Altruism in Protests[*]

Germán Gieczewski[†] and Korhan Kocak[‡]

March 2023.

Abstract

This paper studies a model of collective action in which citizens face repeated opportunities to protest against a regime, and observe noisy signals of the potential gains from attacking at each moment. We depart from the existing literature by assuming that agents are partially altruistic. This assumption has far-reaching implications. First, citizens can be motivated by public benefits in addition to private ones. Second, the continuation value of the status quo influences the citizens' willingness to protest today. As a result, a revolt may be triggered by a mere change in expectations about the future. The same logic can induce a pattern of intermittent attacks, as well introduce a novel source of inefficiency: a temptation to attack later rather than earlier when future attacking opportunities are relatively attractive leads to attacks being inefficiently delayed. Thus, altruistic agents can fall prey to a form of collective procrastination that lowers social welfare.

1 Introduction

Citizens who take part in mass protests, especially those seeking regime change or concessions from a non-democratic government, often expose themselves to significant danger. Collectively, the potential gains from a successful protest can more than

[*] We are thankful to Federica Izzo, Dimitri Landa, seminar participants at Princeton, Sabanci, and UCSD 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.

justify the costs. But each individual protester is unlikely to make a difference, and if she does, most of the value generated accrues to *other* citizens. Thus, a selfish benefit-cost calculus can rarely justify the bulk of individual decisions that translate into mass participation.

The dilemma faced by potential protesters is thus a classic problem of collective action. Formal models of protests have typically solved this dilemma in one of two ways. First, following analogous economic models (e.g., of currency attacks), protesting may be assumed to confer *private* benefits, which accrue only to participants, rather than all citizens, in case of a success.[1] This approach more plausibly captures the calculus of protest leaders and organizers than that of the masses. Second, protesters may be modeled as driven by grievances or expressive "warm glow" benefits which, though plausible, leaves us with a less than fully-specified model, in the absence of a theory of *why* certain events and circumstances aggrieve people.[2]

In this paper, we take a novel approach to modeling protest movements. We assume that citizens—rather than selfish—are partially altruistic. That is, they value the well-being of their fellow citizens, though perhaps less than their own. They may thus protest to contribute to the public good—not to obtain some private or expressive benefits unavailable to fence-sitters (although our model can accommodate such benefits).

The model we present is otherwise simple and canonical. We construct a dynamic global game in which citizens repeatedly choose whether to "attack" a regime, at a cost. Citizens have imperfect information about the profitability of regime change. Each citizen can attack any number of times. The game ends when the regime falls, or after it survives a given number of periods. Success is more likely if more citizens attack. A larger crowd makes the marginal participant more impactful, so that there is a coordination motive.

In some ways, the analysis is standard. Information perturbations, combined with two-sided limit dominance, yield equilibrium uniqueness and sharp predictions: in each period, citizens attack if their information about the profitability of attacking is favorable enough. But the addition of altruism to the model flips on its head much of the conventional logic from existing models of protests.

[1] This type of benefit is related to the concepts of "selective incentives" (Olson, 1965) or "club goods" (Buchanan, 1965).

[2] A third solution is to simply assume the problem away by modeling the protesters as a unitary-decision maker; see, e.g., Acemoglu and Robinson (2001), Shadmehr (2014).

In models with either private or "warm glow" benefits, the citizens' motivation to attack is proportional to the attack's probability of succeeding, and rewarding them with spoils or satisfaction.[3] In contrast, in our model, citizens are motivated by their *contribution* to the probability of success. For instance, suppose a citizen knows that the probability of a successful attack is 0.5001 if she participates, and 0.5 if not, in a population of 10,000. For a selfish citizen chasing private benefits, the incentive to attack would be 0.5001 times her (potential) private benefit. If there is also a public benefit, a selfish citizen multiplies its valuation by $0.5001 - 0.5 = 0.0001$ in her calculation, likely rendering it irrelevant. In contrast, a fully altruistic citizen (who values the well-being of her fellow citizens as much as her own) would multiply the public benefit by $0.0001 \times 10{,}000 = 1$ in her calculation, because in the unlikely event that her participation tips the scales, *all* citizens receive the public benefit thanks to her. More generally, as the number of players goes to infinity, a decrease in each citizen's probability of being pivotal is offset by an increase in the number of players who benefit from a potential success. As a result, public motives can remain relevant for altruistic citizens, even in a large population.

Moreover, since the collective gain from a success is simply the payoff *gap* between the status quo and regime change, incentives to attack respond both to the "carrot" of a better post-revolutionary outcome and the "stick" of a more oppressive status quo. Hence a mere worsening of the status quo can trigger a protest in our model, whereas, in a private benefits setting, it would be the size of excludable spoils that is paramount.

Our main results, however, arise when the model allows for multiple opportunities to protest. In that case, the distinct logic of altruistic protesting has far-reaching ramifications: altruistic citizens realize that, by successfully overthrowing the regime today, they are forfeiting chances to instead overthrow the regime in the future. In other words, a successful protest robs all citizens of the option value of future protests. Their calculus must account for this. As a result, they are more likely to attack today if conditions for an attack are good today *or if they are bad tomorrow*, and vice versa; their behavior reflects forward-looking considerations. By contrast, selfish agents would take the outcome of the protest today as practically outside of their control, and participate if the odds of success are high, regardless of future

[3] Unless expressive payoffs are obtained even in case of failure, in which case the probability of success is irrelevant.

opportunities. Their actions would thus be divorced from future expectations *even if the agents themselves were forward-looking*, as shown by Angeletos, Hellwig and Pavan (2007) and Little (2017).

An immediate implication is that, in our model, protests may be triggered by the mere *expectation* that the status quo will deteriorate if nothing is done, even in the absence of any material changes in the environment (e.g., the regime's strength, the cost of participation, or current flow payoffs). To our knowledge, ours is the first dynamic model of regime change with many players able to explain this phenomenon, which is remarkably common. For example, the 2019 Hong Kong protests began in response to a *proposed* extradition bill that was seen as jeopardizing the region's autonomy from mainland China, long before the law could be ratified, let alone executed; the 2013-2014 Euromaidan revolution was triggered by the Ukrainian government's backpedaling in ongoing negotiations aimed at integration with Europe; the 2013 Gezi Park protests in Turkey were precipitated by the increasingly authoritarian rhetoric of the country's prime minister.[4] These and other examples do not readily fit an image of self-interested protesters opportunistically chasing the spoils of victory, or responding in knee-jerk fashion to material deprivation. They instead suggest citizens who assess the likely evolution of the status quo when deciding whether to protest today.

The welfare consequences of forward-looking protest behavior are ambiguous. It is clear why forward-looking protesters may attain better outcomes. For instance, suppose that the gains from overthrowing the regime are governed by a parameter θ which strongly increases over time (e.g., because the opposition can groom a better candidate for the new government if given more time.) Altruistic agents would wait to attack in later periods when the value of a success is high. Selfish agents, on the other hand, would attack as soon as θ is high enough that an attack would be focal today *in the absence of any future chances.*

Our most surprising result, however, is that when agents are *imperfectly* altruistic—that is, they value others' welfare, but less than their own—the "option value" considerations induced by altruism can lead to excessive and inefficient delay in equilibrium, a form of collective procrastination. More precisely, giving citizens more opportuni-

[4] These examples are discussed further in Section 6. Other examples include the 2008 protests in Argentina against a planned hike on grain export taxes and the 2010–2012 protests in Greece in response to proposed austerity measures.

ties to protest in the future, even *less* attractive ones than the current one, can lower their equilibrium welfare by inducing a sort of "paralysis of options." The logic of this result is related to the intuition behind equilibrium selection in all global games: in global games, agents can sometimes coordinate on an attack, but they need the state of the world to be somewhat better than the bare minimum needed to render an all-out attack profitable. A crowd in a global game thus behaves much like a person with low motivation or willpower to choose a high-effort action. Offering such a person an "out" in the form of a second chance can tempt her to procrastinate, leaving her worse off. Thus an attack may come not when it is most profitable, but rather when there are no second chances left.

Due to the same logic, the equilibrium generally displays a pattern of intermittent protest: like waves crashing against the shoreline, citizens eventually coordinate on an attack, then—if unsuccessful—let several periods pass before trying again, and the process repeats. These waves are strategic and forward-looking: citizens attack when the *anticipated* delay until the next wave crashes is high enough that they become impatient. Intermittent attacks can also arise in existing models (Angeletos et al., 2007; Little, 2017), but driven by a different, backward-looking logic: after a failed attack, citizens know that the regime is strong enough to have survived, and so they give up on attacking again until enough uncertainty has re-accumulated about the regime's changing strength.

The rest of the paper is structured as follows. Section 2 reviews the related literature. Section 3 presents the baseline model and Section 4 solves it. Section 5 shows how to add private benefits and more general uncertainty to the model, and considers a variant in which the citizens are fighting to keep a resistance movement alive rather than attempting to overthrow the regime. Section 6 discusses the results through the lens of recent protest movements in Hong Kong, Ukraine and Turkey. Section 7 concludes the paper.

