsilon, \sigma_\theta)$ (cf. Equation 2). The citizen's effective agency is then $[1+\alpha(n-1)]\frac{f(1)-f(0)}{n}$ (cf. Equation 6), which converges to $\alpha[f(1)-f(0)]$, rather than to 0, for large n. The equilibrium threshold from Equation (3) becomes

$$x_t^* = \frac{n}{1+\alpha(n-1)}\frac{c}{f(1)-f(0)} + \nu_t + \delta\overline{U}_{t+1}. \quad (8)$$

Proposition 3 extends for all $\alpha < 1$—in particular, there are thresholds $\mu_*(\alpha) < \mu^*(\alpha)$ such that there is procrastination and cycles when μ_t lies between them.[16] At the same time, as $n \to \infty$, the equilibrium thresholds x_t^* now converge not to infinity but to $\frac{c}{\alpha[f(1)-f(0)]} + \nu_t + \delta\overline{U}_{t+1}$, which is finite for any $\alpha > 0$. In other words, for any $\alpha \in (0,1)$, incentives to engage in pivotal protesting remain relevant in large populations *and* entail the same general consequences that we have characterized in the baseline model.

Other-regarding preferences have previously been proposed as a solution to the turnout paradox (Feddersen, 2004; Blais, 2000), *i.e.*, an explanation for why rational, selfish voters would vote in large elections where pivotality is unlikely (Edlin, Gelman and Kaplan, 2007; Jankowski, 2007; Fowler, 2006; Myatt, 2015).[17] They have been less explored in formal models of protest.[18] But protesting, like voting, is a form of civic expression, and arguably the closest substitute available in non-democratic societies, so protest participation could plausibly be driven by similar motives.

It is worth noting that the formal literature on protests broadly considers two types of potential motivations for citizens to act: private benefits (Olson, 1965; Tullock, 1971), that is, material or social benefits of regime change that are exclusive to participants; and psychological rewards, such as frustration in response to relative deprivation (Gurr, 1970) and "pleasure in agency" (Wood, 2003).[19] But typically, most participants in mass protests do not receive (or expect) material rewards. If a notion of social or psychological benefits is

[16] This is shown formally in the proof of Proposition 3. In the extreme case $\alpha = 1$, the equilibrium must replicate the social planner's solution, so procrastination disappears.

[17] Models of "ethical voting" (Coate and Conlin, 2004; Feddersen and Sandroni, 2006) are also closely related.

[18] The exception is Shadmehr (2021), which considers protests by altruistic citizens, albeit only in a static model.

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

required to explain participation, then other-regarding preferences are in principle no more or less plausible than the usual operationalization of psychological rewards as expressive "warm glow" payoffs (Persson and Tabellini, 2009; Little et al., 2015; Egorov and Sonin, 2021).

6.2 News Shocks

Mass protests often respond to events that shift expectations about the future, even ones that leave current material conditions unaffected. Examples include public announcements, proposed bills and agreements, and political developments abroad.[20] As our examples in the Introduction and Section 7 show, this is an empirically common and important phenomenon. Unlike other dynamic models of mass protest,[21] our model can provide a natural explanation for it, if augmented to allow for "news shocks".

In our baseline model, μ_t and ν_t are commonly known parameters. We can instead assume that, for each t, μ_t and ν_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 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 $\overline{U}_{t+1}$ would now average over the possible values of μ_{t+1}, ν_{t+1}, the players' equilibrium actions as a function of these parameters, and the next period's continuation value, $\overline{U}_{t+2}$. Any information received at time t about μ_τ or ν_τ (for $\tau > t$) would constitute a "news shock".

For brevity, we illustrate the impact of news shocks in an example. Let $f(l) = \frac{2l+l^2}{8}$, $c = 1$, $\delta = 0.8$, and $n = 10$. Assume that $\mu_t \equiv 0$, but ν_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. 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. As the yellow state is not materially

[20] Consider, for instance, the forward thinking encapsulated in the chilling slogan used by Taiwanese protesters: "Today's Hong Kong, tomorrow's Taiwan." https://foreignpolicy.com/2014/08/19/todays-hong-kong-tomorrows-taiwan/

[21] For instance, in Angeletos et al. (2007) or Little (2017), equilibrium behavior is *always* independent of all expectations about the future, even if the citizens are forward-looking.

worse than the green one, a switch to the yellow state is a pure news shock.

Denote the moment the state turns yellow by t_0. Using Equations (3) and (4), we can show that citizens attack in every red period (from $t_0 + 3$ onwards) if $\underline{\nu} < -12.533$. If $\underline{\nu} < -25.067$, citizens also attack in the last yellow period, $t_0 + 2$. If $\underline{\nu} < -50.133$, they also attack in period $t_0 + 1$ And, if $\underline{\nu} < -100.27$, they also attack in period t_0, as soon as the state becomes yellow. They thus become more prone to protesting the more imminent the red state is. On the other hand, they will not attack in the green state so long as $\underline{\nu} > -1830$, because a switch to the yellow state is not particularly likely. In particular, for $\underline{\nu}$ between -1830 and -101, the citizens are peaceful in the green state, but react to news shocks: they begin protesting as soon as the yellow state is realized, even though their current payoffs are unchanged.

