What Drives Regimes to Manipulate Information: Criticism, Collective Action, and Coordination*

Jakub Redlicki†

30th October 2016

Abstract

Authoritarian states often manipulate information in order to prevent regime change. The model uses the global games framework in which the regime is overthrown if enough citizens attack, however, the citizens are imperfectly informed about its strength and the regime can increase noise in their private information. Inspired by empirical findings in political science (King, Pan and Roberts, 2013; 2014; 2016), this paper illuminates (i) why regimes may aim to prevent collective action rather than criticism of the state *per se*, and (ii) why they might distract their citizens rather than try to persuade them. Unlike boosting the apparent strength by adding a bias, manipulating the noise is effective even if (i) the citizens can observe the regime’s manipulative action and (ii) the regime is no better informed about its strength than they are. I show that under these conditions the regime has no incentives to increase noise for the purpose of improving the citizen’s perception of the state *per se*. At the same time, minimising the size of all collective attacks may be an effective way of preventing regime change, especially when the citizens’ private information is intrinsically imprecise. Finally, I demonstrate that if citizens can coordinate better only at the cost of impeded information aggregation, the regime’s incentives to increase noise may become stronger.

---

*I am grateful to my supervisor, Meg Meyer, for her continuous support and guidance. I also thank Avinash Dixit, Péter Esö, Paweł Gola, Sophie Hatte, Ian Jewitt, Max Kwick, Stephen Morris, Bartosz Redlicki, and Scott Tyson for helpful comments, as well as seminar participants at the University of Oxford; RES-York Symposium on Game Theory; Ce2 Workshop, Warsaw; PET 15, Luxembourg; Summer School of the Econometric Society, University of Tokyo; RES Annual Conference 2016; MPSA 2016; EEA-ESEM 2016. An earlier version of this work was circulated as “Internet Trolls and the Incentives of Authoritarian Regimes to Manipulate Information”.

†Department of Economics, University of Oxford. E-mail: jakub.redlicki@economics.ox.ac.uk.
1 Introduction

Authoritarian regimes often manipulate information available to citizens in order to improve the perception of the ruling government and to prevent protests and revolts that could ultimately result in an overthrow. They have several instruments at their disposal, with the most common being censorship, direct media control, employment of social media commentators, as well as use of political pressure and repressions to influence journalist reporting. Out of sixty-five countries assessed in the “Freedom on the Net 2015” report by the Freedom House, thirty-one engage in censorship of political, social and religious content, twenty-four employ pro-government commentators who manipulate online discussions, and forty have arrested, imprisoned or put journalists into detention for political or social content. Furthermore, the situation is not generally improving with thirty-two countries on a negative trajectory since June 2014.

The scale of information manipulation is extensive in some countries, as exemplified by as many as two million pro-regime online commentators employed in China. This large group of people came to be called the Fifty Cent Party since they were allegedly paid 50 cents (5 jiao) for every post that advances the line of the Communist Party. The Chinese regime also uses a sophisticated system of censorship, which involves both automated mechanisms and human monitors. In recent years, for instance, this apparatus was extensively used in the period surrounding the 25th anniversary of the Tiananmen Square Massacre and during the Umbrella Revolution in Hong Kong.

Regimes also distort the information available to citizens through firm control of media. For example, in Russia, not only does the state control—either directly or through proxies—all five of the major national television networks, but also national radio networks, important national newspapers, and national news agencies. In addition to this, the regime also controls more than 60 percent of the country’s estimated 45,000 regional and local newspapers and other periodicals.

---

1 The Freedom House publishes annual reports on media independence and internet freedom, “Freedom of the Press” and “Freedom on the Net”, which document the ways in which governments influence the information available to citizens across the world.
3 See King, Pan and Roberts (2016).
4 Han (2015) provides a comprehensive list of tasks performed by online commentators. For similar recent developments in Russia, see, e.g.: http://www.nytimes.com/2015/06/07/magazine/the-agency.html; http://www.theguardian.com/world/2015/apr/02/putin-kremlin-inside-russian-troll-house
On a more general level, information manipulation may take different forms. A natural one is one where the regime tries to persuade the citizens who criticise the government and its policies that the situation is better than it really is. Alternatively, regimes may also distract citizens and redirect public attention from discussions on unfavourable topics or events. Interestingly, King, Pan and Roberts (2016) provide empirical evidence on these two forms of information manipulation in the context of government-employed social media commentators in China. They show that, contrary to the common views on the subject, it is not part of the Chinese regime’s strategy to engage in arguments with skeptics and to defend the party and the government. They observe instead that these paid internet trolls do not even discuss controversial issues and mostly write posts that involve cheerleading for China, the revolutionary history of the Communist Party, and other symbols of the regime. King et al. (2016) thus infer that the regime employs online commentators to distract the public and to change the subject of discussion when needed.

Naturally, the choice of specific forms and methods of information manipulation depends on the authoritarian regime’s objectives. When analysing the Chinese censorship programme, King, Pan and Roberts (2013, 2014) have proposed two distinct theories of what constitutes the goals of the regime: the state critique theory and the theory of collective action potential. According to the former, censorship is aimed at restricting any criticism of the government and its policies, while the latter assumes that the objective is to stop the spread of information that could lead to any kind of collective action, regardless of whether or not the expression is in opposition to the state and whether or not it is related to the government’s policies.

In principle, it could be that either or both of the above theories are correct or incorrect. Remarkably, King et al. (2013, 2014) demonstrate empirical evidence that, in the context of China, the theory of collective action potential turns out to be correct, while the state critique theory does not. This suggests that the citizens’ ability to coordinate a collective action matters. One of the aims of this paper is to formally analyse how this ability affects the regime’s incentives to manipulate information.

When modelling information manipulation, one needs to remember that it is usually not entirely costless. For example, one of the main challenges of pro-government online commentators is to write favourably about the regime in such a way that other internet users do not immediately notice that they are being paid to do that.7 Since superiors

provide online commentators with a certain degree of discretion and—at the same time—skilful writing of pro-government comments requires effort, it is not surprising that online commentators need to be provided with pecuniary incentives. They are often compensated on a piece rate basis (which inspired the name of the Fifty Cent Party), but some of them receive salaries or, in the case of students, a work-study compensation. Commentators may also be rewarded implicitly (e.g., in the form of improved career prospects) or in a non-pecuniary way (e.g., in the form of awards for best achievers).\footnote{See Han (2015).}

There are also other—more indirect—negative implications of manipulating information. It has been noted, for instance, that having a system of online commentators may increase the distrust of internet users in any favourable opinion about the regime since, when online commentators are known to be present, any pro-government voice can be suspected to be paid propaganda.\footnote{Link, P., “Censoring the News Before It Happens”, The New York Review of Books Blog, 10 July 2013; \url{http://www.nybooks.com/blogs/nyrblog/2013/jul/10/censoring-news-before-happens-china/}}

The aim of this paper is to formally analyse the incentives of authoritarian regimes to manipulate information and to highlight the role of the citizens’ ability to coordinate a collective action. The model uses the standard global games framework in which a continuum of citizens observe private and heterogeneous signals about the incumbent regime’s strength, and then simultaneously choose whether to attack it. If the aggregate attack is sufficiently large relative to the incumbent’s strength, the regime change is successful. Citizens receive a positive payoff if they participate in a successful attack, however, this participation is costly. The regime is able to manipulate the information available to the citizens by adding noise to their private signals. In the appendix, I present an illustration of how this kind of information manipulation could arise in the context of employing pro-government internet trolls.

The results of the model are closely related to empirical findings of the political science literature, most notably King et al. (2013, 2014, 2016).

In Section \ref{sec:noise} I analyse the regime’s incentives to increase noise in the citizens’ private information. These turn out to be strongest for moderate values of (i) the citizens’ individual cost of attacking the regime, and (ii) the mean strength of the regime. Intuitively, when the regime adds noise to citizens’ private signals, they attach more weight to the prior information when deciding on whether to attack. Consequently, with a sufficiently high individual cost of attacking and the incumbent expected to be strong enough, the regime may ensure that the citizens choose not to attack even for very low realisations of their private signals. On the other hand, as the cost of attacking the regime and the
mean strength of the incumbent become arbitrarily high, the regime is almost bound to survive and the marginal benefit from adding noise diminishes, which makes the incentives non-monotonic. In Sections 3.2 and 3.3 I analyse the regime’s incentives to increase noise further and present two settings with alternative objective functions of the regime, which aim to capture the ideas behind, respectively, the state critique theory and the theory of collective action potential.

Section 3.2 introduces a “state critique” benchmark in which the regime cares about the citizens’ opinion *per se*, rather than about the possibility of a revolt that could result in an overthrow. I show that if it is citizens’ *average* opinion that is the incumbent’s concern, then the regime has no incentives to increase noise in their signals.

In Section 3.3 I analyse the role of collective action in the regime’s incentives to prevent regime change. In particular, I show that the regime’s incentives to minimise all collective attacks and to prevent regime change are related yet distinct. Furthermore, I identify the conditions under which they are equivalent, which helps illuminate the circumstances under which minimising all attacks, regardless of the likelihood of their success, may be a good proxy instrument for preventing an overthrow. I show that this is likely to be true when (i) the citizens have little more information about the regime’s strength than the incumbent has, and (iii) the size of collective attacks—particularly when they are of moderate scale—strongly affects the chances of a successful overthrow.

The contrasting results in Sections 3.2 and 3.3 provide a certain theoretical rationale for the empirical observation of King et al. (2013, 2014) that the Chinese regime’s objective, rather than to suppress criticism of the state *per se*, is to silence comments that may lead to a collective action.

Section 3.4 discusses the relationship of the above results with King et al. (2013, 2014) in more detail, and also suggests a possible linkage to the main empirical finding of King et al. (2016) which says that the Chinese regime employs social media commentators to distract citizens rather than to persuade them. If a regime wants to prevent an overthrow, information manipulation in the form of adding noise (which could capture distraction as opposed to persuasion) turns out to be effective even if (i) the citizens can observe the regime’s manipulative action (i.e. in this case, how much noise is being added), and (ii) the regime is not better informed about its own strength than the citizens are. On the

---

10 For example, a number of papers in the attention allocation literature assume a linear noise reduction technology, i.e. that the attention devoted to an information source is linearly devoted to the precision of the signal. See, e.g., Myatt and Wallace (2012), Mondria and Quintana-Domeque (2013), and Chen and Suen (2016).

11 The assumption in (i) is that the citizens observe how much noise (i.e. the variance of the error term) is being added by the regime to their private signals, but they do not observe the realisation of the extra
other hand, as is discussed in more detail below, Edmond (2013) shows that information manipulation in the form of adding a bias (which could capture persuasion) is effective only if the two informational conditions above do not hold. Thus, information manipulation in the form of adding noise is effective under much less stringent informational requirements than adding a bias.

In Section 4 I focus on another important aspect of the theory of collective action potential: the ability of citizens to coordinate their actions. In particular, I investigate how the regime’s incentives to prevent regime change identified in Section 3.1 are affected by the degree to which the citizens can coordinate a collective attack. I show that a crucial role is played by how well the citizens can aggregate their collective information. If they can perfectly coordinate an attack (i.e. act essentially like one “large” citizen) as well as perfectly aggregate their collective information, the regime has in fact no incentives to increase noise. However, if they can perfectly coordinate their actions only at the cost of impeded information aggregation, the regime’s incentives to increase noise may be even stronger than if the citizens were imperfectly coordinated but could still partially aggregate their information. This will be especially true if the regime is expected to be strong but the prior information about it is imprecise.

The paper is organised as follows. Section 2 outlines the model and characterises the equilibrium. Section 3 analyses the regime’s incentives to manipulate information, with an emphasis on the role of state criticism and collective action, while Section 4 investigates the impact of the citizens’ ability to coordinate their actions. Section 5 concludes.

1.1 Related Literature

The model presented here draws from the extensive literature on global games, which was introduced by Carlsson and van Damme (1993) and further developed by Morris and Shin (1998, 2001, 2002). The subject of the paper is closely related to Edmond (2013), who analyses a coordination game of regime change in which the incumbent can add a bias to the agent’s private signals about the regime’s strength, and to the literature which investigates the impact of private signal precision on the probability of regime change, e.g., Metz (2002), Bannier and Heinemann (2005), Iachan and Nenov (2015), and Szkup and Trevino (2015). I also focus on the role of signal precision and add to the literature by looking further into the impact of collective action and coordination on the regime’s incentives to increase error term, and therefore they cannot back it out.

Morris and Shin (2003) provide an excellent survey of the global games literature.
noise, which leads to interesting observations on the empirical findings of King et al. (2013, 2014, 2016).

In Edmond (2013), the regime is informed about its strength, and its manipulative action—in the form of a linear bias added to the private signals—is unobserved by the citizens. This informational setting allows weaker regimes to mimic the strong ones, however, since information manipulation is costly, only the intermediate regimes do so. Crucially, the manipulation is not backed out by the citizens because they need to infer the regime’s hidden action jointly with their inferences about the incumbent’s strength. As a result, they cannot perfectly disentangle the regime’s strength from the endogenous bias. As I show later, contrasting the implications of the model presented here with those of Edmond (2013) provides interesting linkages to King et al. (2013, 2014, 2016).

Little (2016) presents a theoretical model whose results are also related to King et al. (2013, 2014). In his analysis, he distinguishes between “political coordination” (i.e. a sufficient number of citizens must protest to successfully overthrow) and “tactical coordination” (i.e. they must choose a similar tactic) by combining elements of the “beauty contest” branch of global games (e.g., Morris and Shin, 2002) and the “participation” models (e.g., Angeletos et al., 2007). The model shows that reducing the precision of citizens’ information about tactics is always an effective way to reduce protest, while reducing the precision of the signal about the regime’s popularity is only sometimes. Petrova and Zudenkova (2015) make a related distinction between “tactical censorship”, i.e. one which makes it more costly to participate in protest, and “content censorship”, i.e. one which biases downwards a signal of the regime’s unpopularity. Neither of these papers, however, distinguishes between distraction and persuasion as methods of information manipulation, and thus neither provides insights into their relative merits and the circumstances under which they are likely to be used.

The paper is also related to the growing economic literature on censorship. Shadmehr and Bernhardt (2015) analyse the strategic choices of authoritarian regimes on when to censor the information available to citizens, and they show that bad rulers may prefer media to be strong. This is because when revolution payoffs are high, a ruler values strong media that might uncover good news about the status quo which could prevent a revolution. Although the mechanisms are quite different, one can see parallels to the result in Section 3.1 that the mean strength of the regime must be sufficiently high for the regime to have incentives to increase noise.

Lorentzen (2014) analyses the trade-off of an authoritarian regime in setting the level of censorship, where free media provide the information necessary for the regime to discipline
lower-level bureaucrats, but—at the same time—they may facilitate a coordinated revolt if discontent is revealed to be widespread. The model provides insights into why an increase in uncontrollable information may lead to a reduction in media freedom, for example, in China. The same trade-off is analysed by Egorov, Guriev and Sonin (2009) who show that providing bureaucrats with incentives is less important for authoritarian regimes in resource-rich economies and, therefore, media are more likely to be free in these countries. Guriev and Treisman (2015) on the other hand, investigate how information manipulation, e.g., censorship and state propaganda, may interact with other methods of preventing regime change such as co-opting the elite and equipping the police to repress attempted uprisings.

The idea of information manipulation appears also in the literature on media capture, for example, Besley and Prat (2006) Petrova (2008) Enikopolov, Petrova and Zhuravskaya (2011) and Gehlbach and Sonin (2014). Recent developments in social media have been studied by Enikopolov, Petrova and Sonin (2015) who provide empirical evidence that social media can discipline corruption even in countries with limited political competition. Enikopolov, Makarin, and Petrova (2016) find that increased social media penetration in Russia has boosted both the probability of protest and the number of protesters. They also identify the main channel to be the reduction in the costs of coordination, rather than the spread of information critical of the government.