2 Related Literature

To our knowledge, we are the first to consider altruism in a dynamic model of protests. While the idea that collective action must be motivated by private benefits is widespread (Olson, 1965), there are settings close to ours for which this framework is known to be inadequate. For instance, rational, instrumental models of voter turnout are known to predict unrealistically low levels of turnout (Feddersen, 2004; Blais, 2000). High turnout in large elections is better explained by models including "civic duty" (Feddersen and Sandroni, 2006; Coate and Conlin, 2004) or altruistic motives (Edlin, Gelman and Kaplan, 2007; Jankowski, 2002; Fowler, 2006; Rotemberg, 2009; Fowler and Kam, 2007). Protest participation—like voting—is a form of civic expression, and arguably *the* closest substitute for voting available to citizens in a non-democratic society. With the exception of Shadmehr (2021), which focuses on altruistic protesters in a static setting, whether protesting can be motivated by similar impulses is an understudied question.

The most popular non-instrumental approach in the protest literature has been to assume "warm glow" payoffs (Persson and Tabellini, 2009; Egorov and Sonin, 2021). Our altruism-based approach can be seen as a microfoundation of warm glow: rather than add a fixed additive term to the players' payoffs representing moral concerns, we model how the strength of these concerns would depend on the actual consequences of a player's actions.

We follow the literature on regime change in modeling mass protests using the technical machinery of global games (Carlsson and Van Damme, 1993),[5] first applied by economists interested in coordination games such as bank runs and currency attacks (Morris and Shin, 1998).

The bulk of the literature on regime change studies static games (Morris and Shin, 2004; Edmond, 2013; Bueno De Mesquita, 2010; Barbera and Jackson, 2020), even though the intended applications are often dynamic in nature—for example, investors choose not just whether to run against the currency, but when; protestors can choose when to demonstrate against the regime, etc. There is also a smaller but growing literature that explicitly accounts for the dynamics of collective action, albeit with selfish agents.

Like our paper, Angeletos et al. (2007) models a population of agents who choose whether to attack a regime of *fixed* strength in each of many periods. In the first period, agents play as in a static global game. In the second period, agents infer that—if the game is still going—the regime must have been strong enough to survive the attack faced in the first period. This creates common knowledge that the regime's

[5] In a global game, players obtain noisy information—for example, about the strength of a currency or the stability of a regime—and then act simultaneously. The inability to coordinate behavior perfectly due to slight differences in information typically yields equilibrium uniqueness in static models.

strength is above some threshold, which naturally results in equilibrium multiplicity. In particular, there is always an equilibrium where no attack occurs after the first period. The paper then shows that, if new information arrives or the regime's strength changes over time, repeated attacks are possible once the signal of past survival has lost its relevance. Little (2017) extends their model to allow for payoffs to vary over time and for the game to continue with a new regime after a coup.

Another thread of the literature on dynamic attacks considers a single attack, which agents can join at different times. In this vein, Dasgupta (2007) compares the uptake of a risky action in a multi-period game to its one-shot counterpart. In a setting where early in the game agents face higher uncertainty, but also potentially greater returns, he shows that the option to delay decreases uptake in the first period relative to the static benchmark. But late uptake compensates for this decrease. Thus, the option to delay reduces coordination failure.

Relatedly, Shadmehr and Bernhardt (2019) studies a two-player and two-period model in which both citizens must protest for regime change to occur. A player who protests in the first period encourages a less optimistic partner to join—but also risks being the lone protester. Thus, both citizens are tempted to wait for the other to start a revolution, resulting in a coordination failure. Shadmehr and Bernhardt compare this to an alternative model where only one of the two players, a vanguard, can initiate protests. Here, the incentive to free-ride disappears, and the vanguard is more likely to act than either citizen in the benchmark model.

3 The Model

We model a set N of players who must repeatedly choose whether to "attack" (protest, mobilize) or not. In our main specification, the set of players is a continuum: $N = [0, 1]$. However, to clarify some issues related to the scaling of payoffs and pivotal probabilities as the population grows, we will also discuss the case of a finite population in Section 4.

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

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}, where

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

and $\epsilon_{it} \sim N(0, \sigma_\epsilon^2)$. (The random variables θ_t, ϵ_{it} are independent across all players and periods.)

Each player $i \in N$ then simultaneously chooses whether to attack ($a_{it} = 1$) or not ($a_{it} = 0$). These actions result in the regime being overthrown with probability $f(l_t) = a_0 + a_1 l_t + a_2 l_t^2$, where l_t denotes the fraction of the population who attack in period t. If the regime falls, the players receive some terminal payoffs, to be described below, and the game ends. With probability $1 - f(l_t)$, the game continues on to the next period. (At the end of period T, the game ends even if the regime survives.) We assume that f is convex with $a_0 \geq 0$, $a_1, a_2 > 0$, and $a_0 + a_1 + a_2 \leq 1$.

Payoffs

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

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

In other words, each player puts weight α on the utility of each other individual player, and weight 1 on her own utility. For example, $\alpha = 0$ models completely self-interested players, while $\alpha = 1$ models fully altruistic players that consider the welfare of others just as important as their own, as a social planner would. (Note that Equation 1 only yields a well-defined utility function in the case of a finite population. However, the resultant expression for the *marginal* payoff that players obtain from attacking—which is the key object of interest—extends in a natural way to the case of an infinite population. See Section 4 for details.)

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

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

Hedonic payoffs are as follows. Each agent i who attacks in a period t bears a flow cost of attacking $c > 0$ in that period. If the regime falls in period t, then all agents also receive a one-time payoff 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.[6] 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

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

First, we assume that the benefits from a successful revolt are public. Although there is evidence that both private and public benefits matter in practice (Cantoni, Yang, Yuchtman and Zhang, 2019), models in this literature typically focus on private benefits (Angeletos et al., 2007; Edmond, 2013; Little, 2017). (An exception is Shadmehr (2021).) In Section 5, we show how our model can be extended to account for both private and public benefits; the logic of the model carries through so long as public benefits are present.

Second, we allow for some degree of altruism. This assumption is what keeps public benefits relevant in the agents' benefit-cost calculation as the population becomes large and, hence, the probability of being pivotal goes to zero. Shadmehr (2021) is a rare example that also focuses on pivotality, and also assumes that public benefits scale with the population size.

Third, the payoff from revolution is affected by the state of the world, θ_t, in that period, but the probability of a successful revolt, $f(l_t)$, is not *directly* affected by the state. A natural interpretation is that θ_t measures a determinant of the expected outcome after a revolution—for example, the ideology or the competence of a *de facto* opposition leader—rather than the regime's ability to stave off protesters.

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

This assumption is for simplicity; qualitatively similar results are obtained if there is uncertainty about the function f, or about other payoff parameters such as ν_t or c.

It is worth comparing our setup to the two most popular payoff specifications in global games. First, in many models (Morris and Shin, 2001; Little, 2016), attackers receive $\theta + l - 1$, while non-attackers receive 0. Our model generates a very similar specification for the *marginal* payoff from attacking. But these papers have no concept of an attack being "successful" or "unsuccessful" in a binary sense. Second, in other models (Morris and Shin, 1998; Dasgupta, 2007; Angeletos et al., 2007; Shadmehr, 2021; Little, 2017; Shadmehr and Bernhardt, 2019; Edmond, 2013), attackers receive $1 - c$ if successful and $-c$ if unsuccessful, but are only successful if $l \geq 1 - \theta$. This specification is inconvenient for our purposes because it only yields a supermodular game when pivotality concerns—which are central to our analysis—are absent.

Fourth, we assume that regime change ends the game. This assumption is less substantively restrictive than it might appear: the payoff θ_t represents the citizens' expected continuation utility from a new regime starting in period $t + 1$. The new regime could itself face protests if it is unpopular, as in the 2013 Egyptian protests leading to Morsi's removal a year after he became president. Such possibilities are all captured by the payoff θ_t.

Fifth, the probability of a successful revolt, $f(l_t)$, is strictly increasing and strictly convex in the size of the protest—this is the content of the assumption $a_1, a_2 > 0$. The convexity assumption guarantees supermodularity in the presence of pivotality concerns, by ensuring that the more other agents protest, the higher is the marginal impact of an additional protester.[7,8] The model is not intractable if we instead assume, for example, that f is concave—leading to strategic substitutability, as found by Cantoni et al. (2019)—though the equilibrium strategies will involve some degree of mixing. We keep to the case of strategic complementarity mainly to make our analysis comparable to most of the literature, which models protests as coordination games.

Finally, we assume that the state of the world $(\theta_t)_t$ is drawn independently across periods. This contrasts with Angeletos et al. (2007) and Little (2017), in which the state is fixed over time, or else affected by *persistent* random shocks. Note, however,

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

[8] Shadmehr (2021) assumes $f(l_t)$ to be a step function. In that setting, if success is "easy" in the sense that a few participants can ensure regime change, more participation leads to a lower likelihood any individual will be pivotal and thus actions are strategic substitutes. If the threshold for regime change is high, then actions are strategic complements.

that our model allows the *mean* of the state in each period to follow an arbitrary sequence $(\mu_t)_t$. In Section 5 we allow for general uncertainty and shocks, including persistent ones, as long as they are commonly observed; the key assumption keeping our model tractable is that the *idiosyncratic* uncertainty about θ_t—which supports unique equilibrium selection—must be transient. (In our main results, we will focus on the case of σ_ϵ^2 small, so it will be substantively unimportant whether the idiosyncratic shocks are persistent or transient.)

4 Analysis

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

Let Δ_{it} be an agent i's *marginal* payoff from attacking, given a signal observation x_{it} and expected equilibrium strategies of the other players. In equilibrium, i must attack if $\Delta_{it} > 0$ and not attack if $\Delta_{it} < 0$.