6.3 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 ρ of regime change benefits are private: if the regime falls at time t protesters receive θ_t and abstainers receive only $(1-\rho)\theta_t$. ($\rho \in [0,1]$ is commonly known.) Then i's marginal payoff from protesting at time t becomes

$$\Delta_{it} = -c + E\left[((1-\rho)\theta_t - \nu_t - \delta\overline{U}_{t+1})\left(f\left(\tilde{l}_t + \frac{1}{n}\right) - f\left(\tilde{l}_t\right)\right) + \rho\theta_t f\left(\tilde{l}_t + \frac{1}{n}\right) \mid x_{it}\right],$$

where $\rho\theta_t$, i's private benefit, is received with probability $f\left(\tilde{l}_t + \frac{1}{n}\right)$ if she participates and 0 otherwise, while the additional probability of receiving the net public benefit, $(1-\rho)\theta_t - \nu_t - \delta\overline{U}_{t+1}$, is i's probability of being pivotal, as before. 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, which now converges for small σ_ϵ to

$$x_t^* = \frac{c + \frac{f(1)-f(0)}{n}(\nu_t + \delta\overline{U}_{t+1})}{(1-\rho)\frac{f(1)-f(0)}{n} + \rho\frac{\sum_{j=1}^{n} f\left(\frac{j}{n}\right)}{n}}. \tag{9}$$

A derivation of Equation (9) can be found in the Appendix. Note that, when $\rho = 0$, this simplifies to Equation (3).

From Equation (9) it follows that, even when there are private benefits ($\rho > 0$), or even if *all* benefits are private ($\rho = 1$), pivotality concerns are active: continuation values matter in the citizens' strategic calculus (*i.e.*, x_t^* increases in $\delta\overline{U}_{t+1}$), and similar arguments as in our main model show that protest cycles can still result for any finite n.

However, the higher n is, the closer this model becomes to canonical models of protest (Morris and Shin, 2003; Angeletos et al., 2007; Little, 2017), in which the population is infinite and private benefits are the only driver of behavior. Indeed, setting $\rho > 0$ and taking $n \to \infty$, Equation (9) becomes

$$x_t^* = \frac{c}{\rho \int_0^1 f(l)dl}. \tag{10}$$

The continuation utility, $\overline{U}_{t+1}$, vanishes from the expression: when agents expect to never be pivotal, their behavior becomes *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.

In contrast to the tendency of agency-driven protesters to delay collective action, protesters who are solely (or mostly) motivated by excludable benefits may be inefficiently slow *or* quick to act, precisely because they disregard the future in their calculations. They might prematurely "jump the gun" amid improving conditions (μ_t, ν_t increasing), and conversely may fail to react to an approaching catastrophe (μ_t, ν_t sharply decreasing) if current regime change payoffs are not tempting enough. Moreover, their behavior would not give rise to protest cycles: the threshold in Equation (10) is constant over time, so if μ_t is constant, there will be attacks in all periods (if $\mu_t > x_t^*$) or none ($<$).[22]

7 Discussion

Our model generates several empirical predictions that do not naturally follow from existing formal models of protests. First, public benefits can drive protest behavior, whether private benefits are present or not. Second, cycles of protest can arise even when the underlying grievances are long-standing, with little meaningful change over time. In such instances, periods of delay between protests reflect collective procrastination, as it would be socially preferable to protest in all periods. And third, "news shocks" can be impactful: for instance, threatening bills can cause protests, while a promise to hold elections can defuse them.

We now discuss some examples as suggestive evidence of these phenomena at work. That public benefits can drive protest behavior—in particular, that dissatisfaction with the status

[22] Dynamic models of protests with private benefits (Angeletos et al., 2007; Little, 2017) find that intermittent attacks are possible *if the state is hidden and persistent* (*i.e.*, θ is drawn only once and remains fixed); the agents keep receiving exogenous signals of it; and these signals happen to periodically compensate for the "bad news" that regime survival itself implies. But other outcomes are also possible, such as equilibria in which the citizens give up after a single failed attack. In contrast, in our model, protest cycles are a robust phenomenon.

quo precedes protest movements—is an empirical regularity, satisfied by all of the examples we mention. As discussed in the Introduction, the Chilean protests in 2006, 2008, 2011, and 2019 also display an apparent pattern of protest cycles. Minor events, such as a 4% hike in metro fares in 2019, served as the proximate cause of each outbreak. Yet the roots of discontent lay in issues such as economic inequality, insufficient public education, and perceived disenfranchisement, which had been fixtures of the Chilean landscape since the Pinochet dictatorship (Borzutzky and Perry, 2021). Similarly, the 2013 protests in Brazil were triggered by public transportation fare hikes, but responded more broadly to perennial issues such as the "state of public infrastructure . . . public spending . . . corruption, urban violence, and a 'fed-up-ness' with the state of the country" (Alonso and Mische, 2017), which had also triggered previous protests (Alonso and Mische, 2017, p. 152).[23] It has been argued more generally that contentious movements are inherently cyclical in nature (Tarrow, 2011; Hirschman, 1982). The existing literature proposes various explanations for such waves or cycles, such as cycles of disappointment that shift citizens' focus back and forth between private consumption and public action (Hirschman, 1982), and protest fatigue and responses by the regime (Tarrow, 2011). Our model shows that even in the absence of any such phenomena, the dynamic discouragement effects we characterize induce cycles in agency-driven protesting. Another alternative explanation for delays in collective action, based on *preference falsification* (Kuran, 1989), seems less appropriate for examples such as Chile or Brazil, where severe restrictions on speech were not present and past protests, elections, etc., provided plenty of information about public sentiment.