2 The Model

2.1 Setup

Game of Regime Change. The model uses the global games framework of modelling regime change. There is a continuum of agents of measure 1, indexed by \( i \) and uniformly distributed over \([0, 1]\). Agents simultaneously choose between two actions: they can either attack the regime (e.g., protest, join a revolt) or refrain from attacking. The agents’ payoffs from the two possible actions depend on whether the attack turns out to be successful:
<table>
<thead>
<tr>
<th>Attack is successful</th>
<th>Status quo remains</th>
</tr>
</thead>
<tbody>
<tr>
<td>Attack</td>
<td>$1 - c$</td>
</tr>
<tr>
<td>Do Not Attack</td>
<td>$-c$</td>
</tr>
</tbody>
</table>

If an agent chooses not to attack the regime, she receives a payoff of zero. Parameter $c \in (0, 1)$ denotes an agent’s individual cost of attacking the regime. If an agent attacks and the regime is overthrown, the agent receives a payoff of $1 - c > 0$. Otherwise, she only incurs the cost with no benefit from regime change, and thus her payoff is $-c < 0$. The assumption that $c \in (0, 1)$ ensures that neither action is a dominant strategy for the agent.

The regime is overthrown if $f(A) \geq \theta$, where $A$ is the measure of attacking agents, $\theta \in \mathbb{R}$ is the regime’s strength, and $f : [0, 1] \to \mathbb{R}$ is an increasing, $C^1$ function that determines the productivity of attacks. An agent’s incentive to attack thus increases with the aggregate size of the attack, implying that the citizens’ action choices are strategic complements. For most of the paper, I assume that $f(A) = A$.

**Information Structure.** Agents have heterogeneous information about the regime’s strength, $\theta$. They share the common prior $\theta \sim \mathcal{N}(\bar{\theta}, \sigma^2_{\theta})$, and each agent $i$ receives a private signal $x_i = \theta + \varepsilon_i$, where $\varepsilon_i \sim \mathcal{N}(0, \sigma^2_{\varepsilon})$ is independently and identically distributed across agents and independent of $\theta$. Importantly, the regime can (publicly) control $\sigma^2_{\varepsilon}$, i.e. the variance of noise in agents’ private signals. Given this, an agent $i$’s posterior is:

$$E[\theta | x_i] = \frac{\sigma^2_{\varepsilon}}{\sigma^2_{\theta} + \sigma^2_{\varepsilon}} \bar{\theta} + \frac{\sigma^2_{\theta}}{\sigma^2_{\theta} + \sigma^2_{\varepsilon}} x_i. \quad (1)$$

**Regime’s Objective.** The regime’s objective function is assumed to be

$$U_R(\sigma^2_{\varepsilon}) := \left[ 1 - \Phi_{RC}(\theta^*) \left( \sigma^2_{\varepsilon} \right) \right] - \lambda C \left( \left| \sigma^2_{\varepsilon} - \sigma^2_{\varepsilon} \right| \right), \quad (2)$$

where $\Phi_{RC}(\theta^*)$ is the probability of regime change and function $C(\cdot)$ parametrises the regime’s cost of information manipulation, with $C'(\cdot) > 0$ and $C''(\cdot) > 0$. The intrinsic level of noise in the agents’ private signals is given by $\sigma^2_{\varepsilon}$, and so the cost is higher the

---

$^{13}$It is straightforward to introduce additional features to the agents’ payoffs in this framework, e.g., an agent could receive a spillover payoff, $s \in (0, 1 - c)$, if she does not attack and the regime is still overthrown because a sufficient measure of the other agents have attacked, or an agent could incur an additional cost of being repressed, $r > 0$, if she unsuccessfully attacks. In the appendix, I show that in this framework an increase in either $s$ or $r$ is equivalent to an appropriately scaled increase in $c$.

$^{14}$In the appendix, I show how—in the context of this setting—the regime could increase $\sigma^2_{\varepsilon}$ by employing pro-government social media commentators who have both intrinsic and extrinsic motivations to write favourably about the incumbent.
further from that level is $\sigma^2$ set by the regime. Parameter $\lambda$ measures the relative importance of the cost component relative to the “benefit” component in the regime’s objective function.

**Timeline of the Game.** The game proceeds in the following steps:

1. The nature draws $\theta$ from a normal distribution $\mathcal{N}(\bar{\theta}, \sigma^2_\theta)$, which defines the initial common prior about the incumbent regime’s strength, $\theta$.

2. The regime (publicly) chooses the variance of noise in agents’ private signals, $\sigma^2_\varepsilon$.

3. The regime’s strength, $\theta$, is realised.

4. Each agent $i$ receives a private signal $x_i = \theta + \varepsilon_i$, forms a posterior about the regime’s strength, $\theta$, and decides whether or not to attack the regime.

5. The regime is overthrown, or it survives. All players’ payoffs are realised.

The possible interpretations of this setting are not limited to political change, but can also include self-fulfilling bank runs, currency attacks, and debt crises. In the appendix, I present an application of the model to the context of paid online commentators employed by authoritarian regimes.

### 2.2 Equilibrium

In what follows, we limit our discussion to perfect Bayesian Nash equilibria in monotone strategies, that is, equilibria in which the agents’ strategies are non-decreasing in each agent’s private signal, $x$. The reason for focusing on monotone strategies is twofold. First, the cumulative distribution function of an agent’s posterior about the regime’s strength, $\theta$, is decreasing in her private signal, $x$. Furthermore, as long as the signal is sufficiently low (more precisely, if $x < \bar{x}$, where $c$), the citizens find it strictly dominant to attack. Conversely, if the signal is sufficiently high (more precisely, if $x > \bar{x}$, where $\mathbb{P}(\theta \leq 0 | \bar{x}) = c$), it is strictly dominant for them not to attack.

An equilibrium consists of a critical strength of the regime and a critical private signal, $(\theta^*, x^*)$, such that a citizen attacks whenever her private signal is below the critical signal.\footnote{This implies that making the agents’ private signals more precise is also costly. In this paper, I focus on the regime’s incentives to increase noise, however, its incentives to make the signals more precise will be to a large extent a mirror image.}

\footnote{See, e.g., Angeletos et al. (2007).}

\footnote{Restricting attention to monotone Bayesian equilibria is common in the global games literature, e.g., see Angeletos et al. (2007) and Loeper et al. (2013).}
$x \leq x^*$, and the regime is overthrown whenever its true strength is less than the critical strength, $\theta \leq \theta^*$.

An alternative statement is that there is a critical strength of the regime, $\theta^*$, which generates a distribution of private signals such that the proportion of citizens who observe signals less than or equal to the critical value, $x^*$, is exactly $\theta^*$, resulting in a regime change. In other words, in this equilibrium, whenever $\theta \leq \theta^*$, at least $\theta$ players attack and the regime is overthrown. Conversely, whenever $\theta > \theta^*$, no more than $\theta$ players attack and the regime remains in place.

An equilibrium is determined by two conditions. The first of these is the payoff indifference condition (e.g., [1]), which states that—in equilibrium—an agent must be indifferent between attacking and not attacking when she observes a private signal $x^*$.

With the structure of the equilibrium defined as above, this yields the condition

$$P(\theta \leq \theta^* | x^*) = c.$$ (3)

Given the assumptions on the distribution of $\theta$, the left-hand side can be conveniently rewritten as $\Phi\left((\theta^* - E(\theta | x^*)) / \sqrt{Var(\theta | x^*)}\right)$. The payoff indifference condition is then:

$$c = \Phi\left(\frac{\sigma_\theta^2 + \sigma_\xi^2}{\sigma_\theta^2 \sigma_\xi^2} \left(\theta^* - \bar{\theta} - \frac{\sigma_\theta^2}{\sigma_\theta^2 + \sigma_\xi^2} (x^* - \bar{\theta})\right)\right).$$ (4)

The agent attacks if and only if $x \leq x^*$, where $x^*$ is determined in (4).

The second equilibrium condition is the critical mass condition (e.g., [1]). In equilibrium, whenever the true strength of regime is exactly $\theta^*$, the proportion of agents who receive a signal $x \leq x^*$ should equal precisely $\theta^*$:

$$\theta^* = P(x \leq x^* | \theta^*).$$ (5)

With the distributional assumptions stated above, this can be rewritten as:

$$\theta^* = \Phi\left(\frac{x^* - \theta^*}{\sqrt{\sigma_\xi^2}}\right).$$ (6)

The two equilibrium conditions stated in (4) and (6) determine an equilibrium of the game:

**Proposition 1.** (Morris-Shin) In a Perfect Bayesian Equilibrium of the game, a critical
strength of the regime, $\theta^*$, solves:

$$\theta^* = \Phi \left( \frac{\sqrt{\sigma_\epsilon^2} (\theta^* - \bar{\theta}) - \sqrt{\sigma_\theta^2 + \sigma_\epsilon^2 \Phi^{-1} (c)}}{\sigma_\theta^2} \right).$$

(7)

There is a unique equilibrium if $\sigma_\epsilon^2 \leq 2\pi (\sigma_\theta^2)^2$.

As usual in global games, we will assume that the agents’ private signals are sufficiently precise so that there is a unique equilibrium.

In equilibrium, the regime is overthrown whenever $\theta \leq \theta^*$, so the probability of regime change is

$$\Phi_{RC} (\theta^*) = \Phi \left( \frac{\theta^* - \bar{\theta}}{\sqrt{\sigma_\theta^2}} \right),$$

(8)

with $\theta^*$ given by (7).

3 Regime’s Incentives to Manipulate Noise

In this section, I analyse the regime’s incentives to prevent regime change by manipulating noise in the agents’ private information about the regime’s strength. Section 3.1 outlines these incentives. In Sections 3.2 and 3.3, I analyse them further to provide insights into the empirical findings of King et al. (2013, 2014) and their two theories of what constitutes the goals of the Chinese censorship programme: state critique theory and the theory of collective action potential. King et al. (2013, 2014) show empirical evidence that, in the context of censorship in China, the latter theory is correct, while the former is not; in other words, the censorship in China aims to stop the spread of information that could lead to any kind of collective action, rather than to restrict criticism of the state per se.

If we extrapolate the results of King et al. (2013, 2014) to information manipulation in general, it turns out that the model can give some theoretical rationale for their empirical findings. In Section 3.2, I present a setting in which the regime’s concern is the citizens’ perception of the state per se, while in Section 3.3, I perform a similar exercise in which the incumbent cares solely about minimising collective attacks, without regard to how likely they are to succeed. Thus, Sections 3.2 and 3.3 aim to capture the ideas behind, respectively, the state critique theory and the theory of collective action potential. Finally, in Section 3.4, I discuss how the results in this section may connect with King et al. (2016).
3.1 Preventing Regime Change

The impact of the precision of the agents’ private signals on the chances of regime change has been studied by Metz (2002), Bannier and Heinemann (2005), Iachan and Nenov (2015), and Szkup and Trevino (2015). Here, I evaluate the regime’s incentives to increase noise in agents’ private signals, $\sigma^2$, in the context of the objective function in (2).

It is useful to start the analysis with the following lemma:

Lemma 1. The probability of regime change, $\Phi_{RC}(\theta^*)$, is decreasing in:

(i) the mean strength of the regime, $\bar{\theta}$, i.e. $\frac{\partial \Phi_{RC}(\theta^*)}{\partial \bar{\theta}} < 0 \ \forall \bar{\theta}$,

(ii) the agents’ individual cost of attacking, $c$, i.e. $\frac{\partial \Phi_{RC}(\theta^*)}{\partial c} < 0 \ \forall c$.

These results are in line with common intuition. Looking at (4), we can see that an increase in either the mean strength of the regime, $\bar{\theta}$, or the agents’ individual cost of attacking, $c$, results in a decrease in $x^*$, i.e. the critical value of the private signal that makes an agent indifferent between attacking the regime and not doing so. In fact, as long as the equilibrium uniqueness condition in Proposition 1 is satisfied, the regime’s critical strength, $\theta^*$, is decreasing in both $\bar{\theta}$ and $c$, which ensures that fewer agents choose to attack and the regime becomes less likely to be successfully overthrown.

Another useful observation is:

Lemma 2. The probability of regime change is decreasing in the variance of noise in agents’ private signals, i.e. $\frac{\partial \Phi_{RC}(\theta^*)}{\partial \sigma^2_\varepsilon} < 0$, if and only if $\theta^* < T_{RC}$, where

$$T_{RC} = \bar{\theta} + \sqrt{\frac{\sigma^2_\theta \sigma^2_\varepsilon}{\sigma^2_\theta + \sigma^2_\varepsilon} \Phi^{-1}(c)}.$$  \(9\)

In other words, the regime’s critical strength, $\theta^*$, must be low enough for the probability of regime change to be decreasing in noise in the agents’ private signals, $\sigma^2_\varepsilon$. Note here that, by Lemma 1, $\partial \theta^*/\partial \bar{\theta} < 0$ and $\partial \theta^*/\partial c < 0$, while threshold $T_{RC}$ is increasing in both $\bar{\theta}$ and $c$. Therefore, the higher is the mean strength of the regime, $\bar{\theta}$, or the individual cost of attacking, $c$, the more likely it is that an increase in $\sigma^2_\varepsilon$ lowers the chances of regime change.

The intuition behind the result in Lemma 2 is as follows. When the mean strength of the incumbent (or the cost of attacking the regime) is high, the agents’ prior incentives to attack are rather weak. However, since the fundamentals of the model are normally

---

18 This result is substantially different from a model with common knowledge of the state of the world, in which case there are multiple equilibria and the outcome of the game is not affected by the fundamentals.
distributed, there is always a positive probability that an agent receives a very low private signal inducing her to attack the regime. By lowering the precision of private signals, the regime makes the agents attach less weight to them and more to the prior. This may lead the agents to refrain from attacking even for very low private signals, and a regime change becomes less likely.

Information manipulation is nonetheless costly for the regime. If we incorporate these costs into the analysis of the regime’s incentives, we arrive at the following.\footnote{The result that $\bar{\theta}$ and $c$ must be sufficiently high is parallel to Metz (2002) and Bannier and Heinemann (2005). They consider the central bank’s optimal rules for minimising the probability of currency crises and show that full transparency is not always optimal: when prior beliefs about economic performance are good, the central bank should disclose imprecise information. In the appendix, I provide additional additional results on the behaviour of $\theta^*$ and $\Phi_{RC}$ as a function of $\sigma^2_e$.}

**Proposition 2.** The regime’s incentives to increase the variance of noise in the agents’ private signals, $\sigma^2_e$, are non-monotonic with respect to the mean strength of the regime, $\bar{\theta}$, and the agents’ individual cost of attacking the regime, $c$:

(i) for sufficiently low and sufficiently high values of $\bar{\theta}$ (or $c$), the regime chooses not to raise $\sigma^2_e$;

(ii) for moderate values of $\bar{\theta}$ (or $c$), the regime chooses to raise $\sigma^2_e$ if and only if $\lambda$ is small enough.

Figure 1 presents a simulation of the result stated in Proposition 2.\footnote{A simulation with respect to the agents’ individual cost of attacking, $c$, produces a very similar figure.}

![Figure 1: A simulation of the regime’s optimal choice of the variance of noise in the agents’ private signals, $\sigma^2_e$, as a function of the mean strength of the regime, $\bar{\theta}$.}
The fact that the regime would never choose to increase the noise in the agents’ private signals for sufficiently low values of \( \bar{\theta} \) and \( c \) follows from Lemma 2. On the other hand, we know from Lemma 1 that, as \( \bar{\theta} \) or \( c \) increase to very high levels, the probability of regime change, \( \Phi_{RC} \), becomes very small; as a result, the marginal benefit of increasing the noise in the agents’ private signals falls. In fact, as \( \bar{\theta} \) (or \( c \)) becomes arbitrarily large, the marginal decrease in \( \Phi_{RC} \) caused by raising \( \sigma^2 \) approaches zero. Thus, given that increasing \( \sigma^2 \) is costly, the regime would also not choose to raise \( \sigma^2 \) for sufficiently high values of \( \bar{\theta} \) and \( c \).

If the weighting parameter, \( \lambda \), is small enough, then the benefits from a decrease in \( \Phi_{RC} \) are large relative to the marginal costs, and there exists an interval of moderate values of \( \bar{\theta} \) and \( c \), where the regime optimally chooses to increase \( \sigma^2 \).

3.2 State Critique Benchmark

Suppose that the regime would like the average public opinion to be as high as possible and is not interested in the \textit{ex ante} probability of being overthrown. This captures the idea that the concern is the criticism of the state \textit{per se} and not the possibility of collective action.