To see how marginal payoffs should be calculated in this setting, it is instructive to consider the case of a large but finite population.[9] Suppose that $N = \{1, \ldots, n\}$. Then the marginal payoff from attacking is

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

where $\tilde{l}_t = \frac{\sum_{j \neq i} a_{jt}}{n}$ is the fraction of the population who attacks, assuming that i does not attack. As $n \to \infty$, both $\tilde{l}_t$ and $\tilde{l}_t + \frac{1}{n}$ converge to l_t, while

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

In particular, note that pivotality concerns vanish from the model if and only if $\alpha = 0$: in other words, if the agents are even slightly altruistic, they must take public benefits into account. The reason is that, as the population grows, the impact

[9] It is preferable not to work directly with a finite population in the main model, because the finiteness would reintroduce aggregate uncertainty about the distribution of signals even conditional on the state, which complicates the analysis.

of a successful revolt on total welfare increases proportionally, while the probability of a single agent being pivotal decreases proportionally; these two forces offset each other.[10]

In the case of a continuous population, then, we define

$$\begin{aligned}\Delta_{iT} &= -c + E\left[\alpha(\theta_T - \nu_T) f'(l_T) | x_{iT}\right] \\ &= -c + E\left[\alpha(\theta_T - \nu_T)(a_1 + 2a_2 l_T) | x_{iT}\right].\end{aligned} \tag{2}$$

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

Proposition 1. *Assume σ_ϵ^2 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 at least as high as a threshold $x_T^*(\sigma_\epsilon^2)$. Moreover, as $\sigma_\epsilon^2 \to 0$, we have $x_T^*(\sigma_\epsilon^2) \to x_T^*$, where*

$$x_T^* = \frac{c}{\alpha(a_1 + a_2)} + \nu_T.$$

Two observations are in order. First, when $\alpha = 1$, the agents are fully altruistic, and the equilibrium threshold, $\frac{c}{a_1+a_2} + \nu_t$, coincides with the threshold that would be chosen by a social planner who wants to maximize the agents' expected utility. Indeed, because the agents play a supermodular coordination game, it is always optimal for the social planner to either have everyone attack (yielding a per-agent hedonic payoff $-c + (a_0 + a_1 + a_2)(\theta_T - \nu_T) + \nu_T$) or have no one attack (yielding $a_0(\theta_T - \nu_T) + \nu_T$); the break-even point is when $\theta_T = \frac{c}{a_1+a_2} + \nu_T$.

Second, consider an alternative model in which the agents have full information. As usual in coordination games with multiple equilibria, there is a range of parameter values $[\theta_*, \theta^*]$ such that, if θ_T lies in this interval, there are multiple equilibria; if it is higher than θ^*, then all agents must attack; if lower than θ_*, then no one attacks. It can be shown that $[\theta_*, \theta^*] = \left[\frac{c}{\alpha(a_1+2a_2)} + \nu_T, \frac{c}{\alpha a_1} + \nu_T\right]$.

It follows that the equilibrium threshold in our game, $\frac{c}{\alpha(a_1+a_2)} + \nu_T$, implies a strictly higher propensity to protest than the most peaceful equilibrium, but a strictly *lower* propensity to protest than either the most combative equilibrium *or* the social planner's solution. This result is a typical prediction in the global games literature.

A further implication, which is key to our main results, is that the agents' utility

[10] Edlin et al. (2007); Fowler and Kam (2007); Rotemberg (2009); Loewen (2010) offer similar arguments in the context of large elections.

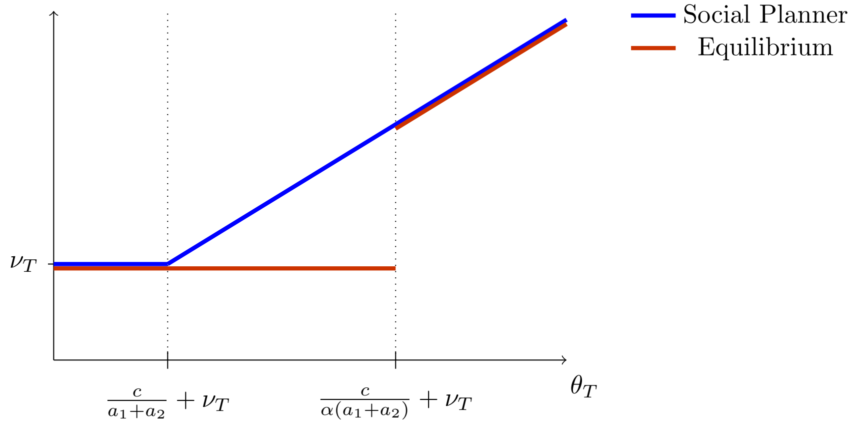

Figure 1: Utility in equilibrium and in the social planner's solution, assuming $f(0) = 0$

in the social planner's solution is a continuous function of the state θ_T, whereas their utility in the equilibrium of our model is a discontinuous function of θ_T: when the state crosses the threshold $\frac{c}{\alpha(a_1+a_2)} + \nu_T$, the agents' equilibrium payoff jumps upwards, since for values of θ in this neighborhood it was already strictly optimal to have everyone protest—just not implementable in equilibrium. This is illustrated in Figure 1. The inability of agents to coordinate on an attack—despite knowing that they would all be better off if they could—results in a lower payoff than the social planner's solution.

Next, we extend the previous arguments to obtain an equilibrium characterization for the full game. For this purpose, denote by $\overline{U}_{t+1}(\sigma_\theta^2, \sigma_\epsilon^2)$ the expected value of any agent's hedonic continuation payoffs from period $t+1$ on, with the expectation taken conditional on the regime having survived until the end of period t, but before any information is revealed about θ_{t+1}.

Proposition 2. *Assume σ_ϵ^2 is small enough. Then the game has a unique equilibrium. In this equilibrium, each player i attacks in period t if and only if x_{it} is at least as high as a threshold $x_t^*(\sigma_\theta^2, \sigma_\epsilon^2)$. Moreover, as $\sigma_\epsilon^2 \to 0$, we have $x_t^*(\sigma_\epsilon^2, \sigma_\theta^2) \to x_t^*(\sigma_\theta^2)$, $\overline{U}_{t+1}(\sigma_\theta^2, \sigma_\epsilon^2) \to \overline{U}_{t+1}(\sigma_\theta^2)$. And as $\sigma_\theta^2 \to 0$, we have $x_t^*(\sigma_\theta^2) \to x_t^*$, $\overline{U}_{t+1}(\sigma_\theta^2) \to \overline{U}_{t+1}$. The sequence of thresholds and continuation utilities $(x_0^*, \ldots, x_T^*; \overline{U}_0, \ldots, \overline{U}_T)$ can be* *found by recursively solving the following system of equations for $t = T, T-1, \ldots, 0$:*

$$x_t^* = \frac{c}{\alpha(a_1 + a_2)} + \nu_t + \delta \overline{U}_{t+1}; \tag{3}$$

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

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

Thus, in equilibrium citizens may attack in periods when the net current payoff of regime change, $\theta_t - \nu_t$, is believed to be high. But profitability of regime change, while necessary, is not sufficient for an attack to occur. When deciding whether to attack in a given period, citizens also take into account that a successful attack in that period would preclude another, possibly more profitable attack in the future, as reflected by the fact that if the continuation utility from staying in the game, $\overline{U}_{t+1}$, increases, so does today's attacking threshold, x_t^*.

In the social planner's solution, a higher continuation payoff always results in a weakly higher payoff; the social planner chooses to wait only when it is the best option. But because the agents in our model cannot coordinate on an attack, a higher $\overline{U}_{t+1}$ may result in a *lower* equilibrium payoff. In other words, changes to the environment which slightly increase the agents' payoffs given any strategy profile—but discourage them from protesting on net—may leave them worse off in equilibrium.

To illustrate, suppose that $f(l) = 0.5l + 0.5l^2$ so that in particular $f(0) = 0$ and $f(1) = 1$, i.e., the regime survives for sure if no one protests, and falls for sure if everyone protests. Suppose, moreover, that $\mu_t \equiv \mu$ and $\nu_t \equiv 0$ are constant, and σ_ϵ^2, σ_θ^2 are both very small, so that θ_t will likely be close to μ in every period. In this setting, there is no reason for the agents to wait for a "better" moment to attack (i.e., to wait for a period with higher θ_t); the socially optimal strategy would be to attack and overthrow the regime immediately if $\mu > c$—yielding a payoff of $\mu - c$ per agent—or never do so if $\mu < c$—yielding 0. Yet, in equilibrium, not only are agents less likely to attack in each period (as x_t^* is always at least $\frac{c}{\alpha} > c$), but they condition their actions today on expected future attacks. For instance, assuming $\mu > \frac{c}{\alpha} = x_T^*$, there will be an attack in period T. But then $x_{T-1}^* = \frac{c}{\alpha} + \delta(\mu - c)$, which exceeds μ if $\alpha < 1$ and δ is close enough to 1: knowing that the regime will be overthrown tomorrow, the agents procrastinate today. Procrastination is individually rational but leads to a lower payoff of $\delta(\mu - c)$ for everyone, as waiting until tomorrow results in the agents' payoffs being discounted with no improvement in the fundamentals. By the same logic, $x^*_{T-k} = \frac{c}{\alpha} + \delta^k(\mu - c)$ for all k until a value k_0 is reached for which $x^*_{T-k_0} < \mu$. At $t = T - k_0$, then, the agents attack, knowing that a failure to overthrow the regime today would lead to a lengthy wait for another (endogenous) opportunity, and the process repeats for lower values of t. The same logic is illustrated in Figure 2, assuming that $f = \frac{l+l^2}{4}$, $T = 5$, $\mu = 3$, $c = \alpha = 0.1$ and $\delta = 0.8$. In equilibrium, there are attacks in periods 0, 3 and 5, each with a 50% chance of success, whereas the social planner's solution would have the citizens attack in every period.