Though our model suggests that collective action will coalesce into cyclical outbreaks even absent obvious triggers, it also predicts that significant exogenous shifts in threats and opportunities (Tilly, 1978) will give shape to these cycles when present (see also Hirschman (1982) (pp. 4-6) on this point). A specific prediction that is novel to our model in the formal literature is that even pure "news shocks" that only concern the future can have such effects.

Examples show that both negative and positive news shocks can be impactful. For example, the 2019 proposal of a bill in Hong Kong that would have allowed extraditions to mainland China prompted marches numbering over a million protesters, which ultimately challenged not just the proposed bill but also the legitimacy of Hong Kong's government and police force. The protests boiled over even before the bill was to be formally discussed in the legislature (Purbrick, 2019), and continued even after the bill was shelved,[24] ending only after mainland China directly imposed a national security law that criminalized dissent.[25]

[23] See also Koopmans (1993) for a discussion of protest cycles in Western Europe during the Cold War.

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

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

Clearly, the protesters reacted before the bill could have had any material consequences; the driver for action was instead what the proposal signaled about Hong Kong's political future. Similarly, an earlier wave of protest in 2003 followed a proposed national security bill, while the 2014 Umbrella Revolution 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."[26] That the protesters' most forceful bid for change only took place in 2019–2020 is arguably a sign of collective procrastination: as the mainland's resolve to bring Hong Kong under its heel had come to harden throughout the 2010s, decisive action might have been more effective had it come earlier.

The 2014 Euromaidan revolution in Ukraine was also triggered by a negative news shock. After years of negotiations with the European Union and promises of European integration, the Yanukovych administration announced in November 2013 that it was suspending plans to sign a broad association agreement with the EU, only a week before the scheduled signing. Instead, Ukraine would seek closer ties with Russia, which had threatened trade sanctions in response to the EU deal. Protesters gathered, spurred by the threat that their chance to finally escape the Russian sphere of influence—to no longer live in "a post-Soviet barrack temporarily repainted in yellow and blue"—would evaporate.[27] Ukraine failed to sign the EU agreement as scheduled, even as both sides claimed that a deal was still on the table.[28] The protests grew in number and scope of demands, and turned into riots even as the government responded with a package of draconian anti-protest laws,[29] and violent crackdowns that killed over 100 protesters in total. Soon, widespread desertion among demoralized police forces forced Yanukovych to flee to Russia.[30] In the Ukrainian case, too, the sources of resentment—poverty, corruption, and low democratic legitimacy—had plagued the country for years, ever since its transition out of communism. But the prospect of European integration had offered hope that the status quo would improve, helping to stave off collective action until the breakdown in negotiations.

The same forces explain why protest movements sometimes stagnate in anticipation of

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

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

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

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

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

elections: elections hollow out the incentive to protest by offering a less costly avenue for change (Hafner-Burton, Hyde and Jablonski, 2018). In particular, then, a surprise call for new elections can serve as a positive news shock that defuses collective action. Indeed, Marsteintredet and Berntzen (2008) argue that calls for early elections were successfully used as a "last resort to solve an ongoing political conflict", *e.g.*, in Bolivia (1995) or the Dominican Republic (1996).

Conversely, a controversial incumbent's reelection is often succeeded by anti-regime protests. This is consistent with the logic of collective procrastination: citizens who were hoping for change at the ballot box may finally coordinate on protesting if, after the election, they perceive the next opportunity to attain change peacefully as too remote. The widespread protests following Orban's reelections in 2018 and 2022—which stood in stark contrast with the relative tranquility in the streets leading up to the elections—can be seen in this light.[31] Another example is the 2017 Women's March in the United States, held the day after Donald Trump's inauguration. In reverse fashion, the repeated postponement of elections in Bolivia in 2020 sparked widespread protests.[32]

8 Conclusions

In this paper we develop a dynamic model of protests in which citizens act driven by the desire to bring about change, even if the benefits from regime change accrue even to non-participants. 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. 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. When such protest cycles occur, the citizens would be better off protesting in all periods, but are tempted to drag their feet in between periods of expected coordination.

The dynamic encouragement and discouragement effects that are central to our analysis are absent from models of repeated protests driven by private benefits. Within our theory, they are the source of predictions that find support in the substantive literature on mass protests and are empirically plausible, yet are novel to the formal literature on the topic.