The fact that the regime cares about the \textit{average} public opinion means that it is hurt by a deterioration in an agent’s opinion in the same way regardless of how good or bad their initial stance had been. This objective of the regime can be illustrated with a utility function whose “benefit” component is linear in the agents’ posterior about the incumbent’s strength:

\[
U^{SC}_{R} (\sigma^2) := \mathbb{E}_{\theta, \varepsilon} \left[ \mu \left( x \left( \sigma^2 \right) \right) \right] - \lambda C \left( |\sigma^2 - \sigma^2| \right),
\]

where \( \mu (x) \) is an agent’s posterior about \( \theta \) given a private signal \( x \), \( \mu (x) = \mathbb{E} [\theta | x] \). As this implies that coordination among agents does not matter, we can assume here—without loss of generality—that there is only one agent.

The regime’s utility function in (10) provides an economic interpretation of what constitutes the objectives of the regime according to the state critique theory, as defined by King et al. (2013):

“(...) \textit{State critique theory} (...) posits that the goal of the Chinese leadership is to suppress dissent, and to prune human expression that finds fault with elements of the Chinese state, its policies, or its leaders. The result is to make the sum total of available
public expression more favorable to those in power.”

The idea here is that, if the regime fears state critique, its utility should change proportionally with the agents’ beliefs about the regime’s strength. Consequently, small changes in the agents’ opinions about (and their criticism of) the regime should not be a cause for major concern—and this should be true across the whole spectrum of the agents’ initial opinions.

It turns out that if the regime fears state critique in the sense of having the utility function stated in (10), then the following holds:

**Proposition 3.** Suppose the regime’s benefit function is linear in the agents’ posterior expectations about the incumbent’s strength and therefore the regime’s objective function is

$$U^SC_R(\sigma^2) := \mathbb{E}_{\theta, \varepsilon}[\mu(x(\sigma^2))] - \lambda C(\sigma^2).$$

Then the regime would never choose to increase the variance of noise in the agents’ private signals, $\sigma^2_{\varepsilon}$.

This result is due to the fact that, regardless of the value of $\sigma^2_{\varepsilon}$ chosen by the regime, the *ex ante* expectation of an agent’s posterior always equals $\bar{\theta}$. Given that the regime is not better informed than citizens about its own strength, the regime’s marginal benefit from increasing $\sigma^2_{\varepsilon}$ is thus always zero. Since the marginal cost is strictly positive, it follows that the regime has no incentives to increase $\sigma^2_{\varepsilon}$. This implication naturally would be different if the regime had more precise information about $\theta$ than the agents, as then it would presumably have incentives to increase $\sigma^2_{\varepsilon}$ when knowing that $\theta$ is likely to be lower than the agents believe.

More generally, Proposition 3 suggests that, in the given informational setting, the regime’s incentives to prevent regime change must be driven by something other than the average public opinion. The model may thus be used to provide a possible theoretical rationale for the empirical finding of King et al. (2013, 2014) that restricting state critique *per se* is not the Chinese regime’s target in censorship. This connection is discussed further in Section 3.4.

### 3.3 Targeting Collective Action

In this section, I analyse the extent to which the regime’s incentives to prevent regime change are driven by the desire to minimise all collective action. In particular, I am interested in identifying the circumstances under which the incentives to prevent regime change are equivalent to minimising all kinds of collective attacks, regardless of how likely they are to overthrow the regime. The purpose of this exercise is to shed more light on
the theory of collective action potential in King et al. (2013, 2014) which says that the regime’s objective is to stop the spread of information that could lead to any kind of collective action.

Suppose here that the regime’s objective function is:

\[ U^{CA}_{R}(\sigma_{\varepsilon}^2) := E[1 - A(\theta^*)] - \lambda C(\sigma_{\varepsilon}^2 - \sigma_{\theta}^2), \]  

(11)
i.e. the benefit component in the regime’s utility function is the expected measure of agents who choose not to participate in a collective attack.

Given that in the equilibrium an agent attacks if and only if \( x \leq x^* \), the expected size of the collective attack is \( \mathbb{P}(x \leq x^*) = \Phi\left(\frac{x^* - \bar{\theta}}{\sqrt{\sigma_{\theta}^2 + \sigma_{\varepsilon}^2}}\right) \), i.e.

\[ \mathbb{E}[A(\theta^*)] = \Phi\left(\theta^* - \bar{\theta} \frac{\sqrt{\sigma_{\theta}^2 + \sigma_{\varepsilon}^2}}{\sigma_{\theta}^2} - \frac{\sigma_{\theta}^2 \Phi^{-1}(c)}{\sigma_{\theta}^2}\right), \]  

(12)

where \( \theta^* \) is given by (7).

Clearly, like the probability of regime change, the expected size of the attack is decreasing in both the mean strength of the regime, \( \bar{\theta} \), and the individual cost of attacking the regime, \( c \). Furthermore, as \( \bar{\theta} \) and \( c \) become arbitrarily large, the expected size of the attack will be zero, and it will be 1 while when they are arbitrarily small. This suggests that the regime’s incentives to minimize the expected size of the attack will follow a similar pattern to those in Proposition 2, i.e. they will be non-monotonic with respect to \( \bar{\theta} \) and \( c \), being highest for moderate values.\(^{21}\)

Comparing (12) with (8), we can see however that, for a sufficiently small value of \( \bar{\theta} \) (or \( c \)) and for \( \sigma_{\varepsilon}^2 > 0 \), the expected size of the attack is larger than the probability of regime change, but any increase in \( \bar{\theta} \) (or \( c \)) always has a more prominent impact on \( \mathbb{E}[A(\theta^*)] \) than it has on \( \Phi_{RC} \). Consequently, for a sufficiently high value of \( \bar{\theta} \) (or \( c \)), the converse holds: \( \mathbb{E}[A(\theta^*)] < \Phi_{RC} \).\(^{22}\) This suggests that the regime’s incentives to minimize all kinds of collective attacks are not exactly the same as its incentives to prevent regime change.

\(^{21}\)I present these results formally, as well as provide other supplementary results for the analysis in this section, in the appendix in the proof of Lemma 3.

\(^{22}\)As \( \sigma_{\varepsilon}^2 \) becomes arbitrarily small, we have \( \mathbb{E}[A(\theta^*)] = \Phi_{RC} \). With \( \sigma_{\varepsilon}^2 \rightarrow 0 \), the agents observe the regime’s strength almost perfectly and attack whenever \( \theta < \lim_{\sigma_{\varepsilon}^2 \rightarrow 0} \theta^* = 1 - c \). This is also precisely when the regime is overthrown.
(i) Individual cost of attacking is relatively low: $c < \frac{1}{2}$.

(ii) Individual cost of attacking is intermediate: $c = \frac{1}{2}$.

(iii) Individual cost of attacking is relatively high: $c > \frac{1}{2}$.

Figure 2: Marginal change in the probability of regime change and the size of the collective attack as a function of the noise in the agents’ private signals, $\sigma^2_\varepsilon$, and the regime’s critical strength, $\theta^*$. 

17
The following lemma is an equivalent of Lemma 2 stated for the objective in (11):

**Lemma 3.** The expected size of the collective attack is decreasing in the variance of noise in agents’ private signals, \( \frac{\partial E[A(\theta^*)]}{\partial \sigma^2_\varepsilon} < 0 \), if and only if \( \theta^* < T_A \), where

\[
T_A = \tilde{\theta} + \sqrt{\frac{\sigma^2_\theta (\sigma^2_\theta + \sigma^2_\varepsilon)}{\sigma^2_\varepsilon}} \Phi^{-1}(c) - 2 \left( \sigma^2_\theta + \sigma^2_\varepsilon \right) \frac{\partial \theta^*}{\partial \sigma^2_\varepsilon}. \tag{13}
\]

Lemmas 2 and 3 imply that the thresholds which determine whether the size of the collective attack and the probability of regime change are decreasing in \( \sigma^2_\varepsilon \) are different: \( T_A \neq T_{RC} \). Figure 2 provides an illustration.

Since the indirect effect captured by the third term in (3) turns out to be second-order in the limit as \( \sigma^2_\varepsilon \to 0 \), the following holds when \( \sigma^2_\varepsilon \to 0 \): (i) \( T_A < T_{RC} \) when \( c < \frac{1}{2} \), (ii) \( T_A = T_{RC} \) when \( c = \frac{1}{2} \), and (iii) \( T_A > T_{RC} \) when \( c > \frac{1}{2} \). In other words, in the limiting case \( \sigma^2_\varepsilon \to 0 \), when the individual cost of attacking the regime is relatively low, \( c < \frac{1}{2} \), we have that \( T_A < T_{RC} \), which implies that there exist values of \( \sigma^2_\varepsilon \) and \( \theta^* \) where the probability of regime change is decreasing in \( \sigma^2_\varepsilon \), but the size of the collective attack is increasing in \( \sigma^2_\varepsilon \).

On the other hand, the reverse is true when the individual cost of attacking the regime is relatively high, \( c > \frac{1}{2} \). In the knife-edge case where \( c = \frac{1}{2} \), the thresholds \( T_A \) and \( T_{RC} \) coincide and therefore we have that \( \frac{\partial \Phi_{RC}(\theta^*)}{\partial \sigma^2_\varepsilon} > 0 \) if and only if \( \frac{\partial E[A(\theta^*)]}{\partial \sigma^2_\varepsilon} > 0 \).

Furthermore, it follows from Lemma 2 and Lemma 3 that \( |T_A - T_{RC}| \) approaches zero as \( \sigma^2_\varepsilon \) becomes arbitrarily large. To see this, note that the second terms in (9) and (3) both approach \( \sqrt{\sigma^2_\varepsilon \Phi^{-1}(c)} \) as \( \sigma^2_\varepsilon \to \infty \), and that the indirect effect captured by the third term in (3) approaches zero. Thus, as the agents’ private signals become more noisy, the agents have less extra information about the regime’s strength compared to what the incumbent has, and so the regime’s incentives to prevent collective attacks and to prevent regime change become similar.

The following proposition states these results more precisely:

**Proposition 4.** Consider the marginal change in the probability of regime change, \( \frac{\partial \Phi_{RC}(\theta^*)}{\partial \sigma^2_\varepsilon} \), and in the expected size of the collective attack, \( \frac{\partial E[A(\theta^*)]}{\partial \sigma^2_\varepsilon} \). Then:

(i) as \( \sigma^2_\varepsilon \to 0 \), \( \frac{\partial \Phi_{RC}(\theta^*)}{\partial \sigma^2_\varepsilon} < 0 \) if and only if \( \tilde{\theta} > 1 - c \), and \( \frac{\partial E[A(\theta^*)]}{\partial \sigma^2_\varepsilon} < 0 \) if and only if \( \tilde{\theta} > \tau_{A,0}(c) \),

(ii) as \( \sigma^2_\varepsilon \to \infty \), both \( \frac{\partial \Phi_{RC}(\theta^*)}{\partial \sigma^2_\varepsilon} < 0 \) and \( \frac{\partial E[A(\theta^*)]}{\partial \sigma^2_\varepsilon} < 0 \) if and only if \( \tilde{\theta} > \tau_{A,\infty} = \tau_{RC,\infty} \), where \( \tau_{A,0}(c) < \tau_{A,\infty}(c) < -1 \), and \( \frac{\partial \Phi_{RC}(\theta^*)}{\partial \sigma^2_\varepsilon} = \frac{\partial E[A(\theta^*)]}{\partial \sigma^2_\varepsilon} = 0 \) for both limiting cases when \( \tilde{\theta} = c = \frac{1}{2} \).

\(^{23}\)I show the latter in the appendix.
It is best to explain the statements made in Proposition 4 with the use of Figure 3. With $c$ on the $x$-axis and $\bar{\theta}$ on the $y$-axis, Figure 3 illustrates the parameter spaces where $\frac{\partial \Phi_{RC}(\theta^*)}{\partial \sigma^2}$ and $\frac{\partial E[A(\theta^*)]}{\partial \sigma^2}$ are positive and negative in two limiting cases, $\sigma^2 \to 0$ and $\sigma^2 \to \infty$. Threshold $\tau_{RC}$ depicts the critical value of $\bar{\theta}$—as a function of $c$—such that whenever $\bar{\theta}$ is larger than this critical value, the probability of regime change in the global game is decreasing in $\sigma^2$, i.e. $\frac{\partial \Phi_{RC}(\theta^*)}{\partial \sigma^2} < 0$. Threshold $\tau_A$ illustrates the equivalent for the size of the collective attack.

Figure 3: Marginal change in the probability of regime change, $\frac{\partial \Phi_{RC}}{\partial \sigma^2}$, and the size of the collective attack, $\frac{\partial E[A(\theta^*)]}{\partial \sigma^2}$, as a function of the individual cost of attacking the regime, $c$, and the mean strength of the regime, $\bar{\theta}$, in the limiting case where (i) $\sigma^2 \to 0$, and (ii) $\sigma^2 \to \infty$.

It follows from Proposition 4 that the slope of threshold $\tau_A$ is steeper than that of threshold $\tau_{RC}$ when $\sigma^2 \to 0$, and that the two thresholds coincide when $\sigma^2 \to \infty$. Furthermore, the very last statement in the proposition implies that the two thresholds cross when $\bar{\theta} = c = \frac{1}{2}$.

In Figure 3(i), we can see that when the agents observe precise private signals, i.e. $\sigma^2$ is small, the regime’s incentives to prevent regime change and to prevent collective attacks are rather different. In order to prevent regime change, the regime should increase noise only when the mean strength of the incumbent, $\bar{\theta}$, and the individual cost of attacking, $c$, are both sufficiently high. This is due to Lemma 2. On the other hand, when the regime aims to minimise all collective attacks—as opposed to only those which are likely to be successful—then it is less concerned about whether any given attack is likely to be successful. Therefore, its optimal strategy of whether to increase noise depends predominantly...
on the individual cost of attacking the regime, and the mean strength of the incumbent has little effect on whether $\frac{\partial \mathbb{E}[\theta^*]}{\partial \sigma^2}$ is positive or negative.

As a result, for high values of $c$ and sufficiently low values of $\bar{\theta}$, the regime has no incentives to increase $\sigma^2$ so as to prevent regime change, but it does have them if the objective is to prevent collective attacks per se. The converse holds when $c$ is low and $\bar{\theta}$ is sufficiently high. This suggests that when the agents’ private signals are intrinsically precise and the regime aims to minimise all collective attacks, it may (i) unnecessarily increase noise when individual cost of attacking the regime is high but the incumbent is expected to be weak, and (ii) increase noise insufficiently when the individual cost of attacking the regime is low but the incumbent is expected to be strong.

The situation is different, however, when the agents’ signals are intrinsically noisy, which is illustrated in Figure 3(ii). The agents’ private signals provide then little additional information about the regime’s strength relative to what the incumbent knows and, consequently, the regime’s incentives to prevent regime change and to prevent collective attacks become similar. In other words, minimising collective attacks is then a very good “proxy instrument” for preventing regime change.

**Role of the productivity of attacks.** An important role in the regime’s incentives to prevent regime change is played by function $f(A)$, which measures how productive an aggregate attack of size $A$ is: the regime is overthrown if and only if $\theta \leq f(A)$. As is the case in most of the global games literature, I have assumed so far that $f(A) = A$. It is worthwhile, however, to consider more general forms of $f(\cdot)$ to shed light on the impact of the productivity of collective attacks on the regime’s incentives to prevent regime change.

I focus here on the limiting case where $\sigma^2 \to 0$, that is, when the discrepancy between the incentives to prevent regime change and those to prevent collective attacks is largest. Unsurprisingly, it turns out that the exact shape of the regions illustrated in Figure 3 depends on the specific functional form of $f(A)$. Figure 4 illustrates this with three examples which assume that, as with $f(A) = A$, the range of values of $f(A)$ is restricted to $[0, 1]$.

Figure 4(i) illustrates a case where the productivity of attacks rises very steeply when they are of moderate size. Under these circumstances, the regime’s incentives to prevent regime change and to prevent collective attacks turn out be similar. This should not be surprising given that, with the given shape of $f(A)$, the success of a collective attack strongly depends on its size.
(i) \( f(A) \) is S-shaped with \( f'(A) \) being high for moderate values of \( A \);

(ii) \( f(A) \) is inverse S-shaped with \( f'(A) \) being low for moderate values of \( A \);

(iii) \( f(A) \) is S-shaped and asymmetric with \( f'(A) \) being high for low values of \( A \).