These examples illustrate the forward-looking logic of *intermittent attacks* in our model: because an expectation of an imminent attack reduces incentives to attack today, if the profitability of attacks is in an intermediate region, attacks must arrive in waves separated by periods of apparent calm, even if the underlying fundamentals—the level of discontent or grievances, the state of the economy, and so on—remain stable. (Of course, when attacks are very profitable, the agents will indeed attack regardless of future expectations, and when they are unprofitable enough, no one will attack.) The following corollary provides a general version of this argument.

Corollary 1. *Suppose the status quo payoff ν_t is equal to ν for all periods $t < T$, with $\nu_T = \frac{\nu}{1-\delta}$.*[11] *Then there are thresholds $\mu_0 < \mu_* < \mu^*$ such that, for $\sigma^2_\epsilon << \sigma^2_\theta$ small enough:*

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

(ii) If $\mu_t = \mu < \mu_0$ for all t, then, in every period, almost no one attacks.

(iii) Generically,[12] *if there is $\eta > 0$ for which $\mu_t \in (\mu_* + \eta, \mu^* - \eta)$ for all t, there are intermittent attacks: if T is large enough, there are periods in which most players attack and periods in which almost no one attacks. (If $T = \infty$, there are infinitely many periods of both types.)*

[11] This is equivalent to assuming status quo payoffs of size ν for period T and all periods thereafter, as noted in Footnote 6.

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

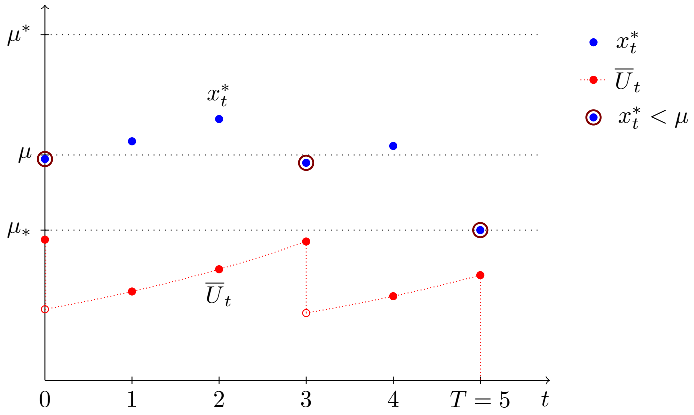

Figure 2: Pattern of attacks when μ_t lies between μ_* and μ^*

Moreover

$$\begin{aligned}
\mu_0 &= \frac{c}{\alpha(a_1+a_2)} + \frac{\nu}{1-\delta},\\
\mu_* &= \frac{c}{\alpha(a_1+a_2)} + \frac{\delta c}{1-\delta}\frac{a_0}{\alpha(a_1+a_2)} + \frac{\nu}{1-\delta},\\
\mu^* &= \frac{c}{\alpha(a_1+a_2)} + \frac{\delta c}{1-\delta}\left[\frac{a_0+a_1+a_2}{\alpha(a_1+a_2)} - 1\right] + \frac{\nu}{1-\delta}.
\end{aligned}$$

Finally, Proposition 3 characterizes the comparative statics of the model in a limited sense. 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 $t \leq t'$, as 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. 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 has no effect on it: $\frac{\partial x_t^*}{\partial \mu_{t'}} > 0$ for all $t' > t$ and $\frac{\partial x_t^*}{\partial \mu_t} = 0$.

Explicit formulas for the derivatives $\frac{\partial x_t^*}{\nu_{t'}}$, $\frac{\partial x_t^*}{\mu_{t'}}$ can be found in the Appendix. The intuition behind the result is as follows: when the status quo payoff, $\nu_{t'}$, or the regime change payoff, $\mu_{t'}$, increases in some future period $t' > t$, it becomes more attractive to let the regime survive at time t, for a chance to receive this increased payoff at time t'.

As a result, 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 in equilibrium (when information is precise) players attack at time t whenever $\mu_t > x_t^*$. The general message, then, is that a more attractive status quo always deters attacks, while a more attractive payoff from regime change in a given period encourages attacks *then* while discouraging attacks *in previous periods*.

When information is precise, the results in Proposition 3 characterize only *latent* changes in the willingness to attack: for example, if $\mu_t < x_t^*$, then there will not be an attack at time t, a conclusion left unaffected by any marginal change to the parameters. If a change to the parameters is large enough, collective behavior will eventually change discontinuously, and perhaps simultaneously in multiple periods. For instance, as we increase $\nu_{t'}$, all the thresholds x_t^* for $t < t'$ will smoothly increase, up to the point when one of them finally crosses μ_t from below. At that point, the players would suddenly switch from attacking the regime at time t to not doing so, and this expectation may in turn galvanize the players to attack in a previous period t'', etc.

5 Extensions

In this Section we briefly present several extensions of the model. The first one shows how the model may accommodate more general forms of uncertainty or informational shocks that the citizens may face. The second one shows how the results change if private benefits (i.e., "club goods") are present in the model in addition to public benefits. The third one shows how to solve an alternative model in which, in contrast to the situation in our main model, the protest movement has no hope of overthrowing the regime, but protests only to keep a resistance alive and stave off a state of permanent repression.

5.1 Generalized Uncertainty

While we assumed, for simplicity and consistency with the literature, that the citizens face idiosyncratic uncertainty about their payoff from regime change, θ_t, we would have obtained similar results had we instead assumed them to face the same kind of uncertainty regarding their status quo payoff if the regime survives, ν_t. In particular, it would still be the case that, as information becomes more precise, most citizens would attack in period t if

$$\mu_t > \frac{c}{\alpha(a_1 + a_2)} + \nu_t + \delta \overline{U}_{t+1},$$

and most would refrain from attacking if the reverse strict inequality holds (with ν_t now denoting the *expected* status quo payoff). Qualitatively similar results are obtained if we instead assume uncertain and time-varying costs of protesting, c, or impacts of protest behavior, (a_0, a_1, a_2).

Perhaps more importantly, we can allow for very general uncertainty and learning about *future* payoff parameters. For example, we can assume that, for each t, 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 even gradually over many periods, with all such signals being revealed equally to all citizens. Because this uncertainty is resolved by time t, it makes no difference for the purpose of characterizing the citizens' equilibrium strategies at time t. The only change to our analysis is that, when calculating a continuation value $\overline{U}_{t+1}$, we must write a more complicated version of Equation (4) taking into account that the parameters μ_{t+1} and ν_{t+1}, the players' equilibrium actions, and the next period's continuation value, $\overline{U}_{t+2}$, may take many possible values that we must take an expectation over.

Adding this kind of uncertainty to the model allows us to think about the equilibrium response to information about future shocks. For example, let $f(l) = \frac{2l+l^2}{8}$, $c = 1$, $\delta = 0.8$, and $\alpha = \frac{4}{33}$. Assume that $\mu_t \equiv 1$, but 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 concentrate power in the hands of the incumbent, and red to after the bill has been ratified. Alternatively, we can think of these colors as capturing different levels of social strife, where green is peaceful, yellow is tension, and red corresponds to conflict. Either way, while the state is green or yellow, $\nu_t = 0$, whereas $\nu_t = \underline{\nu} < 0$ in the red state. If the state is green at time t, then, at time $t+1$, it will still be green with probability 0.98, or it will turn yellow with probability 0.02. If the state turns yellow in period t, it remains in this state for three periods (t,

$t+1$, $t+2$) and then becomes red forever. Note that the yellow state is not materially any worse than the green one—it just denotes that citizens are *aware* of an imminent slide to the red state.

Suppose that the state turns yellow in period t_0. Using Equations (3) and (4), we can show that citizens will attack in every red period (i.e., from $t_0 + 3$ onwards) if $\underline{\nu} < -10$. Moreover, if $\underline{\nu} < -20$, citizens will also attack in the last yellow period, t_0+2; if $\underline{\nu} < -40$, they will also attack in period t_0+1; and if $\underline{\nu} < -80$, they will also attack in period t_0, that is, as soon as the state changes to yellow. Of course, citizens cannot preemptively attack in period $t_0 - 1$ because they do not know when the state will turn yellow until they see it; and they will not attack in the green state so long as $\underline{\nu} > -1800$. A crucial assumption underpinning this example is that $f(1) < 1$ (more specifically, $f(1) = 0.375$): because even an all-out attack is not guaranteed to topple the regime, and the red state is very costly, the citizens would do well to begin attacking ahead of time if they are on a path to the red state. The more costly this state is, the earlier they begin to attack. Only when $\underline{\nu}$ is extremely negative (in particular, $\underline{\nu} < -1800$), the citizens attack even in the *green* state, in an attempt to dodge a future red state that may not ever materialize. A general lesson from this example is that, in our model, the citizens may respond proactively to a threat that the status quo will worsen in the future, or that opportunities to protest may disappear in the future.