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

[31] https://www.nytimes.com/2018/04/14/world/europe/hungary-protest-orban.html

[32] https://www.nytimes.com/2020/08/07/world/americas/bolivia-roadblock-blockade.html

beliefs over time to defuse protests? For example, the government may increase clientelistic transfers when the threat of revolt spikes. Likewise, promises to hold new elections as an alternative to immediate resignation, discussed in the previous section, ought in fact to be modeled as news shocks that are not exogenous, but follow from a strategic choice by the regime.

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.

References

Alonso, Angela and Ann Mische, "Changing Repertoires and Partisan Ambivalence in the New Brazilian Protests," *Bulletin of Latin American Research*, 2017, *36* (2), 144–159.

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.

Borzutzky, Silvia and Sarah Perry, "" It Is Not About the 30 Pesos, It Is About the 30 Years": Chile's Elitist Democracy, Social Movements, and the October 18 Protests," *The Latin Americanist*, 2021, *65* (2), 207–232.

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.

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.

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

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

Hafner-Burton, Emilie M, Susan D Hyde, and Ryan S Jablonski, "Surviving Elections: Election Violence, Incumbent Victory and Post-Election Repercussions," *British Journal of Political Science*, 2018, *48* (2), 459–488.

Hirschman, Albert O., *Shifting Involvements: Private Interest and Public Action*, Princeton University Press, 1982.

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.

Koopmans, Ruud, “The Dynamics of Protest Waves: West Germany, 1965 to 1989,” *American Sociological Review*, 1993, pp. 637–658.

Kuran, Timur, “Sparks and Prairie Fires: A Theory of Unanticipated Political Revolution,” *Public Choice*, 1989, *61* (1), 41–74.

_ , “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., “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.

Marsteintredet, Leiv and Einar Berntzen, “Reducing the Perils of Presidentialism in Latin America Through Presidential Interruptions,” *Comparative Politics*, 2008, *41* (1), 83–101.

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.

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.

Tarrow, Sidney G., *Power in Movement: Social Movements and Contentious Politics*, third ed., Cambridge University Press, 2011.

Tilly, Charles, *From Mobilization to Revolution*, Addison-Wesley, 1978.

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 Online Supporting Information

Contents

A.1 Proof of Remark 1

By Remark 3, the social planner has each player attack in period 2 if $x_{i2} > \frac{c}{1-0} = c$ for σ_ϵ small. If $\mu > c$, then this holds almost surely for σ_θ, σ_ϵ small. If $\mu < c$, then this fails to hold almost surely for σ_θ, σ_ϵ small.

In period 1, the social planner's limit threshold is $x_1^{sp} = c + \delta \max(\mu - c, 0)$, so we again have $x_{1t} > x_1^{sp}$ almost surely if $\mu > c$ and vice versa if $\mu < c$.

□

A.2 Proof of Remark 2

By Proposition 1, the limit threshold x_2^* equals $\frac{c \times 2}{1-0} = 2c$. The limit period 2 utility is then $\overline{U}_2 = (\mu - c)\mathbb{1}_{\mu > c}$. The limit threshold in period 1, x_1^*, is then $\frac{c \times 2}{1-0} + \delta(\mu - c)\mathbb{1}_{\mu > c}$. So we have a likely attack in period 1 ($\mu_1 > x_1^*$) if $\mu > 2c + \delta(\mu - c)\mathbb{1}_{\mu > c}$, or $\mu > \frac{2-\delta}{1-\delta}c$.

□

A.3 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 three 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

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

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 ν_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 σ_ϵ. Third, as we work with a finite number of players, aggregate outcomes are random even conditional on the realized state θ_t.

We write the proofs of our main results to cover the baseline model as well as the case of altruistic citizens, letting $\tilde{\alpha} = \frac{1+(n-1)\alpha}{n}$ and $\tilde{f}(l) := n\left[f\left(l + \frac{1}{n}\right) - f\left(l\right)\right]$.

(i) Supermodularity. Formally, denote j's strategy in period T by A_{jT}, the set of realizations of x_{jT} for which j attacks. Let $(A_{jT})_{j \in N}$, $(\tilde{A}_{jT})_{j \in N}$ 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 (1) 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: 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. We argue that (a) agents never want to attack when their signals are low enough, no matter what others will do, and (b) when restricting attention to strategies that respect this constraint, supermodularity does hold.

Remark 4. (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 5. 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 from the inverse Mills ratio formula (see Greene (2003), p. 759). □

Remark 6. Given a fixed n, suppose that in period t the players (excluding i) each protest with probability z. Then $n\tilde{l}_t \sim B(n-1, z)$.

Remark 7. $\int_0^1 \left[\binom{n-1}{k} z^k (1-z)^{n-1-k}\right] dz = \frac{1}{n}$ for all $k = 0, 1, \ldots, n-1$.