Figure 4: The impact of productivity function \( f(A) \) on the regions illustrated in Figure 3(i).

In Figure 4(ii), on the other hand, the size of the aggregate attack, \( A \), has little impact on the productivity of an attack for a wide range of \( A \), except for very low and very high values. In this case, unsurprisingly, the regime's incentives to prevent regime change and
those to prevent collective attacks prove to be very different. This implies that when the success of an attack depends only weakly on its size, minimising collective attacks serves as a poor proxy instrument for avoiding regime change. In the extreme case where \( f(A) \) is effectively a horizontal line with \( f(A) = \frac{1}{2} \) for all \( A \), these incentives are almost “orthogonal”: the sign of \( \frac{\partial \Phi_{RC}(\theta^*)}{\partial \sigma^2} \) depends predominantly on \( \theta \) but not on \( c \), while the reverse holds for the sign of \( \frac{\partial E[A(\theta^*)]}{\partial \sigma^2} \).

Finally, Figure 4(iii) illustrates an asymmetric variant of the first case. If the productivity of an attack rises steeply already for a low size of the attack, then we see an increase in the size of the region where \( \frac{\partial E[A(\theta^*)]}{\partial \sigma^2} < 0 \) and \( \frac{\partial \Phi_{RC}(\theta^*)}{\partial \sigma^2} > 0 \), i.e. where the regime would increase noise in order to prevent collective attacks, but would not do so if its objective were to prevent regime change. At the same time, the region where \( \frac{\partial E[A(\theta^*)]}{\partial \sigma^2} > 0 \) and \( \frac{\partial \Phi_{RC}(\theta^*)}{\partial \sigma^2} < 0 \) shrinks in size.

### 3.4 Discussion

The contrast between the results of the state critique benchmark (Section 3.2) and the scenario in which the regime minimises the expected size of all attacks (Section 3.3) suggests that the model may help explain the main empirical observation of King et al. (2013, 2014)—which says that the Chinese regime’s objective is not to suppress criticism of the state but to silence comments that may lead to any kind of collective action.

Section 3.2 shows that the regime has no incentives to increase noise for the purpose of improving the citizens’ perception of the state per se. This is because an uninformed regime is not able to restrict its strategy of adding noise to specific levels of its own strength (e.g., only to the cases where the regime’s private information indicates it is low). Consequently, adding noise does nothing to improve the citizens’ average perception of the state. Thus, if the regime wants to avoid being overthrown, its focus should be something else and minimising the size of all collective attacks is, as Section 3.3 shows, an important component in its incentives to prevent regime change. Section 3.3 further identifies the circumstances under which—from the point of view of the incumbent’s incentives—preventing regime change is equivalent to minimising all collective attacks by citizens, regardless of how likely they are to overthrow the incumbent.

As it turns out, the model may also be used to clarify the linkage between this result and that of King et al. (2016), which states that the reason why the Chinese regime fabricates social media posts is not because it wants to engage in arguments with skeptics and to defend the government, but because it aims to distract the public.

In the model presented here, when the regime is concerned about the possibility of
a collective action resulting in regime change, information manipulation in the form of adding noise—which could capture distraction of citizens as opposed to persuasion—is effective under much less stringent informational requirements than adding a bias. In particular, noise manipulation is effective even if the agents are aware of how much noise the regime has introduced and the regime is no better informed about its own strength than the agents are. This contrasts with the setting of Edmond (2013), where the regime’s ability to increase the likelihood of survival by adding a bias to the agents’ private signals stems from the fact that they need to infer the regime’s hidden action jointly with their inferences about the incumbent’s strength which is privately known by the regime. In other words, they cannot perfectly disentangle the strength of the regime from the endogenous bias, and are thus unable to back out the regime’s manipulation.

Thus, as far as the finding of King et al. (2016) is concerned, it may be that the Chinese regime is aware of the fact that preventing a collective attack by adding a bias to the public opinion is informationally rather difficult. It would require, in particular, (i) the citizens to be uninformed about the government’s manipulative actions and (ii) the regime to be better informed about its strength than the citizens are. Adding noise into citizens’ information provides an easier yet effective alternative under these circumstances.

4 Role of Coordination

The empirical findings of King et al. (2013, 2014) suggest that the citizens’ ability to coordinate a collective action may have a significant impact on the incentives of authoritarian regimes to manipulate information. In this section, I investigate how the regime’s incentives to prevent regime change (as analysed in Section 3.1) are affected by the degree to which agents can coordinate their actions.

We begin with two benchmarks in which agents are perfectly coordinated. Suppose that the mass of agents acts like one “large” agent who is able to overthrow the regime with certainty as long as the incumbent’s strength, $\theta$, is less than the “large” agent’s mass, $1$. It turns out that the implications of perfect coordination for the regime’s incentives depend on the extent of information aggregation that we assume perfect coordination entails.

If the agents were able to perfectly aggregate all the collective information available
to them in the private signals, they would be able to back out the regime’s strength, \( \theta \), regardless of the level of noise in their private signals, \( \sigma^2_\varepsilon \). Consequently, the regime would have no incentives to increase noise in the agents’ private signals, which is in stark contrast to the regime’s incentives when agents are imperfectly coordinated, as discussed in Section 3.1. In this scenario, the agents would attack (i.e. \( A_{PCA} = 1 \)) and overthrow the regime whenever \( \theta \leq 1 \), and would not attack it otherwise (i.e. \( A_{PCA} = 0 \)). The \textit{ex ante} probability of regime change would thus be \( \Phi_{PCA} = \Phi \left( \left( 1 - \bar{\theta} \right) / \sqrt{\sigma^2_\theta} \right) \).

It may be more illuminating, however, to look at a benchmark where the agents are perfectly coordinated but are not able to aggregate the information they collectively possess. In other words, suppose there is a “large” agent who receives a single signal, \( x = \theta + \varepsilon \), with the variance of the error term being \( \sigma^2_\varepsilon \) as before. One could interpret this setting as one in which there is a “political leader” who is not better informed than any other agent but is able to make all the agents attack the regime if she so desires.\(^{26}\) In this case, the regime still has incentives to increase noise, which makes the results parallel to Edmond (2013)\(^{27}\).

In the perfect coordination benchmark with no information aggregation, the “large” agent chooses to attack the regime if \( \mathbb{P}(\theta \leq 1 \mid x) \leq c \) and does not attack it otherwise. This means that the critical mass condition from the global game is no longer necessary to establish the equilibrium and it is only the payoff indifference condition that matters. The “large” agent is indifferent between attacking and not attacking if

\[
c = \Phi \left( \sqrt{\frac{\sigma^2_\theta + \sigma^2_\varepsilon}{\sigma^2_\theta \sigma^2_\varepsilon}} \left( 1 - \bar{\theta} - \frac{\sigma^2_\varepsilon}{\sigma^2_\theta + \sigma^2_\varepsilon} (x_{PC} - \bar{\theta}) \right) \right),
\]

which is notably different from (4). Thus, the agent attacks if and only if her private signal is less than the critical value, \( x \leq x_{PC}^* \), where:

\[
x_{PC}^* = \frac{\sigma^2_\theta + \sigma^2_\varepsilon}{\sigma^2_\theta} - \frac{\sigma^2_\varepsilon}{\sigma^2_\theta} \bar{\theta} - \sqrt{\frac{\sigma^2_\varepsilon (\sigma^2_\theta + \sigma^2_\varepsilon)}{\sigma^2_\theta}} \Phi^{-1}(c).
\]

Inspection of (4) shows that \( x_{PC}^* > x^* \) for all \( \theta^* < 1 \).

\(^{26}\)These two benchmarks will be referred to as, respectively, the perfect coordination benchmark with perfect information aggregation (“PCA”) and the perfect coordination benchmark with no information aggregation (“PC”).

\(^{27}\)In Edmond (2013), the regime has no incentives to add a bias to the agents’ private signals when the agents are perfectly coordinated and can perfectly aggregate their information, but it is no longer the case when the agent’s abilities to aggregate information are impeded. The latter is briefly discussed in the appendix to his paper.
The regime faces here aggregate uncertainty since it does not know which value of $x$ will be realised. The size of the attack is 1 if $x \leq x_{PC}^*$ and 0 otherwise. In this benchmark, the expected size of the attack is

$$E \left[ A_{PC} \right] = \Phi \left( (1 - \bar{\theta}) \sqrt{\frac{\sigma_\theta^2 + \sigma_\varepsilon^2}{\sigma_\theta^2}} - \sqrt{\frac{\sigma_\varepsilon^2}{\sigma_\theta^2}} \Phi^{-1} (c) \right),$$

(16)

and it is easy to see that it is always higher than (12), i.e. the expected size of the attack when the agents are imperfectly coordinated.

Perfect coordination with no information aggregation thus, on average, always yields more collective attacks than we observe when agents are imperfectly coordinated. This does not mean, however, that also regime change is more likely.

For the regime to be overthrown, we need the agent to attack, which happens when $x \leq x_{PC}^*$, and the regime’s strength, $\theta$, to be less than 1. The ex ante probability of regime change is then $\Phi_{RC}^{PC} = P (x \leq x_{PC}^* \cap \theta \leq 1)$. In the analysis below, I focus on the case where the negative impact of increasing $\sigma_\varepsilon^2$ on $\Phi_{RC}^{PC}$ is largest:

**Lemma 4.** $\frac{\partial}{\partial \sigma_\varepsilon^2} \Phi_{RC}^{PC}$ is minimised for all $\sigma_\varepsilon^2$ when $\bar{\theta} = 1$ and $c = \frac{1}{2}$.

This is a very useful result here since, in general, calculating $\Phi_{RC}^{PC}$ is non-trivial. In fact, a closed form solution in fact exists only for one specific case—and it turns out to be precisely the one identified in Lemma 4, i.e. $\bar{\theta} = 1$ and $c = \frac{1}{2}$. While being restrictive, the analysis of the special case thus remains insightful; it fixes our attention on the scenario where the negative impact of increasing $\sigma_\varepsilon^2$ on $\Phi_{RC}^{PC}$ is most prominent. Furthermore, it can also reveal interesting observations about how the regime’s incentives to increase noise are affected by the agents’ ability to coordinate actions and aggregate their information.

The following lemma proves useful here:

**Lemma 5.** Suppose $\bar{\theta} = 1$ and $c = \frac{1}{2}$. The probability of regime change in the perfect coordination benchmark with no information aggregation is:

$$\Phi_{RC}^{PC} = \frac{1}{4} + \frac{\arcsin \left( \sqrt{\frac{\sigma_\varepsilon^2}{\sigma_\theta^2 + \sigma_\varepsilon^2}} \right)}{2\pi}.$$  

(17)

Moreover, (i) $\frac{\partial \Phi_{RC}^{PC}}{\partial \sigma_\varepsilon^2} < 0 \forall \sigma_\theta^2, \sigma_\varepsilon^2 > 0$, (ii) $\lim_{\sigma_\varepsilon^2 \to 0} \frac{\partial \Phi_{RC}^{PC}}{\partial \sigma_\varepsilon^2} = -\infty$, and (iii) $\lim_{\sigma_\varepsilon^2 \to \infty} \frac{\partial \Phi_{RC}^{PC}}{\partial \sigma_\varepsilon^2} = 0$.

---

28Generically, there are no closed form solutions for $F (x) = P (X \leq x)$, where $X \sim N (\mu, \Sigma)$. There is, however, an analytical solution for the standardised bivariate normal distribution in the special case where $x = (0, 0)$. Setting $\bar{\theta} = 1$ and $c = \frac{1}{2}$ allows me to use it.
By contrast, when $\overline{\theta} = 1$ and $c = \frac{1}{2}$, the probability of regime change when the agents are imperfectly coordinated (i.e. in the global game) is $\Phi \left( \left( \theta^* - \overline{\theta} \right) / \sqrt{\sigma_\theta^2} \right)$, where

$$\theta^* = \Phi \left( \frac{\sqrt{\sigma_\varepsilon^2}}{\sigma_\theta^2} \left( \theta^* - 1 \right) \right),$$

(18)

which we obtain by substituting $\overline{\theta} = 1$ and $c = \frac{1}{2}$ into (7).

It is worth noting here that parts (i)-(iii) of Lemma 5 are also true for the case where the agents are imperfectly coordinated, which can be seen by inspecting Lemma 2 for $\overline{\theta} = 1$ and $c = \frac{1}{2}$. Thus, on a broad level, the regime’s incentives to increase noise are similar whether or not the agents are perfectly coordinated.

The differences become apparent, however, when we consider the disparities between $\Phi_{RC} (\theta^*)$ and $\Phi_{PC}^{RC}$ for a range of values of $\sigma_\varepsilon^2$:

**Proposition 5.** Suppose $\overline{\theta} = 1$ and $c = \frac{1}{2}$. If $\sigma_\theta^2$ is sufficiently low, then $\Phi_{RC} (\theta^*) < \Phi_{PC}^{RC}$ for all values of $\sigma_\varepsilon^2$. Otherwise, $\Phi_{RC} (\theta^*) < \Phi_{PC}^{RC}$ for sufficiently low values of $\sigma_\varepsilon^2$ and $\Phi_{RC} (\theta^*) > \Phi_{PC}^{RC}$ for sufficiently high values of $\sigma_\varepsilon^2$.

Proposition 5 implies that when $\theta$ is distributed tightly, i.e. $\sigma_\theta^2$ is low, then for a given $\sigma_\varepsilon^2$ the regime is always more likely to survive when agents are imperfectly coordinated rather than when they are perfectly coordinated but cannot aggregate their collective information. The situation is different, however, when $\theta$ is dispersed, i.e. $\sigma_\theta^2$ is high. Then the regime is more likely to survive with imperfectly coordinated agents when $\sigma_\varepsilon^2$ is low, but when $\sigma_\varepsilon^2$ becomes very high, the regime is actually better off with agents who are perfectly coordinated but their ability to aggregate their collective information is impeded.

In order to better understand the results of Proposition 5, consider the limiting cases where $\sigma_\varepsilon^2 \to 0$ and $\sigma_\varepsilon^2 \to \infty$. When agents are imperfectly coordinated, we have that

$$\lim_{\sigma_\varepsilon^2 \to 0} \Phi_{RC} (\theta^*) \mid_{\overline{\theta}=1,c=\frac{1}{2}} = \Phi \left( \frac{-1}{2 \sqrt{\sigma_\theta^2}} \right) \quad \text{and} \quad \lim_{\sigma_\varepsilon^2 \to \infty} \Phi_{RC} (\theta^*) \mid_{\overline{\theta}=1,c=\frac{1}{2}} = \Phi \left( \frac{-1}{\sqrt{\sigma_\theta^2}} \right),$$

(19)

while in the perfect coordination benchmark with no information aggregation, the following holds:

$$\lim_{\sigma_\varepsilon^2 \to 0} \Phi_{PC}^{RC} = \frac{1}{2} \quad \text{and} \quad \lim_{\sigma_\varepsilon^2 \to \infty} \Phi_{PC}^{RC} = \frac{1}{4}.$$

(20)

Note here that the upper limit for the probability of regime change is $\frac{1}{2}$. If the agents were perfectly coordinated and could perfectly aggregate their collective information, they would attack whenever $\theta \leq 1$, and therefore, given that we set $\overline{\theta} = 1$, the probability
of regime change would be $\Phi \left( \frac{(1 - \bar{\theta})}{\sqrt{\sigma_{\bar{\theta}}^2}} \right) = \frac{1}{2}$. This is the same as $\lim_{\sigma_\varepsilon^2 \to 0} \Phi_{RC}^{PC}$. Hence, unsurprisingly, when the private signals become arbitrarily precise, the probability of regime change in the perfect coordination benchmark does not depend on the agents’ abilities to aggregate their collective information. On the other hand, when $\sigma_\varepsilon^2 \to 0$ and the agents are imperfectly coordinated, the probability of regime change is always less than $\frac{1}{2}$.