5.2 Private and Public Benefits

For simplicity, our main model assumes that there are *only* public benefits from protesting, that is, any payoff from regime change benefits all citizens. We could instead allow for the coexistence of public and private benefits. Suppose, as before, that all players receive θ_t if the regime is toppled at time t, but in this case, protesters receive an additional payoff $\rho_t > 0$ which is commonly known. Then i's marginal payoff from protesting at time t would become

$$\Delta_{it} = -c + E\left[\alpha(\theta_t + l_t\rho_t - \nu_t - \delta\overline{U}_{t+1})f'(l_t) + \rho_t f(l_t)|x_{it}\right], \tag{5}$$

where $\rho_t f(l_t)$ is the expected private benefit received by i, and $\alpha l_t \rho_t f'(l_t)$ is i's valuation of the private benefits that i's participation enables *other* protesters to receive. It can be shown that, under the same conditions as in the main model, the game remains one of strategic complements, so the citizens attack iff x_{it} is above a threshold, converging to a limit x_t^* when σ_ϵ^2 is small enough, specifically:

$$x_t^* = \frac{c}{\alpha(a_1+a_2)} + \nu_t + \delta \overline{U}_{t+1} - \rho_t \left[\frac{\frac{a_1}{2} + \frac{2a_2}{3}}{a_1+a_2} + \frac{a_0 + \frac{a_1}{2} + \frac{a_2}{3}}{\alpha(a_1+a_2)} \right]. \tag{6}$$

A derivation of Equation (6) can be found in the Appendix.

5.3 A Model of Fighting to Survive

Consider a variant of the model with the following properties. 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 and the game ends. (Of course, ν_t can represent a discounted sum of payoffs.) Demonstrating still costs c and we make the same assumptions as before regarding altruism. The probability that the movement survives period t is $f(l_t) = a_0 + a_1 l_t + a_2 l_t^2$.

Then the net payoff of demonstrating for the marginal agent is

$$-c + E\left[\alpha(\theta_t + \delta\overline{U}_{t+1} - \nu_t) | x_{it} = x_t^*(\sigma_\epsilon^2)\right],$$

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 $\sigma_\epsilon^2 \to 0$ is now

$$x_t^* = \frac{c}{\alpha(a_1+a_2)} + \nu_t - \delta\overline{U}_{t+1}. \tag{7}$$

As in the main model, agents are reluctant to protest relative to the social planner's solution (because they don't fully internalize the benefits), which means that a marginal change in the future parameters which shifts the equilibrium from not attacking to attacking in a future period will discontinuously increase the players' payoffs. But, in this variant of the model, such an increase in continuation utilities will actually **encourage** more protests today, since the citizens are more likely to accrue that higher continuation utility precisely if they do protest today. (Mechanically, this appears in Equation (7) 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 attacks, there is a single threshold $\mu^* = \frac{c}{\alpha(a_1+a_2)}$ 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.

6 Discussion

Our model of protests fueled by altruistic motives makes several concrete predictions that differ markedly from those made by existing models of protests. We now discuss these contrasts is more detail, both in abstract terms and in the context of contemporary protest movements.

6.1 Theoretical Implications

Some key takeaways of our analysis are the following. In our model, citizens protest only when conditions are quite favorable: their threshold for attacking is higher than the social planner's. Information about the future—about chances to take action, what will happen if nothing is done, and so on—affects current behavior: the bleaker the future looks, the more likely protests are today. Because an expectation of future attacks undermines current incentives to protest, intermittent attacks are likely.

As noted in the Introduction, the canonical models of protests as coordination games (Morris and Shin, 2004; Angeletos et al., 2007; Little, 2017) assume that the players are driven by *private benefits* (i.e., payoffs that are only obtained by participants in a successful attack), and that there are no altruistic concerns (i.e., $\alpha = 0$). It is pedagogically useful to provide a model with these key features within our framework for comparison. (The conclusions we draw from our "selfish" model match those of the canonical models in the literature.)

To this end, set $\alpha = 0$ in Equation (5), yielding $\Delta_{it} = -c + E[\rho_t f(l_t)|x_{it}]$, and assume idiosyncratic uncertainty about the value of ρ_t, e.g., $\rho_t \sim N(\overline{\rho}_t, \sigma^2_\rho)$ and $x_{it} = \rho_t + \epsilon_{it}$. When information is precise (as $\sigma^2_\epsilon \to 0$), this leads to a limit cutoff of the form

$$x^*_t = \frac{c}{a_0 + \frac{a_1}{2} + \frac{a_2}{3}}.$$

Note that the continuation utility, $\overline{U}_{t+1}$, is nowhere present in this expression: agents driven by private motives ought to act *as if* myopic, even when they are forward-looking. The reason is that selfish, atomistic agents must disregard the negligible chance that their actions will change the outcome of a collective endeavor. As a result, information about the future becomes irrelevant in the selfish model of protests.

Although selfish protesters are less motivated to act than a social planner would like in a static setting, they may be inefficiently slow *or* quick to act in a dynamic setting, precisely because they disregard the future in their calculations. For example, in the context of an improving environment (μ_t, ν_t increasing over time) selfish citizens might unnecessarily "jump the gun," chasing a short-term payoff—while they might fail to react to an approaching catastrophe (μ_t, ν_t sharply decreasing) if current conditions to attack are not tempting enough.

Finally, our simple "selfish model" contains no mechanism leading to intermittent attacks (for example if $\overline{\rho}_t$ is constant over time). Existing models (Angeletos et al., 2007; Little, 2017) recover this possibility by assuming that the state of the game is somewhat persistent (i.e., a higher-than-expected ρ_t is a signal of higher $\rho_{t'}$ in the future) but not directly observed, even after the fact. In that context, the linkage across periods is that after a failed attack, citizens know that the regime is strong enough to have survived. This negative signal discourages further attacks until enough uncertainty re-accumulates about the regime's changing strength. In other words, the logic behind waves of protest is informational and backward-looking, in contrast to the strategic, forward-looking logic of our model.

6.2 Examples

We can now compare how either framework would conceptualize the key events of some recent protest movements. Let us begin with the Hong Kong protests, sparked by the February 2019 introduction of a proposed bill that would have allowed extraditions to mainland China.[13] In June, when the bill would have been discussed at the Legislative Council (Purbrick, 2019), the protests peaked, leading to clashes with police. Further protests followed, now against the bill, the police crackdowns and the government's condemnation of the protests as riots. The bill was then suspended

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

indefinitely.[14] Crowds peaked at as many as two million participants. Pro-democracy candidates, previously a minority, captured over 80% of District Council seats at the November 2019 local elections.[15] Protesters described being motivated by a sense that Hong Kong faced a do-or-die fight for its future, saying to the media: "If we don't succeed now, our freedom of speech, our human rights, all will be gone."[16] Newspapers struck a similar tone, with headlines such as "The End of Hong Kong is Almost Here."[17] The conflict echoed massive protests in 2003 in response to a proposed national security bill, as well as the 2014 Umbrella Revolution, which condemned a proposal to implement democratic elections but only between candidates selected by a pro-Beijing committee. These explosions of dissent punctuated a rising collective unease with the mainland's attempts to encroach on Hong Kong's autonomy, described as "the political ground simultaneously shifting and shrinking beneath their feet," all this against the backdrop of a ticking clock, as the terms of the 1997 Sino-British Joint Declaration that delineates the "one country, two systems" framework would formally expire in 2047.[18] Finally, in June of 2020, the mainland National People's Congress, bypassing the local government, imposed a national security law which criminalized dissent in Hong Kong. Described by a Beijing official as "a sharp sword hanging over a minority of people who endanger national security,"[19] the law had an immediate chilling effect on protests and led to the disbandment of pro-democracy parties, raids on media offices, and mass arrests of activists and primary organizers.[20]

It is worth highlighting three key facts. First, the 2003, 2014, and 2019 protests all began in response to *proposed* bills or reforms, which had not yet had any material consequences but which could be taken as *signs* that the future and autonomy of Hong Kong were quickly deteriorating. In other words, current parameters θ_t, ν_t, c_t, f_t were not affected by the proposed bills; what fell was the perceived continuation value $\delta \overline{U}_{t+1}$ offered by the status quo. Thus, citizens demonstrated forward-looking protest participation. (A similar logic is encapsulated in the chilling slogan used by

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

[15] https://www.nytimes.com/2019/11/24/world/asia/hong-kong-election-results.html

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

[17] https://foreignpolicy.com/2019/05/16/the-end-of-hong-kong-is-almost-here/

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

[19] https://www.scmp.com/news/hong-kong/politics/article/3091241/national-security-law-chinese-president-xi-jinping-signs

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

Taiwanese protesters: “Today’s Hong Kong, tomorrow’s Taiwan.”[21]) Second, even though all citizens had common knowledge that the promise of “one country, two systems” would expire in 2047, it took concrete threats that the system would be subverted ahead of schedule, and imminently, to spur them to act. That is, citizens arguably showed signs of collective procrastination, as their strongest attempts at extracting concessions through mobilization took place when the government had already shifted towards a hard-line approach, with Xi Jinping having made bold moves to centralize authority and minimize internal dissent in China throughout the 2010s. Third, protests *grew* in response to police brutality, which could signal that c was higher than expected, but also be a further sign that the worst was yet to come, and so the time to act was now.