Proof. For $k = n-1$, this reduces to $\int_0^1 z^{n-1} dz = \frac{1}{n}$, which is trivial. Then we can prove the claim by induction, as

$$\begin{aligned}
&\binom{n-1}{k}\int_0^1 z^k(1-z)^{n-1-k}dz\\
=&\binom{n-1}{k}\left[\frac{1}{k+1}z^{k+1}(1-z)^{n-1-k}\Big|_0^1+\frac{n-1-k}{k+1}\int_0^1 z^{k+1}(1-z)^{n-2-k}dz\right]\\
=&\binom{n-1}{k}\frac{n-1-k}{k+1}\int_0^1 z^{k+1}(1-z)^{n-2-k}dz=\binom{n-1}{k+1}\int_0^1 z^{k+1}(1-z)^{n-2-k}dz=\frac{1}{n}.
\end{aligned}$$

□

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

Proof. Rewriting Equation (7) with the notation for $\tilde{\alpha}$ and $\tilde{f}$, we find that, for any $x_{iT} \leq \frac{c}{2\tilde{\alpha}f'(1)} + \nu_T$,

$$\begin{aligned}
\Delta_{iT}(x_{iT}) &= -c + \tilde{\alpha}E((\theta_T-\nu_T)\tilde{f}(\tilde{l}_T)|x_{iT})\\
&\leq -c + \tilde{\alpha}E((\theta_T-\nu_T)\tilde{f}(\tilde{l}_T)\mathbb{1}_{\theta_T\geq\nu_T}|x_{iT})\\
&\leq -c + \tilde{\alpha}f'(1)E((\theta_T-\nu_T)\mathbb{1}_{\theta_T\geq\nu_T}|x_{iT})\\
&\leq -c + \tilde{\alpha}f'(1)(E(\theta_T|x_{iT},\theta_T\geq\nu_T)-\nu_T)\\
&\leq -c + \tilde{\alpha}f'(1)\left[\max\left(\frac{\sigma^2_\epsilon\mu_T+\sigma^2_\theta x_{iT}}{\sigma^2_\theta+\sigma^2_\epsilon},\nu_T\right)+\sqrt{\frac{2}{\pi}}\frac{\sigma_\theta\sigma_\epsilon}{\sqrt{\sigma^2_\theta+\sigma^2_\epsilon}}-\nu_T\right]\\
&\leq -c + \tilde{\alpha}f'(1)\left[\max\left(\frac{\sigma^2_\epsilon(\mu_T-\nu_T)+\sigma^2_\theta\frac{c}{2\tilde{\alpha}f'(1)}}{\sigma^2_\theta+\sigma^2_\epsilon},0\right)+\sigma_\epsilon\right].
\end{aligned}$$

(Note that $\tilde{f}(\tilde{l}) \leq f'(1)$ for $\tilde{l} \leq \frac{n-1}{n}$ by the convexity of f.)

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

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

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

Now assume $\sigma_\epsilon^2 < \overline{\sigma}_{\epsilon 1}^2$ and consider a modified game in which, when an agent i receives a signal $x_{iT} < \frac{c}{2\tilde{\alpha} f'(1)} + \nu_T$, she is forced mechanically to abstain, while for $x_{iT} \geq \frac{c}{2\tilde{\alpha} f'(1)} + \nu_T$ 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 σ_ϵ^2 small enough.

Lemma 3. *Assume that* $\sigma_\epsilon^2 < \overline{\sigma}_{\epsilon 2}^2 = \min\left(\overline{\sigma}_{\epsilon 1}^2, \sigma_\theta^2 \frac{c}{4\tilde{\alpha} f'(1)} \frac{1}{|\mu_T - \nu_T - \frac{c}{4\tilde{\alpha} f'(1)}|}, \left(\frac{c}{4\tilde{\alpha} f'(1)}\right)^2 \frac{1}{\ln(\overline{f''}) - \ln(\underline{f''})}\right)$, *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* $x_{iT} \geq \frac{c}{2\tilde{\alpha} f'(1)} + \nu_T$, $\Delta_{iT}(x_{iT})$ *is weakly increasing in* A_T.

Proof. Consider two strategy profiles A_T, $\hat{A}_T \subseteq [\frac{c}{2\tilde{\alpha} f'(1)} + \nu_T, \infty) \times N$ such that $A_{jT} \subseteq \hat{A}_{jT}$ for all j. For any i and any $x_{iT} \geq \frac{c}{2\tilde{\alpha} f'(1)} + \nu_T$, we will compare $\Delta_{iT}(x_{iT})$ to $\hat{\Delta}_{iT}(x_{iT})$. To simplify notation, we will drop the T indices. Denote by $g(\theta|x)$ the posterior density of the state given i's signal x_i. We then have

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

It is enough to show that this last expression is at least zero.[34] Now note that, for all θ,

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

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

$$\overline{f''} \int_{-\infty}^{\nu} (\theta - \nu) \phi\left(\frac{\theta - x}{\sigma_\epsilon}\right) g(\theta|x_i) d\theta + \underline{f''} \int_{\nu}^{\infty} (\theta - \nu) \phi\left(\frac{\theta - x}{\sigma_\epsilon}\right) g(\theta|x_i) d\theta \geq 0. \tag{11}$$

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 $x = x_i = \frac{c}{2\tilde{\alpha} f'(1)} + \nu$ then it will automatically follow for all other x,

[34] The last inequality follows from the fact that $f(y+a) - f(y) - f(x+a) + f(x) = \int_x^{x+a} \left[\int_{\tilde{x}}^{\tilde{x}+y-x} f''(z) dz\right] d\tilde{x}$.