(i) strength of the regime, $\theta$, is distributed tightly, i.e. $\sigma_\theta^2$ is low;

(ii) strength of the regime, $\theta$, is dispersed, i.e. $\sigma_\theta^2$ is high.

Figure 5: The probability of regime change in the global game, $\Phi_{RC}$, and in the perfect coordination benchmark with no information aggregation, $\Phi_{RC}^{PC}$, as a function of the noise in the agents’ private signals, $\sigma_\varepsilon^2$.

More generally, when the “large” agent’s signal reveals the true strength of the regime,
\( \theta \), almost perfectly, the role of information aggregation disappears, and therefore the limiting case \( \sigma^2 \varepsilon \rightarrow 0 \) illustrates the role of pure coordination.

As \( \sigma^2 \varepsilon \) grows, however, information aggregation plays a more and more significant role. The fact that imperfectly coordinated agents in the global game observe many conditionally independent private signals means that collectively they have more information than the “large” agent who observes only a single signal \( x = \theta + \varepsilon \). What Proposition 5 shows is that, when the private signals are sufficiently noisy, the ability of these agents to aggregate their collective information may allow them to overthrow the regime more often than the “large” agent does. In particular, this will be the case when \( \sigma^2 \theta \) is high enough, i.e. \( \theta \) is dispersed, which would mean that not only the agent’s private signals but also their prior information about the regime’s strength is imprecise.

This suggests that when \( \sigma^2 \theta \) is high, the regime may have relatively stronger incentives to increase noise when agents are perfectly coordinated but cannot aggregate their collective information than when they are imperfectly coordinated. The converse holds when \( \sigma^2 \theta \) is sufficiently low. Figure 5 illustrates these observations.

In sum, the analysis in this section shows that coordination and information aggregation are both important—yet distinct—factors affecting collective action. In particular, the fact that citizens are better coordinated need not always weaken the regime’s incentives to manipulate information, and this will be especially true when better coordination is associated with impeded information aggregation. This could occur, for example, when there is a political leader who can make all the citizens attack the incumbent if she so desires, but she cannot observe their information about the regime’s strength. As a result, the regime may actually be more likely to survive with perfectly rather than imperfectly coordinated citizens.

5 Conclusion

In this paper, I have used the global games framework to investigate the incentives of authoritarian regimes to manipulate information, with a particular emphasis on the role of the citizens’ ability to coordinate a collective action. In the model, the regime is able to manipulate the information available to the agents by increasing the noise in their private signals.

One of the aims of this analysis has been to provide a theoretical rationale for the empirical findings of [King et al. (2013, 2014, 2016)] The first two of these papers show that, when censoring content, the Chinese regime aims to silence comments that may lead
to a collective action and is not concerned about mere criticism of the state. The last one, on the other hand, demonstrates that—when fabricating social media posts—the strategy of the Chinese regime is to distract the public rather than to engage in discussions with skeptics and try to improve their opinion of the government.

These empirical observations are consistent with a formal setting in which the regime’s objective is to remain in power, but at the same time the incumbent is aware that doing this by adding a bias to the public opinion is informationally difficult. In particular, the citizens would need to be uninformed about the regime’s manipulative actions and the incumbent would need to be better informed about its strength than the citizens are.

On the other hand, adding noise into citizens’ information provides an effective alternative even when these two conditions do not hold. Under these circumstances, the regime’s incentives to increase noise turn out to be non-monotonic in (i) the citizens’ individual cost of attacking the regime and (ii) the mean strength of the incumbent, being strongest for moderate values of these parameters.

Furthermore, the paper presents two alternative settings which aim to capture the ideas behind the two theories of King et al. (2013, 2014): the state critique theory and the theory of collective action potential. First, the model shows that when the manipulative action is not hidden and the incumbent is not privately informed about its strength, the regime has no incentives to increase noise so as to improve the citizen’s average perception of the government. Second, the model demonstrates that minimising all collective attacks by citizens may be an effective way of preventing regime change, and that this will be the case especially when (i) the citizens’ private signals about the regime’s strength are intrinsically imprecise (i.e. they have little more information than the regime), and (ii) the size of collective attacks—particularly when they are of moderate scale—strongly affects the chances of success.

Finally, the paper analyses how the citizens’ ability to coordinate their actions affects the regime’s incentives to manipulate information. A crucial role is played here by the degree to which the citizens can aggregate the information they collectively possess. If they can perfectly coordinate an attack but rely only on the signal of their leader, the regime’s incentives to increase noise may in fact be stronger than if the citizens were imperfectly coordinated but could partially aggregate their collective information. Consequently, this kind of coordination with no information aggregation may be counterproductive in the sense that it would boost the likelihood that the regime remains in power. This is especially likely to be true when the regime is expected to be strong but the prior information about it is imprecise.
As far as avenues for further research are concerned, it is worth noting that information manipulation is not only the domain of authoritarian regimes. For example, a phenomenon similar to employing pro-government online commentators is fairly common among private companies. Dellarocas (2010) offers a theoretical analysis of how firms manipulate consumer perceptions by posting costly anonymous messages that praise their products. On the empirical front, Mayzlin, Dover and Chevalier (2014) analyse two hotel review websites and determine the characteristics of firms that are most likely to write or buy fake reviews. It would be worthwhile to investigate to what extent the incentives to manipulate information are different in the market setting as this could shed further light on the incentives of authoritarian regimes.

Appendix

A.1 Omitted Proofs

Proof of Proposition 1

Proposition 1 is a standard result in global games, see, e.g., Morris and Shin (2003). According to the payoff indifference condition, an agent is indifferent between attacking the regime and not attacking it when she observes a private signal $x^*$. Hence, we must have

$$\mathbb{P}(\text{attack successful} \mid x^*) = c. \quad (21)$$

Given the structure of the equilibrium, this yields

$$\mathbb{P}(\theta \leq \theta^* \mid x^*) = c, \quad (22)$$

where the left-hand side is equal to $\Phi \left( (\theta^* - \mathbb{E}[\theta \mid x^*]) / \sqrt{\text{Var}(\theta \mid x^*)} \right)$ due to the assumptions on the distribution of $\theta$. Applying the standard results of the normal learning model, the agent’s posterior expectation of the regime’s strength, $\theta$, conditional on a given private signal, $x$, is

$$\mathbb{E}[\theta \mid x] = \bar{\theta} + \frac{\sigma^2_{\theta}}{\sigma^2_{\theta} + \sigma^2_z} \left( x - \bar{\theta} \right). \quad (23)$$

This is another way of stating (1). The more noisy the agents’ private signals are, i.e. the higher $\sigma_z^2$ is, the less informative the private signal is about $\theta$, which makes an agent put more weight on the mean strength of the regime, $\bar{\theta}$, and less on the private signal, $x$, when forming her posterior.

The conditional variance of $\theta$ given $x$ is given by:

$$\text{Var}(\theta \mid x) = \sigma^2_{\theta} \left( 1 - (\text{Corr}(\theta, x))^2 \right) \quad (24)$$

$$= \sigma^2_{\theta} \left( 1 - \left( \frac{\sigma^2_{\theta}}{\sqrt{\sigma^2_{\theta} + \sigma^2_z}} \right)^2 \right) \quad (25)$$

$$= \frac{\sigma^2_{\theta} \sigma^2_z}{\sigma^2_{\theta} + \sigma^2_z} \quad (26)$$

Adding in the results on conditional variance, the payoff indifference condition is:

$$c = \Phi \left( \sqrt{\frac{\sigma^2_{\theta} + \sigma^2_z}{\sigma^2_{\theta} \sigma^2_z}} \left( \theta^* - \bar{\theta} - \frac{\sigma^2_{\theta}}{\sigma^2_{\theta} + \sigma^2_z} (x^* - \bar{\theta}) \right) \right). \quad (27)$$
We can rearrange (27) to obtain the following expression for the “critical” value of the private signal, $x^\ast$:

$$x^*_{PI} = \frac{\sigma_\theta^2 + \sigma_\varepsilon^2 \theta^*}{\sigma_\theta^2} - \frac{\sigma_\varepsilon^2}{\sigma_\theta^2} \theta^* - \sqrt{\frac{\sigma_\varepsilon^2 (\sigma_\theta^2 + \sigma_\varepsilon^2)}{\sigma_\theta^2}} \Phi^{-1} (c) . \quad (28)$$

The critical mass condition states that, in equilibrium, whenever the true strength of strength is exactly $\theta^*$, the proportion of citizens who receive a signal $x \leq x^*$ should equal precisely $\theta^*$:

$$\theta^* = \mathbb{P} (x \leq x^* \mid \theta^*) = . \quad (29)$$

$$= \Phi \left( \frac{x^* - \mathbb{E}[x \mid \theta^*]}{\sqrt{\text{Var}(x \mid \theta^*)}} \right) \quad (30)$$

$$= \Phi \left( \frac{x^* - \theta^*}{\sqrt{\sigma_\varepsilon^2}} \right) . \quad (31)$$

where, given the distributional assumptions on $\theta$, the right-hand side is:

$$\mathbb{P} (x \leq x^* \mid \theta^*) = \Phi \left( \frac{x^* - \mathbb{E}[x \mid \theta^*]}{\sqrt{\text{Var}(x \mid \theta^*)}} \right) \quad (32)$$

$$= \Phi \left( \frac{x^* - \theta^*}{\sqrt{\sigma_\varepsilon^2}} \right) . \quad (33)$$

This leads us to the critical mass condition:

$$x^*_{CM} = \theta^* + \sqrt{\sigma_\varepsilon^2 \Phi^{-1} (\theta^*)} . \quad (34)$$

The two equilibrium conditions stated in (28) and (34) yield an equilibrium value of the critical state of the world, $\theta^*$, as stated in Proposition 1.

As far as the statement on equilibrium uniqueness is concerned, by substituting (34) into (28), we obtain the following condition:

$$U^{st} (\theta \cdot) = 1 - \Phi \left( \sqrt{\frac{\sigma_\theta^2 + \sigma_\varepsilon^2}{\sigma_\theta^2}} \left( \frac{\sigma_\varepsilon^2}{\sigma_\theta^2 + \sigma_\varepsilon^2} (\theta - \theta^*) + \frac{\sigma_\theta^2 \sqrt{\sigma_\varepsilon^2}}{\sigma_\theta^2 + \sigma_\varepsilon^2} \Phi^{-1} (\theta) \right) \right) - c \quad (35)$$

I follow here the steps of the proof for the equilibrium uniqueness condition in Angeletos, Hellwig and Pavan (2007).
\[ 1 - \Phi \left( \sqrt{\frac{\sigma_\theta^2}{\sigma_\theta^2 + \sigma_\varepsilon^2}} \left( \frac{\sqrt{\sigma_\theta^2}}{\sigma_\theta^2} (\bar{\theta} - \theta) + \Phi^{-1} (\theta) \right) \right) - c = 0. \]  

(35)

By inspection, we observe that (i) \( U^{st}(\theta; \cdot) \) is continuous and differentiable in \( \theta \in (0, 1) \), (ii) \( \lim_{\theta \to 0} U^{st}(\theta; \cdot) = 1 - \frac{c}{\bar{\theta}} > 0 \), and (iii) \( \lim_{\theta \to 1} U^{st}(\theta; \cdot) = -c < 0 \). This implies that a solution to \( U^{st}(\theta; \cdot) = 0 \) always exists. Furthermore, note that

\[
\frac{\partial U^{st}(\theta; \cdot)}{\partial \theta} = -\sqrt{\frac{\sigma_\theta^2}{\sigma_\theta^2 + \sigma_\varepsilon^2}} \phi (\cdot) \left( \frac{1}{\phi(\Phi^{-1}(\theta))} - \frac{\sigma_\varepsilon^2}{\sigma_\theta^2} \right). 
\]

(36)

By the properties of the normal distribution, \( \min_{\theta \in (0, 1)} \frac{1}{\phi(\Phi^{-1}(\theta))} = \sqrt{2\pi} \). This means that a necessary and sufficient condition for \( U^{st}(\theta; \cdot) \) to be monotonic in \( \theta \) is that \( \sqrt{2\pi} \geq \frac{\sigma_\varepsilon^2}{\sigma_\theta^2} \), which can be rearranged to yield the statement on equilibrium uniqueness in Proposition 1.

Finally, it is worth noting that I have assumed so far the productivity of attacks is given by function \( f (A) = A \), i.e the regime is overthrown if and only if \( \theta \leq f (A) = A \). Consider now the general case where the productivity is given by \( f (A) \), where \( f : [0, 1] \to \mathbb{R} \) is an increasing, \( C^1 \) function, with \( |f'(A)| \leq F \) for some \( F \in \mathbb{R}_+ \). Then an equilibrium value of \( \theta^* \) (the equivalent of the one stated in Proposition 1 for \( f (A) = A \)) is:

\[
\theta^* = f \left( \Phi \left( \frac{\sqrt{\sigma_\theta^2}}{\sigma_\theta^2} (\theta^* - \bar{\theta}) - \frac{\sqrt{\sigma_\theta^2 + \sigma_\varepsilon^2}}{\sigma_\theta^2} \Phi^{-1} (c) \right) \right). 
\]

(37)

Furthermore, the equilibrium uniqueness condition can then be written as: \( \sigma_\varepsilon^2 \leq \frac{2\pi (\sigma_\theta^2)^2}{F} \).

**Proof of Lemma 1**

For part (i), note that by partially differentiating \( \theta^* \) from (7) with respect to \( \bar{\theta} \), we obtain:

\[
\frac{\partial \theta^*}{\partial \bar{\theta}} = \frac{-\phi(\cdot) \frac{\sigma_\varepsilon^2}{\sigma_\theta^2}}{1 - \phi(\cdot) \frac{\sigma_\varepsilon^2}{\sigma_\theta^2}}. 
\]

(38)

where \( \phi (\cdot) = \phi \left( \frac{\sqrt{\sigma_\theta^2}}{\sigma_\theta^2} (\theta^* - \bar{\theta}) - \frac{\sqrt{\sigma_\theta^2 + \sigma_\varepsilon^2}}{\sigma_\theta^2} \Phi^{-1} (c) \right) \). The numerator is clearly negative. By the properties of the normal distribution, the maximum value of \( \phi (\cdot) \) is \( 1/\sqrt{2\pi} \). The equilibrium uniqueness condition stated in Proposition 1 ensures that the denominator is positive. The probability of regime survival is \( \Phi_{RC} = \Phi \left( \frac{\theta^* - \bar{\theta}}{\sqrt{\sigma_\theta^2}} \right) \). Thus, \( \frac{\partial \Phi_{RC}}{\partial \theta} < 0 \) if and
only if $\frac{\partial \theta^*}{\partial \bar{\theta}} > 0$.

The steps for part (ii) are analogous. First, partially differentiate the expression for $\theta^*$ in (7) with respect to $c$:

$$\frac{\partial \theta^*}{\partial c} = -\phi(\cdot) \sqrt{\frac{\sigma_\theta^2 + \sigma_\varepsilon^2}{\sigma_\theta^2}} \frac{\partial \Phi^{-1}(c)}{\partial (c)} \left( \theta^* - \bar{\theta} \right),$$

(39)

where $\phi(\cdot) = \phi \left( \sqrt{\frac{\sigma_\theta^2}{\sigma_\varepsilon^2}} \left( \theta^* - \bar{\theta} \right) - \sqrt{\frac{\sigma_\theta^2 + \sigma_\varepsilon^2}{\sigma_\theta^2}} \Phi^{-1}(c) \right)$. The numerator is negative, whereas the denominator is again positive by the properties of normal distribution and the condition for equilibrium uniqueness stated in Proposition 1.

**Proof of Lemma 2**

By partially differentiating the expression for $\theta^*$ in (7) with respect to $\sigma_\varepsilon^2$, we obtain:

$$\frac{\partial \theta^*}{\partial \sigma_\varepsilon^2} = \frac{\phi(\cdot) \left( \frac{1}{2\sigma_\varepsilon^2} \left( \theta^* - \bar{\theta} \right) \right) - \frac{1}{2\sigma_\varepsilon^2} \Phi^{-1}(c)}{1 - \phi(\cdot) \sqrt{\frac{\sigma_\theta^2 + \sigma_\varepsilon^2}{\sigma_\theta^2}}},$$

(40)

where $\phi(\cdot) = \phi \left( \sqrt{\frac{\sigma_\theta^2}{\sigma_\varepsilon^2}} \left( \theta^* - \bar{\theta} \right) - \sqrt{\frac{\sigma_\theta^2 + \sigma_\varepsilon^2}{\sigma_\theta^2}} \Phi^{-1}(c) \right)$.