Thus, through the lens of our model, we can see the protests as an increasingly desperate resistance which responded to future threats but only when the prospect of disaster became imminent. In contrast, to explain the 2019 surge in protest behavior, models of citizens driven by private benefits would need to allege an increase in private benefits from success, ρ_t; a decrease in the cost of protesting, c_t; or a weakening of the regime (f_t shifting upwards) which made it a tempting target. While private benefits may well explain the behavior of leaders and activists, it cannot reasonably explain the participation of millions of people, and the other explanations run counter to the facts (as the protesters faced a regime that had dug in its heels and was willing to respond with violence). Finally, though in a coordination game some idiosyncratic event could always trigger collective action by shifting the players’ “focal point,” this explanation would chalk the consistent response to threatening legislation up to coincidence. In particular, in these models, there is no avenue for continuation utilities to play a role, *even if* individual citizens are rational and forward-looking. Finally, while one may plausibly explain the observed protest behavior as a pure emotional response to grievances (Passarelli and Tabellini, 2017; Gibilisco, 2021; Correa, Nandong and Shadmehr, 2021), this explanation is incomplete without a model of *why* certain events aggrieve people.

The 2014 Euromaidan revolution in Ukraine is another recent example of an impactful protest movement that we can examine through the lens of our model. After years of negotiations with the European Union and promises of European integration, the Yanukovych administration announced in November of 2013 that it was suspend-

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

ing plans to sign a broad political and economic association agreement with the EU, only a week before the scheduled signing. Instead, Ukraine would seek closer ties with Russia, which had threatened retaliatory trade sanctions in response to the EU deal.[22] Protesters gathered the same day, angry that their hopes to finally escape the Russian sphere of influence—to no longer live in "a post-Soviet barrack temporarily repainted in yellow and blue"—were quickly vanishing.[23] Ukraine failed to sign the EU agreement as scheduled, even as both sides claimed that a deal was still on the table.[24] The protests grew in number and scope of demands, and turned into riots after the government responded with a violent crackdown.[25] The situation worsened further after the government passed a package of draconian anti-protest laws in January,[26] and reached its nadir in February, with over 100 protesters being killed by police amid escalating clashes. Soon, widespread desertion among demoralized police forces forced Yanukovych to flee to Russia.[27] Again, a pattern emerges of citizens protesting in response to an expectation of a worsening future (or a dashed hope of improvement), after years of inaction despite a bleak outlook, and strengthening their resolve in the face of violence and draconian measures signaling a turn towards dictatorship.

A similar set of circumstances spurred the Gezi protests in Turkey in the summer of 2013. The increasingly authoritarian style of Erdogan drove millions into the streets out of concern that if unchecked, the country would soon be overtaken by the "creeping political authoritarianism" (Özel, 2014). The government had already curbed the powers of two of the only remaining checks on its power: the constitutional court (Özbudun, 2014) and the media (Kocak and Kıbrıs, 2022). In the period leading up to Gezi, the government restricted the sale of alcohol and passed other bills limiting various freedoms of people living in Turkey—corresponding to a slight decline in

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

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

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

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

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

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

contemporaneous ν_t (Özbudun, 2014). But what was a bigger influence on many protesters' decisions to turn out was the shifting style of Erdogan and his government and the sense that the worst was yet to come—a large decline in $\delta\overline{U}_{t+1}$. Before the Gezi protests, a government agency dropped the phrase "Republic of Turkey," which was interpreted as paving the way for a move away from democracy. Such fears were compounded when some Republic Day celebrations were banned. In an Islamist turn that accompanied the alleged democratic backslide, artists and public intellectuals were sentenced to jail for "insulting the religious beliefs" of the society, and members of the governing party decried public displays of affection and called for the creation of single-sex beaches. Another symbolic move by the government was naming a public project after Selim the Grim, a move broadly interpreted as an insult to Turkey's large Alevi minority because of the historical massacre during his rule (Yörük, 2014). Erdogan's language also took an increasingly exclusionary and dismissive tone towards anyone who wasn't Sunni and pious. He variously called anybody who did not fit his image of an ideal citizen "marginals," "thugs," "looters," or "drunkards" (Göle, 2013). His increasingly aggressive and insulting words made it clear to everybody else that soon they would not feel welcome in their own country.

In response to fears that civic freedoms would continue to shrink at an increasing rate if the government went unchallenged, activists started to protest more often and in greater numbers even before Gezi. They were met with increasing use of excessive force by the police—an increase in c_t. Things came to a head on May 28, 2013 when a few dozen peaceful protesters organizing a sit-in at Gezi Park were cleared with tear gas and pepper spray. This was a wake up call for others in Turkey that one of the last remaining outlets for political expression was all but taken away. What followed were the largest anti-government protests in the history of the Turkish Republic. Thus, Gezi was in large part prospective—protesters were responding not only to what the regime had done, but also to what it signaled it would do. Ultimately, these protests failed to change the regime yet succeeded in delaying—but not preventing—Turkey's democratic backslide.

7 Conclusions

In this paper we developed a dynamic model of protests in which citizens are partially altruistic, and hence may choose to act even if the benefits from regime change accrue to all citizens, including non-participants. Our model shows that altruistic citizens—unlike selfish ones—face pivotality concerns: they care about the (minuscule) probability that their participation will change the outcome, not just the probability that they will get to share in the spoils. In a dynamic context, this implies that their willingness to protest responds not just to contemporaneous benefits and costs, but also to the future ramifications of regime change or its absence. In particular, altruistic citizens may protest in response to an event that increases the cost of protesting if it also paints a bleak picture of the future, as is the case with police crackdowns or authoritarian measures. Because an expectation of future collective action makes present collective action less urgent, and vice versa, spikes in social turmoil are self-limiting and may arrive in waves, even if the underlying material and social conditions are stable over time. Moreover, because *partially* altruistic citizens act only when collective action is socially beneficial by a wide enough margin, the mere existence of future chances to act may tempt them to drag their feet today, leaving them worse off. However, if the goal of collective action is just to keep a movement alive, then future action encourages present action, leading to a pattern of either sustained action or disintegration of the movement.

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 novel predictions which, in our view, translate into more natural conceptualizations of many protest processes. (Though we discussed three prominent examples, the emergence of protests in response to negative expectations and state violence appears to be general phenomenon.) In assuming limited altruism, our aim is to build a parsimonious model that keeps *ad hoc* assumptions to a minimum while capturing, in some form, notions such as public-mindedness, grievances, and other moral considerations that undeniably play a role in civic behavior.

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

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

Another possible avenue for future work is to translate our partial altruism framework to model other examples of civic behavior, such as voting, deliberation, and campaigning activity. In these domains, there is also a tension between observations that citizens are more engaged than rational models would predict, yet still strategic (in the sense that, for example, some voters will stop supporting a candidate with no chance to win) in a way that defies purely emotional or expressive explanations. Our modeling approach provides a potential way to bridge this gap.

References

Acemoglu, Daron and James A Robinson, "A Theory of Political Transitions," *American Economic Review*, 2001, *91* (4), 938–963.

Angeletos, George-Marios, Christian Hellwig, and Alessandro Pavan, "Dynamic global games of regime change: Learning, multiplicity, and the timing of attacks," *Econometrica*, 2007, *75* (3), 711–756.

Barbera, Salvador and Matthew O. Jackson, "A Model of Protests, Revolution, and Information," *Quarterly Journal of Political Science*, 2020, *15* (3), 297–335.

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

Buchanan, James M, "An Economic Theory of Clubs," *Economica*, 1965, *32* (125), 1–14.

Bueno De Mesquita, Ethan, "Regime Change and Revolutionary Entrepreneurs," *American Political Science Review*, 2010, *104* (3), 446–466.

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.

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

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

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

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

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

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

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

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

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

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

Fowler, James H. and Cindy D. Kam, "Beyond the Self: Social Identity, Altruism, and Political Participation," *The Journal of Politics*, 2007, *69* (3), 813–827.

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

Göle, Nilüfer, "Gezi-anatomy of a public square movement," *Insight Turkey*, 2013, *15* (3), 7.

Jankowski, Richard, "Buying a Lottery Ticket to Help the Poor: Altruism, Civic Duty, and Self-interest in the Decision to Vote," *Rationality and Society*, 2002, *14* (1), 55–77.

Kocak, Korhan and Özgür Kıbrıs, "Social Media and Press Freedom," *British Journal of Political Science*, 2022, pp. 1–23.

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

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

Loewen, Peter John, "Affinity, Antipathy and Political Participation: How Our Concern For Others Makes Us Vote," *Canadian Journal of Political Science*, 2010, *43* (3), 661687.

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** _ , "Rethinking multiple equilibria in macroeconomic modeling," in "NBER Macroeconomics Annual 2000, Volume 15," MIT Press, 2001, pp. 139–182.

_ **and** _ , "Coordination risk and the price of debt," *European Economic Review*, 2004, *48* (1), 133–153.

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

Özbudun, Ergun, "AKP at the crossroads: Erdoğan's majoritarian drift," *South European Society and Politics*, 2014, *19* (2), 155–167.

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

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

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

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

Rotemberg, Julio J, "Attitude-dependent altruism, turnout and voting," *Public Choice*, 2009, *140* (1), 223–244.

Shadmehr, Mehdi, "Mobilization, repression, and revolution: Grievances and opportunities in contentious politics," *The Journal of Politics*, 2014, *76* (3), 621–635.

_ , "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.

Yörük, Erdem, "The long summer of Turkey: The Gezi uprising and its historical roots," *South Atlantic Quarterly*, 2014, *113* (2), 419–426.