x_i. The reason is that, if Equation (11) holds, then

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

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 $g(\theta|x_i)=\frac{\sqrt{\sigma_\theta^2+\sigma_\epsilon^2}}{\sigma_\theta\sigma_\epsilon}\phi\left(\frac{\theta-\frac{\sigma_\epsilon^2\mu+\sigma_\theta^2x_i}{\sigma_\theta^2+\sigma_\epsilon^2}}{\frac{\sigma_\theta\sigma_\epsilon}{\sqrt{\sigma_\theta^2+\sigma_\epsilon^2}}}\right)$ are both MLRP-increasing in x and x_i, respectively (*i.e.*, $\frac{g(\theta|x_i')}{g(\theta|x_i)}$ is increasing in θ 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 θ for $x'>x$).

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

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

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

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

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

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

where $x_0=\nu+\frac{c}{4\tilde{\alpha}f'(1)}$. 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 $\theta=\nu+\frac{c}{4\tilde{\alpha}f'(1)}+r$, *i.e.*, it is enough to show

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

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

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

and hence enough to show

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

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

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

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

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

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

where we have used that $\tilde{f}(0) \geq f'(0)$ because f is convex. This expression is positive whenever $\sigma_\epsilon^2\tilde{\alpha}f'(0)(\mu_T - \nu_T) + 2c\sigma_\theta^2 > c(\sigma_\theta^2 + \sigma_\epsilon^2)$, or equivalently $\sigma_\epsilon^2(c - \tilde{\alpha}f'(0)(\mu_T - \nu_T)) < \sigma_\theta^2 c$ or $\sigma_\epsilon^2\left(\frac{c}{\tilde{\alpha}f'(0)} - \mu_T + \nu_T\right) < \sigma_\theta^2\frac{c}{\tilde{\alpha}f'(0)}$. □

(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 $x, x' \geq \frac{c}{2\tilde{\alpha}f'(1)} + \nu_T$ and $\sigma \in (0, \overline{\sigma}_{\epsilon 3})$.*

Proof. Recall that

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

Note that $(\theta - \nu_T)E\left[\tilde{f}(\tilde{l})|\theta\right]$ is negative for $\theta < \nu_T$. It is positive and strictly increasing in θ for $\theta > \nu_T$ (because $\theta - \nu_T$ is strictly increasing in θ; $\tilde{f}$ is increasing, by the convexity of f; and $\tilde{l}$ is FOSD-increasing in θ because, taking the ϵ_{jt} fixed, each $x_{jt}|\theta$ is in fact pointwise increasing in θ). 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 θ, and (b) for each $\theta_0 < \nu_T$, $g(\theta_0|x)$ is decreasing in x for all $x \geq \frac{c}{2\tilde{\alpha}f'(1)} + \nu_T$. (a) follows from Remark 4. (b) holds because $\phi(z)$ 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 $\frac{\sigma_\epsilon^2\mu_T+\sigma_\theta^2 x}{\sigma_\theta^2+\sigma_\epsilon^2} \geq \nu_T + \frac{c}{4\tilde{\alpha}f'(1)} \geq \nu_T > \theta_0$ by Lemma 4.

□

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

(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 σ_ϵ 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

$$\Delta(x, x, \sigma_\epsilon) = -c + \tilde{\alpha}\int_{-\infty}^{\infty}(\theta - \nu)E\left[\tilde{f}(\tilde{l})|\theta\right]\frac{\sqrt{\sigma_\theta^2+\sigma_\epsilon^2}}{\sigma_\theta\sigma_\epsilon}\phi\left(\frac{\theta - \frac{\sigma_\epsilon^2\mu+\sigma_\theta^2 x}{\sigma_\theta^2+\sigma_\epsilon^2}}{\frac{\sigma_\theta\sigma_\epsilon}{\sqrt{\sigma_\theta^2+\sigma_\epsilon^2}}}\right)d\theta.$$

Since all agents $j \neq i$ attack iff $x_j \geq x$, conditional on θ, each agent attacks with probability $z := \Phi\left(\frac{\theta-x}{\sigma_\epsilon}\right)$, drawn independently. Then $\tilde{l}|\theta \sim \frac{B(n-1,z)}{n}$ by Remark 6, *i.e.*, for $k = 0, \ldots, n-1$,

$$\Pr\left(\tilde{l} = \frac{k}{n}\right) = \binom{n-1}{k}z^k(1-z)^{n-1-k}.$$

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 our previous expression as:

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

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

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, μ, ν, σ_ϵ, σ_θ that is well defined and C^∞ 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 $h(\cdot)E[\tilde{f}(\tilde{l})|z]$ is uniformly bounded by an integrable function for all σ_ϵ below a threshold. Indeed, $\tilde{f} \leq n$. 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 σ_ϵ^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 h can be bounded by a function of the form $\frac{a+|\ln(z)|}{z^\beta}$, and is well behaved for any σ_ϵ such that the exponent β 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 Δ is a continuous function of its arguments, in particular at $\sigma_\epsilon = 0$, where

$$
\begin{aligned}
\Delta(x,x,0) &= -c + \tilde{\alpha}\int_0^1 (x-\nu)E[\tilde{f}(\tilde{l})|z]dz \\
&= -c + \tilde{\alpha}(x-\nu)\int_0^1 \sum_{k=0}^{n-1} \Pr\left(\tilde{l} = \frac{k}{n}\Big| z\right)\left(f\left(\frac{k+1}{n}\right) - f\left(\frac{k}{n}\right)\right)dz \\
&= -c + \tilde{\alpha}(x-\nu)\sum_{k=0}^{n-1}\left(f\left(\frac{k+1}{n}\right) - f\left(\frac{k}{n}\right)\right)\int_0^1 \Pr\left(\tilde{l} = \frac{k}{n}\Big| z\right)dz \\
&= -c + \tilde{\alpha}(x-\nu)\sum_{k=0}^{n-1}\left(f\left(\frac{k+1}{n}\right) - f\left(\frac{k}{n}\right)\right)\frac{1}{n} \\
&= -c + \tilde{\alpha}(x-\nu)[f(1) - f(0)],
\end{aligned}
$$

which yields the limit threshold x_t^*. (Note that we have used Remark 7 in the fourth step.) But we need to go a step further. To prove that Δ is strictly increasing in x for σ_ϵ small, we will show that $\frac{\partial\Delta}{\partial x}(x,x,\sigma_\epsilon)$ converges uniformly to $\frac{\partial\Delta}{\partial x}(x,x,0) \equiv \tilde{\alpha}[f(1) - f(0)] > 0$ 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

$$
\frac{\partial E[\tilde{f}(\tilde{l})|z]h(\cdot)}{\partial x} = E[\tilde{f}(\tilde{l})|z]\frac{\sqrt{\sigma_\theta^2+\sigma_\epsilon^2}}{\sigma_\theta}\frac{\phi\left(w\right)}{\phi(\psi(z))} + (x-\nu+\sigma_\epsilon\psi(z))E[\tilde{f}(\tilde{l})|z]\frac{\sigma_\epsilon}{\sigma_\theta^2}\frac{\phi\left(w\right)}{\phi(\psi(z))}w.
$$

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 σ_ϵ is small. Hence this expression is bounded (uniformly for x, μ, ν, σ_θ, and σ_ϵ in any closed intervals, with σ_θ strictly positive) by an integrable function. The Leibniz integral rule then implies that $\frac{\partial\Delta}{\partial x}(x,x,\sigma_\epsilon) \equiv \tilde{\alpha}\int_0^1 \frac{\partial E[\tilde{f}(\tilde{l})|z]h(\cdot)}{\partial x}dz$. Moreover, for any convergent sequence $Y_k = (x_k, \mu_k, \nu_k, \sigma_{\theta k}, \sigma_{\epsilon k})$ with limit Y_∞, 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 $E[\tilde{f}(\tilde{l})|z]h(\cdot)$ 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, $\frac{\partial\Delta}{\partial x}(x,\mu,\nu,\sigma_\theta,\sigma_\epsilon) - \frac{\partial\Delta}{\partial x}(x,\mu,\nu,\sigma_\theta,0) < \frac{\tilde{\alpha}}{2}[f(1) - f(0)]$, whence $\frac{\partial\Delta}{\partial x}(x,\mu,\nu,\sigma_\theta,\sigma_\epsilon) > 0$. In particular, taking the range of x to contain $\left[\underline{\nu} + \frac{c}{2\tilde{\alpha}f'(1)}, \overline{\nu} + \frac{2c}{\tilde{\alpha}f'(0)}\right]$, 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 $\sigma_\epsilon \to 0$, $x^*(\sigma_\epsilon) \to \frac{c}{\tilde{\alpha}[f(1)-f(0)]} + \nu$; indeed, if not, there would be $\eta_0 > 0$ and a sequence $\sigma_k \to 0$ such that either $x^*(\sigma_k) \geq \frac{c}{\tilde{\alpha}[f(1)-f(0)]} + \nu + \eta_0$ for all k or $x^*(\sigma_k) \leq \frac{c}{\tilde{\alpha}[f(1)-f(0)]} + \nu - \eta_0$ for all k. But our formula for $\Delta(x, x, 0)$ and the continuity of Δ would imply that, for k high enough, $\Delta(x, x, \sigma_k) > 0$ at any $x \geq \frac{c}{\tilde{\alpha}[f(1)-f(0)]} + \nu + \eta_0$, and $\Delta(x, x, \sigma_k) < 0$ at any $x \leq \frac{c}{\tilde{\alpha}[f(1)-f(0)]} + \nu - \eta_0$, a contradiction.

□

A.4 Proof of Proposition 1

The marginal payoff from attacking in period t is given by the expression

$$\Delta_{it} = -c + E\left[\tilde{\alpha}(\theta_t - \nu_t - \delta\overline{U}_{t+1})\tilde{f}(\tilde{l}_t)|x_{it}\right].$$

By the same argument as in Lemma 1, for σ_ϵ 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, ν) 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 σ_ϵ; hence the continuation value $\delta\overline{U}_{t+1}$ is a function of σ_ϵ. 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 ν 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 Δ (in particular with respect to both σ_ϵ and ν), 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

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

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 σ_ϵ and σ_θ, 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

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

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

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

as we wanted.