Note that by the properties of the normal distribution, the maximum value of $\phi(\cdot)$ is $1/\sqrt{2\pi}$. As long as the equilibrium uniqueness condition stated in Proposition 1 is satisfied, this ensures that the denominator is positive. However, the sign of the numerator is ambiguous and is strictly positive if and only if

$$\frac{1}{\sigma_\theta^2} \left( \theta^* - \bar{\theta} \right) - \frac{1}{\sqrt{\sigma_\theta^2 (\sigma_\theta^2 + \sigma_\varepsilon^2)}} \Phi^{-1}(c) > 0.$$  

(41)

Given that $\Phi_{RC} = \Phi \left( \frac{\theta^* - \bar{\theta}}{\sqrt{\sigma_\theta^2}} \right)$, and so $\frac{\partial \Phi_{RC}}{\partial \sigma_\varepsilon^2} < 0$ if and only if $\frac{\partial \theta^*}{\partial \sigma_\varepsilon^2} > 0$, this can be rearranged to yield the condition in 2.

**Proof of Proposition 2**

By Lemma 2, the probability of regime change is decreasing in the variance of noise in agents’ private signals, $\frac{\partial \Phi_{RC}}{\partial \sigma_\varepsilon^2} < 0$, if and only if the critical strength of the regime is low enough:

$$\theta^* < T_{RC},$$

(42)
where \( T_{RC} = \bar{\theta} + \sqrt{\frac{\sigma^2_{\sigma^2} + \sigma^2_\varepsilon}{\sigma^2_\sigma}} \Phi^{-1}(c) \).

To see why \( \bar{\theta} \) (or \( c \)) must be high enough for the regime to have incentives to increase \( \sigma^2_{\varepsilon} \), note that the left-hand side of (42) is decreasing in \( \bar{\theta} \) and in \( c \) (by Lemma 1), while the right-hand side is increasing in \( \bar{\theta} \) and in \( c \). Hence, a sufficiently high value of \( \bar{\theta} \) (or \( c \)) ensures that the condition in (42) is satisfied.

It remains to show that the regime would not choose to increase \( \sigma^2_{\varepsilon} \), i.e. \( \frac{\partial U_R}{\partial \sigma^2_{\varepsilon}} \leq 0 \), if \( \bar{\theta} \) (or \( c \)) is exceedingly high. The key here is to use two properties of the normal distribution: (i) \( \phi'(x) = x\phi(x) \) and (ii) \( \lim_{x \to -\infty} \phi'(x) = 0 \). The claim here is that:

\[
\lim_{\theta \to \infty} \frac{\partial \theta^*}{\partial \sigma^2_{\varepsilon}} = \lim_{\theta \to \infty} \frac{\phi(\cdot) \left( \frac{1}{2\sigma^2_\sigma} \frac{\partial}{\partial \theta} \left( \theta^* - \bar{\theta} \right) - \frac{1}{2\sqrt{\sigma^2_\sigma} \sigma^2_{\sigma} + \sigma^2_\varepsilon} \Phi^{-1}(c) \right)}{1 - \phi(\cdot) \sqrt{\frac{\sigma^2_{\sigma} + \sigma^2_\varepsilon}{\sigma^2_\sigma}}} = 0,
\]

where \( \phi(\cdot) = \phi \left( \sqrt{\frac{\sigma^2_{\sigma}}{\sigma^2_\sigma}} \left( \theta^* - \bar{\theta} \right) - \sqrt{\frac{\sigma^2_{\sigma} + \sigma^2_\varepsilon}{\sigma^2_\sigma}} \Phi^{-1}(c) \right) \), and an analogous statement holds for \( \lim_{c \to 1} \frac{\partial \sigma^2_{\varepsilon}}{\partial \sigma^2_{\varepsilon}} \).

To see this, note that if \( \bar{\theta} = \frac{1}{2} \) and \( c = \frac{1}{2} \), then \( \theta^* = \frac{1}{2} \) for all values of \( \sigma^2_{\varepsilon} \). The numerator is then zero for all \( \sigma^2_{\varepsilon} \), while the denominator is strictly positive as long as the equilibrium uniqueness condition in Proposition 1 is satisfied, which clearly leads to (43). If, on the other hand, we have that \( \bar{\theta} \neq \frac{1}{2} \) or \( c \neq \frac{1}{2} \), then \( \theta^* \) approaches 1 as \( \bar{\theta} \to \infty \) (or as \( c \to 1 \)). In that case, the result in (43) follows from the two properties of the normal distribution, \( \phi'(x) = x\phi(x) \) and \( \lim_{x \to -\infty} \phi'(x) = 0 \).

Since \( \Phi_{RC} = \Phi \left( \frac{\theta^* - \bar{\theta}}{\sqrt{\sigma^2_{\sigma}}} \right) \), this implies that \( \lim_{\theta \to \infty} \frac{\partial \Phi_{RC}}{\partial \sigma^2_{\varepsilon}} = 0 \) and \( \lim_{c \to 1} \frac{\partial \Phi_{RC}}{\partial \sigma^2_{\varepsilon}} = 0 \).

Thus, given that \( C'(\cdot) > 0 \), the regime will never choose to increase \( \sigma^2_{\varepsilon} \) if \( \bar{\theta} \) (or \( c \)) takes a sufficiently high value.

Hence, for a given value of \( c \), if \( \lambda \) is small enough (so that the role of costs in the regime’s utility function is sufficiently small), there exists an interval of moderate values of \( \bar{\theta} \), \( [\bar{\theta}_1, \bar{\theta}_2] \), for which the regime chooses to increase \( \sigma^2_{\varepsilon} \): \( \frac{\partial U_R}{\partial \sigma^2_{\varepsilon}} > 0 \). Similarly, for a given value of \( \bar{\theta} \), if \( \lambda \) is small enough, there exists an interval of moderate values of \( c \), \( [c_1, c_2] \), for which the regime chooses to increase \( \sigma^2_{\varepsilon} \): \( \frac{\partial U_R}{\partial \sigma^2_{\varepsilon}} > 0 \).

While the arguments above already prove the statement in Proposition 2 to provide further insights, I also present here results on the behaviour of \( \theta^* \) and \( \Phi_{RC} \) as a function of \( \sigma^2_{\varepsilon} \). I denote \( \theta^*_0 = \lim_{\sigma^2_{\varepsilon} \to 0} \theta^* = 1 - c \) for the ease of notation, and analyse three cases: \( c < \frac{1}{2}, \; c = \frac{1}{2} \), and \( c > \frac{1}{2} \). Recall also from (8) that an increase (decrease) in \( \theta^* \) always implies an increase (decrease) in the probability of regime change, \( \Phi_{RC} \).
Lemma 6. Suppose that $c < \frac{1}{2}$. Then there exists $\tilde{\sigma}_e^2$ and $\omega' \in \left[\tilde{\theta}, \hat{\theta} + \sqrt{\tilde{\sigma}_e^2 \Phi^{-1}(c)}\right]$ such that:

1. if $\theta^* < \omega$, then $\frac{\partial \theta}{\partial \sigma_e^2} < 0 \forall \sigma_e^2$;
2. if $\theta^* \in \left[\omega, \tilde{\theta}\right)$, then $\frac{\partial \theta}{\partial \sigma_e^2} < 0$ for $\sigma_e^2 < \tilde{\sigma}_e^2$ and $\frac{\partial \theta}{\partial \sigma_e^2} < 0$ for $\sigma_e^2 > \tilde{\sigma}_e^2$;
3. if $\theta^* > \hat{\theta}$, then $\frac{\partial \theta}{\partial \sigma_e^2} > 0 \forall \sigma_e^2$.

To see this, recall that $T_{RC} = \tilde{\theta} + \sqrt{\tilde{\sigma}_e^2 \Phi^{-1}(c)}$ and note that, given $c < \frac{1}{2}$, $T_{RC}$ is strictly decreasing in $\sigma_e^2$ with $\lim_{\sigma_e^2 \to 0} T_{RC} = \tilde{\theta}$ and a lower bound of $\tilde{\theta} + \sqrt{\hat{\sigma}_e^2 \Phi^{-1}(c)}$. Thus, the condition that $\theta^* > \hat{\theta}$ ensures that $\frac{\partial \theta}{\partial \sigma_e^2} > 0$ for all $\sigma_e^2$.

Now, analyse the case where $\hat{\theta} + \sqrt{\tilde{\sigma}_e^2 \Phi^{-1}(c)} < \theta^* < \tilde{\theta}$. Note that the derivative $\frac{\partial \theta}{\partial \sigma_e^2}$ is negative for sufficiently small $\sigma_e^2$. If $\theta^*$ is small enough, say $\theta^* < \omega$, then as we increase $\sigma_e^2$, $\theta^*$ decreases and reaches $\tilde{\theta} + \sqrt{\tilde{\sigma}_e^2 \Phi^{-1}(c)}$ and therefore there is no $\sigma_e^2$ such that $\theta^* (\sigma_e^2) = T_{RC} (\sigma_e^2)$. As a result, $\theta^*$ is decreasing in $\sigma_e^2$ for all $\sigma_e^2$. On the other hand, if $\theta^*$ is high enough, $\theta^* \geq \omega$, then there exists $\sigma_e^2$ such that $\theta^* (\sigma_e^2) = T_{RC} (\sigma_e^2)$, and hence $\theta^*$ is decreasing in $\sigma_e^2$ for $\sigma_e^2 < \tilde{\sigma}_e^2$ and increasing in $\sigma_e^2$ for $\sigma_e^2 > \tilde{\sigma}_e^2$.

Lemma 7. Suppose that $c = \frac{1}{2}$. Then:

1. if $\theta^* < \tilde{\theta}$, then $\frac{\partial \theta}{\partial \sigma_e^2} < 0 \forall \sigma_e^2$;
2. if $\theta^* = \tilde{\theta}$, then $\frac{\partial \theta}{\partial \sigma_e^2} = 0 \forall \sigma_e^2$;
3. if $\theta^* > \hat{\theta}$, then $\frac{\partial \theta}{\partial \sigma_e^2} > 0 \forall \sigma_e^2$.

In this knife-edge case, we have $T_{RC} = \tilde{\theta}$, i.e. it does not vary with $\sigma_e^2$; which greatly simplifies the analysis. The condition from Lemma 6 is then that $\frac{\partial \theta}{\partial \sigma_e^2} < (>) 0$ if and only if $\theta^* < (>) \tilde{\theta}$, which yields the statement in the lemma.

Lemma 8. Suppose that $c > \frac{1}{2}$. Then there exists $\tilde{\sigma}_e^2$ and $\omega \in \left[\tilde{\theta}, \hat{\theta} + \sqrt{\tilde{\sigma}_e^2 \Phi^{-1}(c)}\right]$ such that:

1. if $\theta^* \leq \tilde{\theta}$, then $\frac{\partial \theta}{\partial \sigma_e^2} < 0 \forall \sigma_e^2$;
2. if $\theta^* \in \left[\tilde{\theta}, \omega\right)$, then $\frac{\partial \theta}{\partial \sigma_e^2} > 0$ for $\sigma_e^2 < \tilde{\sigma}_e^2$ and $\frac{\partial \theta}{\partial \sigma_e^2} < 0$ for $\sigma_e^2 > \tilde{\sigma}_e^2$;
3. if $\theta^* > \omega$, then $\frac{\partial \theta}{\partial \sigma_e^2} > 0 \forall \sigma_e^2$.

The analysis here is analogous to that in Lemma 6. First, note that, given $c > \frac{1}{2}$, $T_{RC}$ is strictly increasing in $\sigma_e^2$ with $\lim_{\sigma_e^2 \to 0} T_{RC} = \tilde{\theta}$ and an upper bound of $\tilde{\theta} + \sqrt{\hat{\sigma}_e^2 \Phi^{-1}(c)}$. Thus, the condition that $\theta^* \geq \tilde{\theta} + \sqrt{\hat{\sigma}_e^2 \Phi^{-1}(c)}$ ensures that $\frac{\partial \theta}{\partial \sigma_e^2} > 0$ for all $\sigma_e^2$.

Now, analyse the case where $\tilde{\theta} < \theta^* < \tilde{\theta} + \sqrt{\tilde{\sigma}_e^2 \Phi^{-1}(c)}$. Note that the derivative $\frac{\partial \theta}{\partial \sigma_e^2}$ is positive for sufficiently small $\sigma_e^2$. If $\theta^*$ is high enough, say $\theta^* > \omega$, then as we increase $\sigma_e^2$, $\theta^*$ reaches $\tilde{\theta} + \sqrt{\tilde{\sigma}_e^2 \Phi^{-1}(c)}$ and so there is no $\sigma_e^2$ such that $\theta^* (\sigma_e^2) = T_{RC} (\sigma_e^2)$. As a result,
\( \theta^* \) is increasing in \( \sigma^2 \) for all \( \sigma^2 \). On the other hand, if \( \theta^*_0 \) is low enough, \( \theta^*_0 \leq \omega' \), then there exists \( \sigma^2 \) such that \( \theta^*_0 (\sigma^2) = T_{RC} (\sigma^2) \), and hence \( \theta^* \) will be increasing in \( \sigma^2 \) for \( \sigma^2 < \sigma_0^2 \) and decreasing in \( \sigma^2 \) for \( \sigma^2 > \sigma_0^2 \).

Finally, since with \( c > \frac{1}{2} \) we have that \( \bar{\theta} \leq T_{RC} \), the condition that \( \theta^*_0 < \bar{\theta} \) clearly ensures that \( \theta^* \) is decreasing in \( \sigma^2 \) for all \( \sigma^2 \).

**Proof of Proposition 3**

Using the results of the normal learning model, we can rewrite the regime’s utility function as:

\[
U_{\text{SC}} (\sigma^2) = \mathbb{E}_{\theta, \varepsilon} \left[ \left( \bar{\theta} + \frac{\sigma^2}{\sigma^2 + \sigma^2} (x - \bar{\theta}) \right) - \lambda C (\sigma^2) \right].
\]

By the law of iterated expectations, this simplifies to \( U_{\text{SC}} (\sigma^2) = \bar{\theta} - \lambda C (\sigma^2) \). Regardless of the value of \( \sigma^2 \) chosen by the regime, the ex ante expectation of an agent’s posterior always equals \( \bar{\theta} \). The regime’s marginal benefit from increasing \( \sigma^2 \) is thus always zero, while the marginal cost, \( C' (\cdot) \), is strictly positive. It follows that the regime would never choose to increase the noise, \( \sigma^2 \).

**Proof of Lemma 3**

Note that

\[
\frac{\partial}{\partial \sigma^2} \mathbb{E} [A (\theta^*)] = \phi (\cdot) \left( \frac{\partial \sigma^2}{\partial \sigma^2} \sqrt{\sigma^2 + \sigma^2} + \left( \theta^* - \bar{\theta} \right) \frac{1}{2 \sigma^2 \sqrt{\sigma^2 + \sigma^2}} - \frac{1}{2 \sigma^2} \Phi^{-1} (c) \right),
\]

where \( \phi (\cdot) = \phi \left( \left( \theta^* - \bar{\theta} \right) \sqrt{\sigma^2 + \sigma^2} \Phi^{-1} (c) \right) \). This is positive if and only if the term in brackets is positive, which can rearranged to yield \( \theta^* < T_{A} \), where \( T_{A} \) is given by \( 13 \).

While this already proves the statement in Lemma \( 3 \), I present here supplementary results for further intuition. These show that, on a broad level, the regime’s incentives to minimise collective attacks are similar to those to prevent regime change, despite the difference between \( T_{RC} \) from \( 9 \) and \( T_{A} \) from \( 13 \). The following is an equivalent of Lemma \( 1 \).