A Appendix

Proof of Proposition 1. First, we note that the game is supermodular in actions, i.e., if the other players' strategies change so that they are more likely to attack than before, then the incentive to attack of any player i increases. More formally, denote j's strategy in period T by A_{jT}, the set of realizations of x_{jT} for which j attacks. Let $(A_{jT})_{j\in[0,1]}$, $(A'_{jT})_{j\in[0,1]}$ be two strategy profiles such that $A_{jT} \subseteq A'_{jT}$ for all j. Let l_T, l'_T be the equilibrium proportion of attackers under each strategy profile (these are, of course, random variables). Then it is clear that $l_T \leq l'_T$. It follows immediately from Equation 2 that $\Delta_{iT}(x_{iT}) \leq \Delta'_{iT}(x_{iT})$ for all i, x_{iT}.

Supermodularity in actions implies the existence of a greatest equilibrium and a smallest equilibrium (Milgrom and Roberts, 1990). Second, we will show that both of these equilibria are symmetric and in threshold strategies. 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), it is sufficient to 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 $\Delta_{iT}(x, x', \sigma^2)$ 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 $\sigma_\epsilon^2 = \sigma^2$. Then we want to show that $\Delta_{iT}(x, x', \sigma^2)$ is strictly increasing in x for all x', σ^2. Note that f' is strictly increasing, and l_T is a strictly increasing (and deterministic) function of θ_T, as in fact

$$x_{jT} \geq x' \iff \frac{\epsilon_{jT}}{\sigma_\epsilon} \geq \frac{x' - \theta_T}{\sigma_\epsilon},$$

implying $l_T = \Phi\left(\frac{\theta_T - x'}{\sigma_\epsilon}\right)$. Then $\alpha(\theta_T - \nu_T)f'(l_T)$ is a deterministic, strictly increasing function of θ_T. Then the claim follows from the fact that the conditional distribution of θ_T given x_{iT} is strictly FOSD-increasing in x_{iT}, as $\theta_T | x_{iT} \sim N\left(\frac{\sigma_\epsilon^2 \mu_T + \sigma_\theta^2 x_{iT}}{\sigma_\theta^2 + \sigma_\epsilon^2}, \frac{\sigma_\theta^2 \sigma_\epsilon^2}{\sigma_\theta^2 + \sigma_\epsilon^2}\right)$ (see DeGroot, 1970, Theorem 9.5.1).

In addition, note that $\alpha(\theta_T - \nu_T)f'(l_T) \in [\alpha(\theta_T - \nu_T)a_1, \alpha(\theta_T - \nu_T)(a_1 + 2a_2)]$, so any x which is a best-response threshold for i to some strategy profile for the other players must satisfy $E(\theta_T | x_{iT} = x) - \nu_T \in \left[\frac{c}{\alpha(a_1+2a_2)}, \frac{c}{\alpha a_1}\right]$; hence any such x must lie in a compact interval I, namely

$$\left[\left(\frac{c}{\alpha(a_1+2a_2)}+\nu_T\right)\left(1+\frac{\sigma_\epsilon^2}{\sigma_\theta^2}\right)-\frac{\sigma_\epsilon^2}{\sigma_\theta^2}\mu_T, \left(\frac{c}{\alpha a_1}+\nu_T\right)\left(1+\frac{\sigma_\epsilon^2}{\sigma_\theta^2}\right)-\frac{\sigma_\epsilon^2}{\sigma_\theta^2}\mu_T\right].$$

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^2 small enough, $\Delta_{iT}(x,x,\sigma_\epsilon^2)$ is continuous and strictly increasing in x, so there must be a unique $x^*(\sigma_\epsilon^2)$ for which $\Delta_{iT}(x,x,\sigma_\epsilon^2)=0$, as required. Denote $\tilde{\theta}=\hat{\theta}+\frac{\sigma_\theta^2}{\sigma_\theta^2+\sigma_\epsilon^2}x$, where $\hat{\theta}\sim N\left(\frac{\sigma_\epsilon^2\mu_T}{\sigma_\theta^2+\sigma_\epsilon^2},\frac{\sigma_\theta^2\sigma_\epsilon^2}{\sigma_\theta^2+\sigma_\epsilon^2}\right)$ is independent of x. Then $\tilde{\theta}$ has the same distribution as θ_T conditional on $x_{iT}=x$. We can then write

$$\Delta_{iT}(x,x,\sigma_\epsilon^2)=-c+E\left[\alpha(\tilde{\theta}-\nu_T)\left(a_1+2a_2\Phi\left(\frac{\tilde{\theta}-x}{\sigma_\epsilon}\right)\right)\right]$$

$$\frac{\partial\Delta_{iT}(x,x,\sigma_\epsilon^2)}{\partial x}=\alpha E\left[\frac{\sigma_\theta^2}{\sigma_\theta^2+\sigma_\epsilon^2}\left(a_1+2a_2\Phi\left(\frac{\tilde{\theta}-x}{\sigma_\epsilon}\right)\right)-\frac{\sigma_\epsilon}{\sigma_\theta^2+\sigma_\epsilon^2}(\tilde{\theta}-\nu_T)2a_2\phi\left(\frac{\tilde{\theta}-x}{\sigma_\epsilon}\right)\right],$$

where $\frac{\tilde{\theta}-x}{\sigma_\epsilon}=\frac{\hat{\theta}}{\sigma_\epsilon}-\frac{\sigma_\epsilon}{\sigma_\theta^2+\sigma_\epsilon^2}x\sim N\left(\frac{\sigma_\epsilon}{\sigma_\theta^2+\sigma_\epsilon^2}(\mu_T-x),\frac{\sigma_\theta^2}{\sigma_\theta^2+\sigma_\epsilon^2}\right)$. Clearly $\Delta_{iT}(x,x,\sigma_\epsilon^2)$ and $\frac{\partial\Delta_{iT}(x,x,\sigma_\epsilon^2)}{\partial x}$ are continuous on the domain $\mathbb{R}\times(0,+\infty)$, with continuous extensions to $\mathbb{R}\times[0,+\infty)$. In particular, we can extend $\frac{\partial\Delta_{iT}(x,x,\sigma_\epsilon^2)}{\partial x}$ like so:

$$\frac{\partial\Delta_{iT}(x,x,0)}{\partial x}:=\alpha E\left[a_1+2a_2\Phi\left(z\right)\right]=\alpha(a_1+a_2)>0,$$

where z is a random variable with distribution $N(0,1)$, and $E(\Phi(z))=\frac{1}{2}$ due to the symmetry of the normal distribution. As $\frac{\partial\Delta_{iT}(x,x,\sigma_\epsilon^2)}{\partial x}$ is continuous on $I\times[0,1]$, it is also uniformly continuous on this set by the Heine-Cantor theorem. Then there is $\overline{\sigma}_\epsilon^2>0$ such that, if $\sigma_\epsilon^2<\overline{\sigma}_\epsilon^2$, $|\frac{\partial\Delta_{iT}(x,x,\sigma_\epsilon^2)}{\partial x}-\frac{\partial\Delta_{iT}(x,x,0)}{\partial x}|<\alpha(a_1+a_2)$ for all $x\in I$, so that $\frac{\partial\Delta_{iT}(x,x,\sigma_\epsilon^2)}{\partial x}>0$ for $x\in I$. This implies that $\Delta_{iT}(x,x,\sigma_\epsilon^2)$ is strictly increasing, as we wanted.

Finally, to find x_T^*, we note that

$$\Delta_{iT}(x,x,0)=-c+\alpha(x-\nu_T)(a_1+a_2),$$

which vanishes only at $\frac{c}{\alpha(a_1+a_2)}+\nu_T$. By continuity, any point of accumulation of

$x^*(\sigma_\epsilon^2)$ as $\sigma_\epsilon^2 \to 0$ must solve the equation $\Delta_{iT}(x,x,0) = 0$, i.e., it must equal $\frac{c}{\alpha(a_1+a_2)} + \nu_T$. Because the image of $x^*(\sigma_\epsilon^2)$ is contained in a compact interval, this implies that $x^*(\sigma_\epsilon^2) \xrightarrow[\sigma_\epsilon^2 \to 0]{} \frac{c}{\alpha(a_1+a_2)} + \nu_T =: x_T^*$, as we wanted. □

Proof of Proposition 2. The marginal payoff from attacking in period t is given by the expression

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

in the general case. Under the assumption that f is quadratic this boils down to

$$= -c + E\left[\alpha(\theta_t - \nu_t - \delta\overline{U}_{t+1})(a_1 + 2a_2 l_t)|x_{it}\right].$$

By the same argument as in Proposition 1, for σ_ϵ^2 small enough, this game has a unique equilibrium, which is symmetric and in threshold strategies. In fact, this game is in a sense equivalent to the game from period T: if we denote $\tilde{\theta} = \theta_t - \overline{U}_{t+1} \sim N(\mu_t - \overline{U}_{t+1}, \sigma_\theta^2)$, then the stage game played in period t is equivalent to the one-shot game with $\tilde{\mu} = \mu_t - \overline{U}_{t+1}$.