□

A.5 Proof of Proposition 2

The generalizations of μ_0, μ_* and μ^* to the case of $\alpha > 0$ are

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

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

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

Since $\mu < \mu_0$, as σ_θ 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

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

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

$$\begin{aligned}\overline{U}_t &= f(0)\frac{1 - \delta^{T-t+1}(1 - f(0))^{T-t+1}}{1 - \delta(1 - f(0))}\mu + \left[1 - f(0)\frac{1 - \delta^{T-t+1}(1 - f(0))^{T-t+1}}{1 - \delta(1 - f(0))}\right]\frac{\nu}{1-\delta} \\ x_{t-1}^* &= \frac{c}{\tilde{\alpha}[f(1) - f(0)]} + \nu + \delta\overline{U}_t,\end{aligned}$$

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 μ and $\frac{\nu}{1-\delta}$, hence higher than μ.

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 $\tilde{\alpha} = 1$ 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.*,

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

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 σ_ϵ 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 μ_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

$$\overline{U}_t = -c + f(1)\mu_t + (1-f(1))(\nu + \delta\overline{U}_{t+1}) \tag{12}$$

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

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

Equation (12) implies that $\overline{U}_t$, and $\overline{U}_{t'}$ for all $t' < t$, are increasing functions of μ_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

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

$$\iff \mu^* \leq \underline{\mu}.$$

Indeed, by construction, μ^* 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

$$\overline{U}_t = f(0)\mu_t + (1 - f(0))(\nu + \delta \overline{U}_{t+1}) \tag{13}$$

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

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

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

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

$$\iff \mu_* \geq \overline{\mu}.$$

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

□

A.6 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 $\overline{U}_{t+2}$, ..., which only depends on $\nu_{t'}$ through $\overline{U}_{t'}$. So

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

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 μ_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 $t' > t$,

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

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 $\mu_{t'} < x_{t'}^*$.

□

A.7 Proof of Remark 3

The social planner aims to maximize the sum of the citizens' (expected) ex ante utilities, $\sum_{i=1}^{n} \sum_{t=1}^{T} \delta^t u_{it}$. The threshold in Equation (5) then follows from the following argument.

Let σ^* be a strategy profile that maximizes social welfare. (It is not hard to prove existence.) Consider now the game G^α with altruistic citizens (Section 6.1) with altruism parameter α. (The solution of this model is given in the proofs of Lemma 1 and Proposition 1.) Take $\alpha = 1$, so that every citizen has the same payoff function, which coincides with the planner's.

We claim that σ^* must be an equilibrium of G^1. Indeed, if it is not, then, starting from the conjectured strategy profile σ^*, there must be a player i with a profitable deviation from σ_i^* to another strategy $\tilde{\sigma}_i$. But, since all the players have the same payoff function as the planner, this means that social welfare must be higher under $(\tilde{\sigma}_i, \sigma_{-i}^*)$ than under σ^*, contradicting the optimality of σ^*. Since G^1 has a unique equilibrium (Proposition 1), the threshold x_t^{sp} then comes simply from setting $\alpha = 1$ in Equation (8).

(i) follows from the fact that the planner's payoff from any fixed strategy profile weakly increases as μ_t or ν_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).

(ii) follows from the fact that, per Proposition 2, $\mu_0(1) = \mu_*(1) = \mu^*(1)$ if $f(0) = 0$. Then there are protests in every period if $\mu > \mu^*$ and never if $\mu < \mu^*$.

□

A.8 Derivation of Equation (9)

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

$$
\begin{aligned}
0 = & -c + E\left[((1-\rho)\theta_t - \nu_t - \delta\overline{U}_{t+1})\left(f\left(\tilde{l}_t + \frac{1}{n}\right) - f\left(\tilde{l}_t\right)\right) + \rho\theta_t f\left(\tilde{l}_t + \frac{1}{n}\right)|x_{it} = x\right] \\
\xrightarrow[\sigma_\epsilon \to 0]{} & -c + ((1-\rho)x - \nu_t - \delta\overline{U}_{t+1})\frac{f(1) - f(0)}{n} + \rho x \frac{\sum_{k=1}^{n} f\left(\frac{k}{n}\right)}{n},
\end{aligned}
$$

which yields Equation (9).

□

A.9 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 θ_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 ν_t *once* and the game ends. (Of course, ν_t can represent a discounted sum of payoffs.) Demonstrating still costs c. The probability that the movement survives period t is $f(l_t)$.

Then the net payoff of demonstrating for the marginal agent is

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

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

$$x_t^* = \frac{c}{\tilde{\alpha}[f(1) - f(0)]} + \nu_t - \delta\overline{U}_{t+1}. \tag{14}$$

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 **encourages** 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 (14) 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 $\mu^* = \frac{c}{\tilde{\alpha}[f(1)-f(0)]}$ 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 μ_t or ν_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.