**Lemma 9.** The expected size of the collective attack, \( \mathbb{E} [A (\theta^*)] \), is decreasing in:

(i) the mean strength of the regime, \( \bar{\theta} \), i.e. \( \frac{\partial \mathbb{E} [A (\theta^*)]}{\partial \bar{\theta}} < 0 \) \( \forall \bar{\theta} \);

(ii) the agents’ individual cost of attacking, \( c \), i.e. \( \frac{\partial \mathbb{E} [A (\theta^*)]}{\partial c} < 0 \) \( \forall c \).
Proof. Note that
\[
\frac{\partial \mathbb{E}[A(\theta^*)]}{\partial \theta} = \phi(\cdot) \left( \frac{\partial \theta^*}{\partial \theta} - 1 \right) \sqrt{\frac{\sigma_\theta^2}{\sigma_\theta^2 + \sigma_2^2}} ,
\]
where \( \phi(\cdot) = \phi\left( \left( \theta^* - \hat{\theta} \right) \sqrt{\frac{\sigma_\theta^2 + \sigma_2^2}{\sigma_\theta^2}} - \sqrt{\frac{\sigma_2^2}{\sigma_\theta^2}} \Phi^{-1}(c) \right) \). This is negative \( \forall \hat{\theta} \) since \( \frac{\partial \theta^*}{\partial \hat{\theta}} < 0 \) from Lemma 1.

Similarly,
\[
\frac{\partial \mathbb{E}[A(\theta^*)]}{\partial c} = \phi(\cdot) \left( \frac{\partial \theta^*}{\partial c} \sqrt{\frac{\sigma_\theta^2 + \sigma_2^2}{\sigma_2^2}} - \sqrt{\frac{\sigma_2^2}{\sigma_\theta^2}} \frac{d\Phi^{-1}(c)}{dc} \right) ,
\]
where \( \phi(\cdot) = \phi\left( \left( \theta^* - \hat{\theta} \right) \sqrt{\frac{\sigma_\theta^2 + \sigma_2^2}{\sigma_\theta^2}} - \sqrt{\frac{\sigma_2^2}{\sigma_\theta^2}} \Phi^{-1}(c) \right) \), is negative \( \forall c \) since \( \frac{d\Phi^{-1}(c)}{dc} > 0 \) and we know from Lemma 1 that \( \frac{\partial \theta^*}{\partial \hat{\theta}} < 0 \).

Unsurprisingly, the intuition behind the results in Lemma 1 and Lemma 9 is similar. Note from (4) that an increase in either \( \hat{\theta} \) or \( c \) leads to a decrease in \( x^* \), i.e. the critical value of the private signal that makes an agent indifferent between attacking the regime and not doing so. This means that fewer agents choose to attack the regime and, as long as the equilibrium uniqueness condition is satisfied, also to regime change being less likely.

Looking at Lemma 2 and Lemma 3, we can see that the impact of an increase in \( \sigma_2^2 \) on the size of the collective attack is also similar to that on the probability of regime change, yet a little more complex due to an indirect effect captured by the third term in (13). But since this indirect effect turns out to be second order in the limit as \( \sigma_2^2 \rightarrow 0 \) and approaches zero as \( \sigma_2^2 \rightarrow \infty \), the following is true (which is parallel to Proposition 2):

**Proposition 6.** Suppose the regime’s objective function is \( U^{CA}_R(\sigma_2^2) := \mathbb{E}[1 - A(\theta^* (\sigma_2^2))] - \lambda C(\sigma_2^2 - \sigma_\epsilon^2) \). Then the regime’s incentives to increase the variance of noise in the agents’ private signals, \( \sigma_2^2 \), are non-monotonic with respect to the mean strength of the regime, \( \hat{\theta} \), and the agents’ individual cost of attacking the regime, \( c \):

(i) For sufficiently low and sufficiently high values of \( \hat{\theta} \) (or \( c \)), the regime chooses not to raise \( \sigma_2^2 \);

(ii) For moderate values of \( \hat{\theta} \) (or \( c \)), the regime chooses to raise \( \sigma_2^2 \) only if \( \lambda \) is small enough.

Proof. Note from (45) that if \( \hat{\theta} \) is sufficiently high then the long term in brackets is negative. To see this, note that (i) \( \frac{\partial \theta^*}{\partial \sigma_2^2} \) is negative if \( \hat{\theta} \) is sufficiently high (by Lemma 2), and (ii) \( \frac{\partial \theta^*}{\partial \hat{\theta}} < 0 \ \forall \hat{\theta} \) (by Lemma 1) so the second term in the long bracket is also decreasing in \( \hat{\theta} \). The same line of argument applies to the analysis with respect to \( c \). Thus, if \( \theta \) (or \( c \)) is sufficiently high, the expected size of the collective attack, \( \mathbb{E}[A(\theta^*)] \), is decreasing in \( \sigma_2^2 \).
A converse argument can be made to show that $\mathbb{E} [A (\theta^*)]$ is increasing in $\sigma^2$ if $\tilde{\theta}$ (or $c$) is sufficiently low.

At the same time, however, we know from the proof of Proposition 2 that $\lim_{\tilde{\theta} \to \infty} \frac{\partial A^*}{\partial \sigma^2} = 0$, so using the properties of the normal distribution, $\phi' (x) = x \phi (x)$ and $\lim_{x \to \pm \infty} \phi' (x) = 0$, it follows that $\lim_{\tilde{\theta} \to \infty} \frac{\partial [\mathbb{E} [A (\theta^*)]]}{\partial \sigma^2} = 0$. Given that information manipulation is costly, $C' (\sigma^2) > 0$, the regime will thus choose not to increase $\sigma^2$ if $\tilde{\theta}$ takes extremely high values. Again, the same applies to $c$. 

Analogously to Proposition 2 it follows that if $\lambda$ is small enough, i.e. if the cost of increasing $\sigma^2$ is sufficiently insignificant relative to the benefit from smaller size of attacks, then there will exist a range of moderate values of $\tilde{\theta}$ (and $c$) where it is optimal for the regime to increase $\sigma^2$ for the purposes of minimising collective attacks: $\frac{\partial [\mathbb{E} [A]]}{\partial \sigma^2} > 0$.

**Proof of Proposition 4**

For part (i), note that:

$$\lim_{\sigma^2 \to 0} \theta^* = \lim_{\sigma^2 \to 0} \Phi \left( \sqrt{\frac{\sigma^2}{\bar{\sigma}^2}} (\theta^* - \bar{\theta}) - \sqrt{\frac{\sigma^2}{\bar{\sigma}^2}} \Phi^{-1} (c) \right)$$

$$= \Phi \left( -\Phi^{-1} (c) \right)$$

$$= 1 - c. \quad (48)$$

First, consider the case where $c > \frac{1}{2}$ and $\bar{\theta} > 1 - c = \lim_{\sigma^2 \to 0} \theta^*$. Then $\Phi^{-1} (c) > 0$ and, by Lemma 2, $\frac{\partial \theta^*}{\partial \sigma^2} < 0$. This implies that both the second and the third term in the expression for $T_A$ (see (13) in Lemma 3) are positive, and so if $\theta^* < \bar{\theta}$, then it must also be the case that $\theta^* < T_A$, and hence $\frac{\partial A (\theta^*)}{\partial \sigma^2} < 0$. Thus, when $c > \frac{1}{2}$ and $\bar{\theta} > 1 - c$, the size of the collective attack (like the probability of regime change) is decreasing in $\sigma^2$.

An analogous argument holds for the case where $c < \frac{1}{2}$ and $\bar{\theta} < 1 - c$, in which case we have $\frac{\partial A (\theta^*)}{\partial \sigma^2} > 0$.

Second, note that when $c > \frac{1}{2}$ and $\bar{\theta} < 1 - c = \lim_{\sigma^2 \to 0} \theta^*$, the second and the third term in the expression for $T_A$ (see (13) in Lemma 3) have opposing signs, and they both approach infinity as $\sigma^2 \to 0$. However, using L'Hôpital’s rule, we can see that the third term is second-order in the limit $\sigma^2 \to 0$: the second term approaches $+\infty$ at rate $1/\sqrt{\sigma^2}$, while the third term approaches $-\infty$ at rate $g (\sigma^2) / \sqrt{\sigma^2}$, where the numerator is decreasing as $\sigma^2$ gets arbitrarily close to 0. Hence, the second term “dominates” the third and we have that $\frac{\partial A (\theta^*)}{\partial \sigma^2} < 0$ as long as $\bar{\theta} > \tau_{A,0} (c)$, where $\tau_{A,0} (c)$ is a steeper function of $c$ than $\lim_{\sigma^2 \to 0} \theta^* = 1 - c$ is: $\tau_{A,0} (c) < -1$. 

39
For part (ii), note that as $\sigma_{\varepsilon}^2 \to \infty$, the second term in the expression for $T_A$ goes to $\sqrt{\sigma_{\varepsilon}^2} \Phi^{-1}(c)$, and so does the third term. To see the latter, use the properties of the normal distribution, $\phi'(x) = x\phi(x)$ and $\lim_{x \to \pm\infty} \phi'(x) = 0$, and note that the third term approaches $\pm\infty$ at rate $\phi'\left(\sqrt{\sigma_{\varepsilon}^2} h(\sigma_{\varepsilon}^2)\right) \sqrt{\sigma_{\varepsilon}^2} h(\sigma_{\varepsilon}^2)$, where $h(\sigma_{\varepsilon}^2)$ is increasing in $\sigma_{\varepsilon}^2$. Consequently, $\lim_{\sigma_{\varepsilon}^2 \to \infty} T_A = \bar{\theta} + \sqrt{\sigma_{\varepsilon}^2} \Phi^{-1}(c)$. At the same time, we clearly also have $\lim_{\sigma_{\varepsilon}^2 \to \infty} T_{RC} = \bar{\theta} + \sqrt{\sigma_{\varepsilon}^2} \Phi^{-1}(c)$, since the second term in the expression for $T_{RC}$ in (6) also approaches $\sqrt{\sigma_{\varepsilon}^2} \Phi^{-1}(c)$ as $\sigma_{\varepsilon}^2 \to \infty$. Thus we have $\tau_{A,\infty}(c) = \tau_{RC,\infty}(c)$, which satisfy $\tau'_{A,0}(c) < \tau_{A,\infty}(c) < -1$.

For the very last statement in the proposition, first note that when $\bar{\theta} = \frac{1}{2}$ and $c = \frac{1}{2}$, then $\theta^* = \frac{1}{2}$. Furthermore, note that then $\Phi^{-1}(c) = 0$ and so $\partial \theta^*/\partial \sigma_{\varepsilon}^2 = 0$ (since $\theta^* = \frac{1}{2}$ and $T_{RC} = \bar{\theta} = \frac{1}{2}$). This implies in turn that $T_A = \frac{1}{2}$ as well, so we have $\frac{\partial \Phi_{RC}(\theta^*)}{\partial \sigma_{\varepsilon}^2} = \frac{\partial A(\theta^*)}{\partial \sigma_{\varepsilon}^2} = 0$.

**Proof of Lemma 4**

The probability of regime change in the perfect coordination benchmark with no information aggregation is $\Phi_{RC}^{PC} = P(x \leq x_{PC}^* \cap \theta \leq 1)$. Since $\theta$ and $x = \theta + \varepsilon$ represent a bivariate normal distribution with correlation coefficient $\rho = \rho(\theta, x) = \sqrt{\sigma_{\theta}^2 / \sigma_{\theta}^2 + \sigma_{\varepsilon}^2}$, we can use the properties of multivariate normal distributions.

Note from (13) that setting $\bar{\theta} = 1$ and $c = \frac{1}{2}$ implies that $x_{PC}^* = \bar{\theta} = 1$. Consider then variables

$$x_1 = (x - x_{PC}^*) / \sqrt{\sigma_{\theta}^2 + \sigma_{\varepsilon}^2}, \quad (51)$$

and

$$x_2 = (\theta - \bar{\theta}) / \sqrt{\sigma_{\theta}^2}. \quad (53)$$

Variables $x_1$ and $x_2$ follow a standardised bivariate normal distribution, i.e. one with $(\mu_1, \mu_2) = (0, 0)$ and $(\sigma_1, \sigma_2) = (1, 1)$, and have a correlation coefficient $\rho = \sqrt{\sigma_{\theta}^2 / \sigma_{\theta}^2 + \sigma_{\varepsilon}^2}$.

The value of the CDF for bivariate standard normal distribution is given by:

$$F(a_1, a_2) = \frac{1}{2\pi \sqrt{1 - \rho^2}} \int_{-\infty}^{a_2} \int_{-\infty}^{a_1} \exp \left( -\frac{x_1^2}{2} + \rho x_1 x_2 \frac{x_2^2}{2} \right) dx_1 dx_2. \quad (55)$$

40
By carrying out the integration with respect to \( x_1 \) and then differentiating with respect to \( \rho \), we obtain the following:

\[
-\frac{1}{2\pi (1 - \rho^2)} \int_{-\infty}^{\infty} \exp \left( a_1 - 2\rho a_1 x_2 + \frac{x_2^2}{2(\rho^2 - 1)} \right) (x_2 - \rho a_1) \, dx_2.
\]

This can be integrated directly to yield the probability distribution function:

\[
f(a_1, a_2) = \frac{1}{2\pi\sqrt{1 - \rho^2}} \exp \left( -\frac{a_1^2}{2} - \rho a_1 a_2 + \frac{a_2^2}{2} \right). \tag{57}
\]

We may reverse the order of integration and differentiation, which means that for all \((a_1, a_2)\) we have \(\frac{\partial}{\partial \rho} F(a_1, a_2) = f(a_1, a_2)\). Since \(f(a_1, a_2) > 0\) everywhere, this implies that in fact \(F(a_1, a_2)\) is strictly increasing in \(\rho\). We are interested, however, in the values of \(a_1\) and \(a_2\) for which \(\frac{\partial}{\partial \rho} F(a_1, a_2) = f(a_1, a_2)\) is maximised. To obtain these, we simply need to find the minimum of function

\[
g(a_1, a_2) = \frac{a_1^2}{2} - \rho a_1 a_2 + \frac{a_2^2}{2}. \tag{58}
\]

Taking the first order condition with respect to \(a_1\) and \(a_2\) shows that the only critical point is \((a_1^*, a_2^*) = (0, 0)\). Furthermore, \(\frac{\partial}{\partial a_1} g(a_1^*, a_2^*) = \frac{\partial}{\partial a_2} g(a_1^*, a_2^*) = 1\) and \(\frac{\partial}{\partial a_1 \partial a_2} g(a_1^*, a_2^*) = -\rho\). Since

\[
D(a_1^*, a_2^*) = \frac{\partial}{\partial a_1^2} g(a_1^*, a_2^*) \frac{\partial}{\partial a_2^2} g(a_1^*, a_2^*) - \left( \frac{\partial}{\partial a_1 \partial a_2} g(a_1^*, a_2^*) \right)^2 = 1 - \rho^2 > 0 \tag{59}
\]

and \(\frac{\partial}{\partial a_1^2} g(a_1^*, a_2^*), \frac{\partial}{\partial a_2^2} g(a_1^*, a_2^*) > 0\), it follows that \((a_1^*, a_2^*) = (0, 0)\) is indeed the minimum of \(g(a_1, a_2)\). Hence, \(\frac{\partial}{\partial \rho} F(a_1, a_2)\) is maximised for \((a_1^*, a_2^*) = (0, 0)\) for all \(\rho\).

The fact that with \(\theta = 1\) and \(c = \frac{1}{2}\) we have \(x_{PC}^* = \bar{\theta} = 1\) for all \(\sigma^2_{\varepsilon}\) implies that we can apply this analysis to our model. Note that we can write

\[
\mathbb{P}(x < x_{PC}^*, \theta < 1) = \mathbb{P} \left( \frac{x - \bar{\theta}}{\sqrt{\sigma^2_{\theta} + \sigma^2_{\varepsilon}}} < \frac{x_{PC}^* - \bar{\theta}}{\sqrt{\sigma^2_{\theta} + \sigma^2_{\varepsilon}}}, \frac{\theta - \bar{\theta}}{\sqrt{\sigma^2_{\theta}}} < \frac{1 - \bar{\theta}}{\sqrt{\sigma^2_{\theta}}} \right) \tag{60}
\]

\[
= \mathbb{P}(x_1 < a_1, x_2 < a_2), \tag{61}
\]

and \(x_{PC}^* - \bar{\theta} = 0\) and \(\bar{\theta} - 1 = 0\) (so that \(a_1 = 0\) and \(a_2 = 0\)) holds if and only if \(\bar{\theta} = 1\) and \(c = \frac{1}{2}\). Therefore, \(\frac{\partial}{\partial \rho} \mathbb{P}(x < x_{PC}^*, \theta < 1)\) is maximised for \(\bar{\theta} = 1\) and \(c = \frac{1}{2}\). Finally,
since $\rho = \sqrt{\frac{\sigma_\theta^2}{\sigma_\theta^2 + \sigma_\varepsilon^2}}$, i.e. it is strictly decreasing in $\sigma_\varepsilon^2$, we have that $\frac{\partial}{\partial \sigma_\varepsilon^2} (x < x_{PC}, \theta < 1)$ is minimised for all $\sigma_\varepsilon^2$ when $\tilde{\theta} = 1$ and $c = \frac{1}{2}$.