To characterize the sequence of thresholds x_t^* and expected continuation payoffs $\overline{U}_t$, denote i's marginal payoff from attacking by $\Delta_{it}(x, x', \sigma_\epsilon^2, \sigma_\theta^2)$ if she sees $x_{it} = x$ and other players attack iff $x_{jt} \geq x'$. By the same argument as in Proposition 1, the function $\Delta_{it}(x, x, \sigma_\epsilon^2, \sigma_\theta^2)$, with domain $\mathbb{R} \times (0, +\infty) \times (0, +\infty)$ admits a continuous extension to $\mathbb{R} \times [0, +\infty) \times (0, +\infty)$, given by:

$$\Delta_{it}(x, x, 0, \sigma_\theta^2) = -c + \alpha(x - \nu_t - \delta\overline{U}_{t+1})(a_1 + a_2),$$

and $x_t^*(\sigma_\theta^2)$ is the unique value of x that solves the equation

$$0 = -c + \alpha(x_t^*(\sigma_\theta^2) - \nu_t - \delta\overline{U}_{t+1})(a_1 + a_2),$$

which implies Equation 3, as in fact $x_t^*(\sigma_\theta^2)$ is constantly equal to x_t^* for all σ_θ^2.

As for Equation 4, for general values of σ_ϵ^2 and σ_θ^2, let $U_t(x, \sigma_\epsilon^2, \sigma_\theta^2)$ be the expected continuation hedonic utility in equilibrium of an agent i starting at time t, conditional on seeing $x_{it} = x$, and $\overline{U}_t(\sigma_\epsilon^2, \sigma_\theta^2)$ be i's expected continuation hedonic utility before seeing x_{it} (both of which, by symmetry, are the same for all agents). Then we have

$$U_t(x)(\sigma_\epsilon^2,\sigma_\theta^2) = -c\mathbb{1}_{\{x\geq x_t^*\}} + E\left[\left(\theta_t - \nu_t - \delta\overline{U}_{t+1}(\sigma_\epsilon^2,\sigma_\theta^2)\right) f(l_t(\theta_t))|x_{it}=x\right] + \nu_t + \delta\overline{U}_{t+1}(\sigma_\epsilon^2,\sigma_\theta^2)$$

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

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

As $\sigma_\epsilon^2 \to 0$, $\overline{U}_t(\sigma_\epsilon^2,\sigma_\theta^2)$ converges to

$$\overline{U}_t(\sigma_\theta^2) = -c\Phi\left(\frac{\mu_t - x_t^*}{\sigma_\theta}\right) + E\left[\left(\theta_t - \nu_t - \delta\overline{U}_{t+1}(\sigma_\theta^2)\right) f\left(\mathbb{1}_{\{\theta_t > x_t^*\}}\right)\right] + \nu_t + \delta\overline{U}_{t+1}(\sigma_\theta^2).$$

As $\sigma_\theta^2 \to 0$, $\overline{U}_t(\sigma_\theta^2)$ converges to

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

as we wanted. Intuitively, if $\mu_t > x_t^*$ then, for σ_θ^2 and σ_ϵ^2 small enough, it is almost certain that $\theta_t > x_t^*$ and that almost all agents protest, so l_t is close to 1. Conversely, if $\mu_t < x_t^*$, then it is almost certain that $\theta_t < x_t^*$ and that almost nobody protests, so l_t is close to 0. □

Proof of Corollary 1. For part (i), assume that $\mu_t = \mu$ for all t, with $\mu < \mu_0$. Then, using Equation (3), we can calculate

$$x_T^* = \frac{c}{\alpha(a_1+a_2)} + \frac{\nu}{1-\delta}.$$

Since $\mu < \mu_0$, as σ_θ^2 goes to zero, for $\sigma_\epsilon^2(\sigma_\theta^2)$ small enough, we are in the limit equilibrium characterized in Proposition 2 in the case $\mu_t < x_t^*$, in which $\theta_t < x_t^*$ with probability going to one, and l_t converges in probability to zero. Hence

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

We can then calculate

$$x_{T-1}^* = \frac{c}{\alpha(a_1+a_2)} + \nu + \delta f(0)\mu + \delta(1-f(0))\frac{\nu}{1-\delta}.$$

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

$$\overline{U}_t = f(0)\frac{1-\delta^{T-t+1}(1-f(0))^{T-t+1}}{1-\delta(1-f(0))}\mu + \left[1 - f(0)\frac{1-\delta^{T-t+1}(1-f(0))^{T-t+1}}{1-\delta(1-f(0))}\right]\frac{\nu}{1-\delta}$$
$$x^*_{t-1} = \frac{c}{\alpha(a_1+a_2)} + \nu + \delta\overline{U}_t,$$

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

$$\mu^* \geq \frac{c}{a_0+a_1+a_2} + \frac{\nu}{1-\delta} = \frac{c}{f(1)} + \frac{\nu}{1-\delta},$$

with equality iff $\alpha = 1$ and $a_0 = 0$, and $\mu > \mu^*$, we know that $\mu - \frac{c}{f(1)} > \frac{\nu}{1-\delta}$, so $\overline{U}_{t_0} > \ldots > \overline{U}_T > \frac{\nu}{1-\delta}$ and $x^*_{t_0-1} > \ldots > x^*_T$. For most players to attack in equilibrium at time $t_0 - 1$, we need $x^*_{t_0-1} < \mu$.

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

$\mu > \lim_{t \to -\infty} x_t^*$, i.e.,

$$
\begin{aligned}
\mu &> \frac{c}{\alpha(a_1+a_2)} + \nu + \delta\frac{-c+f(1)\mu}{1-\delta(1-f(1))} + \delta\frac{(1-\delta)(1-f(1))}{1-\delta+\delta f(1)}\frac{\nu}{1-\delta} \\
\iff \frac{1-\delta}{1-\delta+\delta f(1)}\mu &> \frac{c}{\alpha(a_1+a_2)} - \frac{\delta c}{1-\delta+\delta f(1)} + \frac{\nu}{1-\delta+\delta f(1)} \\
\iff \mu &> \frac{c}{\alpha(a_1+a_2)}\left(1+\frac{\delta f(1)}{1-\delta}\right) - \frac{\delta c}{1-\delta} + \frac{\nu}{1-\delta} = \mu^*.
\end{aligned}
$$

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 2, for $\sigma_\theta^2 \to 0$ with σ_ϵ^2 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

$$
\overline{U}_t = -c + f(1)\mu_t + (1-f(1))(\nu + \delta\overline{U}_{t+1}) \tag{8}
$$

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

$$
\frac{c}{\alpha(a_1+a_2)} + \nu + \delta \liminf \overline{U}_t \leq \underline{\mu}.
$$

Equation (8) 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

$$\frac{c}{\alpha(a_1+a_2)}+\nu+\delta\frac{-c+f(1)\underline{\mu}}{1-\delta+\delta f(1)}+\delta\frac{(1-f(1))\nu}{1-\delta+\delta f(1)}\leq\underline{\mu}$$
$$\iff \mu^*\leq\underline{\mu}.$$

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

$$\overline{U}_t = f(0)\mu_t + (1-f(0))(\nu+\delta\overline{U}_{t+1}) \tag{9}$$

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

$$\frac{c}{\alpha(a_1+a_2)}+\nu+\delta\limsup\overline{U}_t\geq\overline{\mu}.$$

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

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

Then we must have

$$\frac{c}{\alpha(a_1+a_2)}+\nu+\delta\frac{f(0)\overline{\mu}}{1-\delta+\delta f(0)}+\delta\frac{(1-f(0))\nu}{1-\delta+\delta f(0)}\geq\overline{\mu}$$
$$\iff \mu_*\geq\overline{\mu}.$$

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

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

Derivation of Equation (6). By analogous arguments to those used in the proof of Proposition 1, $x_t^*(\sigma_\epsilon^2, \sigma_\theta^2)$ is the unique value of x that solves the equation

$$\begin{aligned} 0 =& \Delta_{it}(x, x, \sigma_\epsilon^2, \sigma_\theta^2) = -c + E\left[\alpha(\theta_t + l_t\rho_t - \nu_t - \delta\overline{U}_{t+1})f'(l_t) + \rho_t f(l_t) | x_{it} = x\right] \\ =& -c + \alpha E(\theta_t f'(l_t) | x_{it} = x) - \alpha(\nu_t + \delta\overline{U}_{t+1}) E(f'(l_t) | x_{it} = x) + \\ & + \alpha\rho_t E(l_t f'(l_t) | x_{it} = x) + \rho_t E(f(l_t) | x_{it} = x) \\ \xrightarrow[\sigma_\epsilon^2 \to 0]{} & -c + \alpha x(a_1 + a_2) - \alpha(\nu_t + \delta\overline{U}_{t+1})(a_1 + a_2) + \\ & + \alpha\rho_t E(a_1 l_t + 2a_2 l_t^2 | x_{it} = x) + \rho_t E(a_0 + a_1 l_t + a_2 l_t^2 | x_{it} = x) \end{aligned}$$

As shown in Proposition 1, $E(l_t | x_{it} = x_t^*(\sigma_\epsilon^2, \sigma_\theta^2))$ converges to $E(\Phi(z)) = \frac{1}{2}$, where $z \sim N(0, 1)$, as $\sigma_\epsilon^2 \to 0$. By the same arguments, we can show that $E(l_t^2 | x_{it} = x_t^*(\sigma_\epsilon^2, \sigma_\theta^2))$ converges to $E(\Phi(z)^2)$ as $\sigma_\epsilon^2 \to 0$. Moreover, $\Phi(z)$ is clearly distributed $U[0, 1]$, since $P(\Phi(z) \leq z_0) = P(z \leq \Phi^{-1}(z_0)) = \Phi(\Phi^{-1}(z_0)) = z_0$ by definition, for any $z_0 \in [0, 1]$. Hence $E(\Phi(z)^2) = \int_0^1 x^2 dx = \frac{1}{3}$. Substituting these identities into the above and rearranging yields Equation (6). □