**Proof of Lemma 5**

Unfortunately, in general, there are no closed form solutions for the cumulative distribution function $F(x) = \mathbb{P}(X \leq x)$, where $X \sim \mathcal{N} (\mu, \Sigma)$. There is, however, an analytical solution for the standardised bivariate normal distribution in the special case where $x = (0, 0)$. Setting $\tilde{\theta} = 1$ and $c = \frac{1}{2}$ allows us to use this property in the model.

The quadrant probability in this special case is then given analytically by:

$$\mathbb{P}(x_1 \leq 0, x_2 \leq 0) = \int_{-\infty}^{0} \int_{-\infty}^{0} P(x_1, x_2) \, dx_1 \, dx_2$$

$$= \frac{1}{4} + \frac{\arcsin \rho}{2\pi}.$$  \hspace{1cm} (64)

From the proof of Lemma 4, we know that variables $x_1$ and $x_2$ follow a standardised bivariate normal distribution, i.e. one with $(\mu_1, \mu_2) = (0, 0)$ and $(\sigma_1, \sigma_2) = (1, 1)$, and have a correlation coefficient $\rho = \sqrt{\frac{\sigma_\theta^2}{\sigma_\theta^2 + \sigma_\varepsilon^2}}$. Applying this to (64), we obtain the following:

$$\Phi_{PC} = \mathbb{P}(x \leq x_{PC}, \theta \leq 1)$$

$$= \mathbb{P}(x_1 \leq 0, x_2 \leq 0)$$

$$= \frac{1}{4} + \frac{\arcsin \left( \sqrt{\frac{\sigma_\theta^2}{\sigma_\theta^2 + \sigma_\varepsilon^2}} \right)}{2\pi}.$$  \hspace{1cm} (67)

Parts (i)-(iii) are easy to see once we consider the first derivative of $\Phi_{PC}^{RC}$ with respect to $\sigma_\varepsilon^2$ at $\tilde{\theta} = 1$ and $c = \frac{1}{2}$:

$$\frac{\partial}{\partial \sigma_\varepsilon^2} \Phi_{PC}^{RC} \bigg|_{\tilde{\theta}=1, c=\frac{1}{2}} = -\frac{\sqrt{\sigma_\theta^2}}{4\pi (\sigma_\theta^2 + \sigma_\varepsilon^2) \sqrt{\sigma_\varepsilon^2}}.$$  \hspace{1cm} (68)

This is clearly negative $\forall \sigma_\theta^2, \sigma_\varepsilon^2 > 0$. Furthermore, $\lim_{\sigma_\varepsilon^2 \to 0} \frac{\partial \Phi_{PC}^{RC}}{\partial \sigma_\varepsilon^2} = -\infty$ and $\lim_{\sigma_\varepsilon^2 \to \infty} \frac{\partial \Phi_{PC}^{RC}}{\partial \sigma_\varepsilon^2} = 0$.

\textsuperscript{32}See, e.g., Rose and Smith (1996, 2002) and Stuart and Ord (1998).
Proof of Proposition 5

Suppose that when $\bar{\theta} = 1$ and $c = \frac{1}{2}$. Note from the proof of Lemma 5 that then $\frac{\partial \Phi_{PC}}{\partial \sigma^2_{\varepsilon}} < 0 \quad \forall \sigma^2_{\varepsilon} > 0$. On the other hand, the probability of regime change in the global game (i.e. with imperfectly coordinated) is $\Phi_{RC} = \Phi \left( \frac{(\theta^* - \bar{\theta})}{\sqrt{\sigma^2_{\theta}}} \right)$, where $\theta^* = \Phi \left( \left( \sqrt{\sigma^2_{\varepsilon}} / \sigma^2_{\theta} \right) (\theta^* - 1) \right)$. Consequently, we have

$$\frac{\partial \Phi_{RC}}{\partial \sigma^2_{\varepsilon}} = \phi \left( \frac{\theta^* - \bar{\theta}}{\sqrt{\sigma^2_{\theta}}} \right) \frac{1}{\sqrt{\sigma^2_{\theta}}} \frac{\partial \theta^*}{\partial \sigma^2_{\varepsilon}}. \quad (69)$$

where

$$\frac{\partial \theta^*}{\partial \sigma^2_{\varepsilon}} = \phi \left( \frac{\sqrt{\sigma^2_{\varepsilon}}}{\sigma^2_{\theta}} (\theta^* - 1) \right) \frac{\theta^* - 1}{2 \sqrt{\sigma^2_{\theta} \sigma^2_{\varepsilon}} \left( 1 - \frac{\sqrt{\sigma^2_{\varepsilon}}}{\sigma^2_{\theta}} \right)}. \quad (70)$$

This is also negative since $\theta^* \in (0, 1)$ and $\sqrt{\sigma^2_{\varepsilon}} / \sigma^2_{\theta}$ by the equilibrium uniqueness condition stated in Proposition 1 $\sigma^2_{\varepsilon} \leq 2 \pi (\sigma^2_{\theta})^2$. Hence, when $\bar{\theta} = 1$ and $c = \frac{1}{2}$, both $\frac{\partial \Phi_{PC}}{\partial \sigma^2_{\varepsilon}}$ and $\frac{\partial \Phi_{RC}}{\partial \sigma^2_{\varepsilon}}$ are negative $\forall \sigma^2_{\varepsilon} > 0$.

To arrive at the result in Proposition 5 note that

$$\lim_{\sigma^2_{\varepsilon} \to 0} \Phi_{RC} (\theta^*) \mid_{\bar{\theta}=1,c=\frac{1}{2}} = \Phi \left( \frac{-1}{2 \sqrt{\sigma^2_{\theta}}} \right) \quad (71)$$

$$\lim_{\sigma^2_{\varepsilon} \to \infty} \Phi_{RC} (\theta^*) \mid_{\bar{\theta}=1,c=\frac{1}{2}} = \Phi \left( \frac{-1}{\sqrt{\sigma^2_{\theta}}} \right) \quad (72)$$

$$\lim_{\sigma^2_{\varepsilon} \to 0} \Phi_{PC} = \frac{1}{2} \quad (73)$$

$$\lim_{\sigma^2_{\varepsilon} \to \infty} \Phi_{PC} = \frac{1}{4}. \quad (74)$$

Since $\Phi \left( -\frac{1}{2} / \sqrt{\sigma^2_{\theta}} \right) < \frac{1}{2}$, there always exists a value of $\sigma^2_{\varepsilon,l}$ such that whenever $\sigma^2_{\varepsilon} < \sigma^2_{\varepsilon,l}$, the probability of regime change in the imperfect coordination model will be greater than that in the perfect coordination benchmark with no information aggregation.

Furthermore, if $\Phi \left( -1 / \sqrt{\sigma^2_{\theta}} \right) > \frac{1}{4}$, i.e. $\sigma^2_{\theta} > \left( \Phi^{-1} \left( \frac{1}{4} \right) \right)$, then there exist a value of $\sigma^2_{\varepsilon,h}$ such that whenever $\sigma^2_{\varepsilon} > \sigma^2_{\varepsilon,h}$, the probability of regime change in the imperfect coordination model will be greater than that in the perfect coordination benchmark with no information aggregation.
A.2 Further Discussion

Spillovers and Repressions

Suppose the agents’ payoffs are as follows:

<table>
<thead>
<tr>
<th></th>
<th>Attack is successful</th>
<th>Status quo remains</th>
</tr>
</thead>
<tbody>
<tr>
<td>Attack</td>
<td>$1 - c$</td>
<td>$-c - r$</td>
</tr>
<tr>
<td>Do Not Attack</td>
<td>$s$</td>
<td>$0$</td>
</tr>
</tbody>
</table>

As before, parameter $c \in (0, 1)$ denotes the agents’ individual cost of attacking the regime.

We introduce here an additional cost, $r$, when an agent attacks the regime and the attack turns out to be unsuccessful. This reflects the fact that an agent may then be repressed by the incumbent regime.

Furthermore, we also allow for a positive payoff of $s$ for a citizen who does not attack the regime in a situation when the attack turns out to be successful because a sufficient measure of other agents have attacked the regime. We assume that $s \in [0, 1 - c)$, where $s$ illustrates a spillover effect. The difference $1 - s$ measures the extent to which, once the regime is overthrown, those who had been among the attackers gain privileged status relative to those who held back.\(^{33}\)

By assuming that $s < 1 - c$, we ensure that the game does not collapse to a prisoner’s dilemma. If it were not the case, then not attacking would be a strictly dominant strategy. It is also worth noting that, as long as $s > 0$, there are positive externalities of attacking: if an agent attacks the current regime and the attack turns out to be successful, the non-attackers are also better off. In principle, one could also imagine a situation where $s < 0$, i.e. that the agents who do not participate in the attack are worse off after the regime change, for example, because of prosecution by the new regime, ostracism leading to diminished career prospects, etc.

With the payoffs specified as above, the agent’s payoff indifference condition is:

\[
(1 - c) \Pr(\text{attack succesful} \mid x^*) + (-c - r) (1 - \Pr(\text{attack succesful} \mid x^*)) = s \Pr(\text{attack succesful} \mid x^*)
\]

This simplifies to:

\[
\Pr(\text{attack succesful} \mid x^*) = \frac{c + r}{1 - s}, \tag{75}
\]

\(^{33}\)The idea of a privileged status of attackers appears in political science literature, e.g., Bueno de Mesquita (2010)
With the structure of the equilibrium defined as in Section 2, this yields

\[ Pr(\theta \leq \theta^* | x^*) = \frac{c + r}{1 - s}, \quad (76) \]

and the payoff indifference condition is then:

\[ \frac{c + r}{1 - s} = \Phi \left( \sqrt{\frac{\sigma_\theta^2 + \sigma_\varepsilon^2}{\sigma_\theta^2 \sigma_\varepsilon^2}} \left( \theta^* - \bar{\theta} - \frac{\sigma_\theta^2}{\sigma_\theta^2 + \sigma_\varepsilon^2} \left( x^* - \bar{\theta} \right) \right) \right). \quad (77) \]

Given that the new payoff structure does not affect the critical mass condition, an equilibrium value of \( \theta^* \) is given by:

\[ \theta^* = \Phi \left( \frac{\sqrt{\sigma_\varepsilon^2}}{\sigma_\theta^2} \left( \theta^* - \bar{\theta} \right) - \sqrt{\frac{\sigma_\theta^2 + \sigma_\varepsilon^2}{\sigma_\theta^2}} \Phi^{-1} \left( \frac{c + r}{1 - s} \right) \right). \quad (78) \]

As we can see, the only way in which the introduction of parameters \( r \) and \( s \) affects \( (7) \) is through the ratio \( \frac{c + r}{1 - s} \). In other words, in this model, the impact of \( r > 0 \) or \( s > 0 \) is equivalent to an appropriately scaled increase in \( c \).

**Example: Online Commentators**

I present here how the model, with a few additions, could be applied to the context of online commentators (or “internet trolls”) who are employed by authoritarian regimes to post comments that are favourable about the incumbent. This exercise is helpful in that it highlights the kind of assumptions that have to be made to apply the familiar setting outlined above to a particular context.

Suppose that, in addition to the regime and the agents, there is a continuum of online commentators of measure 1, who are indexed by \( i \) like the agents and uniformly distributed over \([0, 1]\). This means that there is a one-to-one correspondence (bijective function) between commentators and agents, i.e. each agent follows only one commentator and each commentator is listened to by only one agent.

Each commentator \( i \) has a certain intrinsic valuation for money, which is denoted by \( v \). Their utility from praising the government at a level \( b_i \in R \) is given by:

\[ U_o := (\theta + v_i y) b_i - c_o(b_i), \quad (79) \]

where \( \theta \) is the state of the world and \( y \) is a payment that the online commentator receives for an additional unit of \( b_i \). The praise level, \( b_i \), can be interpreted as the number of
**pro-government posts written by an online commentator.** Intuitively, the stronger is the regime, $\theta$, the more eager is a commentator to write positive comments about the government (*intrinsic motivation*). This may also reflect the fact that it is easier for the commentator to praise the government when the regime is strong. But the degree to which he praises the government depends also on the payment, $y$, which he receives from the government (*extrinsic motivation*). The extent to which a commentator is responsive to this payment depends on his intrinsic valuation for money, $v$.

However, online commentators incur also a cost of writing comments, $c_o(b_i) = \frac{1}{2}kb_i^2$, which is increasing in the praise level, $b_i$. Parameter $k$ measures here the relative importance of the cost of effort in the online commentator’s utility function. The optimal choice of the number of pro-government comments written is then $b_i^* = (\theta + viy)/k$. For simplicity, I assume that $k = 1$.

In order to make the global games model applicable to this context, I assume that online commentators are the only players to observe the exact values of $\theta$ and $v_i$; the former is realised after the contract is signed, and the latter is each commentator’s private information. However, the distribution of $\theta$ and $v$ is common knowledge among all players, with $v \sim \mathcal{N}(\bar{v}, \sigma_v^2)$ and no correlation between $\theta$ and $v$.

Each agent $i$ observes the online commentator’s action, $b_i$, and uses it to update her information about the prior. Given that and the publicly known value of the payment, $y$, the posterior of an agent $i$ is then

$$\mathbb{E}[\theta | b_i] = \frac{y^2\sigma_v^2}{\sigma_\theta^2 + y^2\sigma_v^2} \bar{\theta} + \frac{\sigma_\theta^2}{\sigma_\theta^2 + y^2\sigma_v^2} (b_i - \bar{v}y),$$

which is equivalent to (I) since $b_i - \bar{v}y$ is normally distributed with mean zero and variance $y^2\sigma_v^2$. In other words, an increase in $y$ acts in exactly the same way as an appropriately scaled increase in $\sigma_v^2$. Furthermore, given the bijective relationship between commentators and agents, the noise is identically and independently distributed across agents. Finally, no correlation between $\theta$ and $v$ ensures, given their normal distributions, that noise is

---

34In fact, an online commentator’s optimal action, $b_i$, would be a measure of both the quantity of pro-government posts and the extent to which they praise the regime (their “intensity”).

35These preferences of online commentators are inspired by Bénabou and Tirole (2003, 2006), who study the behaviour of agents choosing the extent of their participation in some prosocial activity, for example, contribution to a public good.

36This means that there is no correlation between the regime’s strength and the commentators’ intrinsic valuation for money. It could be argued, however, that allowing for $\text{Cov}(\theta, v) > 0$ would be more realistic. For example, when the regime is indeed strong, online commentators might feel less guilty about being responsive to the payment they receive, and thus their intrinsic valuation for money may also be more likely to be high.
independent of $\theta$. The setting with online commentators thus becomes equivalent to the global games model of Section 2.1 with the following timeline:

1. The nature draws $\theta$ from a normal distribution $\mathcal{N}(\bar{\theta}, \sigma^2 \theta)$, which defines the initial common prior about the regime’s strength, $\theta$.

2. The government (publicly) offers a common contract to each commentator: a payment of $y$ per unit of praise $b$.

3. Each commentator $i$ observes the regime’s strength, $\theta$, and based on that and on his (private) intrinsic valuation for money, $v$, he chooses his level of praise for the government, $b$.

4. Each citizen $i$ observes the commentator’s level of praise, $b$, forms a posterior about the regime’s strength, $\theta$, and decides whether to attack the regime or not.

5. The regime is overthrown, or it survives. All players’ payoffs are realised.

References


